Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An EMT-Based Five-Gene Prognostic Model and Single-Cell Analysis of SPP1 Expression in Hepatocellular Carcinoma

Xinyuan Chen
Guangxi University of Chinese Medicine
Lu Lu
Guangxi University of Chinese Medicine
Chengting Wu
Guangxi University of Chinese Medicine
Yuanqin Du
Guangxi University of Chinese Medicine
Yu Zhang
Guangxi University of Chinese Medicine
Jiazhi Li
The People's Hospital of Guangxi Zhuang Autonomous Region
Yanhua Lai
The People's Hospital of Guangxi Zhuang Autonomous Region
Jingjing Huang
55869563@qq.com
The First Affiliated Hospital of Guangxi University of Traditional Chinese Medicine
Hongna Huang
The First Affiliated Hospital of Guangxi University of Traditional Chinese Medicine

Research Article

Keywords: Hepatocellular carcinoma (HCC), Epithelial–mesenchymal transition (EMT), Prognostic model, Single-cell RNA sequencing, SPP1, Immunohistochemistry

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11160878/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

An EMT-Based Five-Gene Prognostic Model and Single-Cell Analysis of SPP1 Expression in Hepatocellular Carcinoma

Xinyuan Chen [1,†], Lu Lu[1,†], Chengting Wu [1], Yuanqin Du[1], Yu Zhang [1], Jiazhi Li[4], Yanhua Lai[4], Jingjing Huang[2,3,*] , and Hongna Huang[2,*]

[1]Guangxi University of Chinese Medicine, Nanning, China;

[2]The First Affiliated Hospital of Guangxi University of Chinese Medicine, Nanning, China;

[3]Guangxi Key Laboratory of Translational Medicine for Treating High-Incidence Infectious Diseases with Integrative Medicine, Nanning, China;

[4]The People's Hospital of Guangxi Zhuang Autonomous Region, Nanning, China;

[*]Corresponding author(s): Jingjing Huang, Ph.D, E-mail: 55869563@qq.com; Tel.: +8613481186074, Fax: 8607715361260; Hongna Huang, M.D, E-mail: 18776985262@163.com. Tel: +8618776985262, Fax: +8607715361260

Contributing authors: 839029828@qq.com；460967080@qq.com；337155890@qq.com；820971789@qq.com；279737750@qq.com；drlijiazhi1@163.com；drlaiyanhua1@163.com; 55869563@qq.com； 18776985262@163.com；

[†]These authors have contributed equally to this work.

Abstract

Background and Objective: Hepatocellular carcinoma (HCC) remains a major cause of cancer-related mortality, and reliable prognostic biomarkers are still needed. Epithelial‐mesenchymal transition (EMT) is involved in HCC progression, and EMT-related genes may serve as candidates for prognostic modeling. This study aimed to construct an EMT-based five-gene prognostic model for HCC and to investigate the cellular expression pattern of SPP1 using single-cell and

immunohistochemical analyses.**Methods:** EMT-related genes were obtained from the dbEMT database and intersected with differentially expressed genes and WGCNA hub genes in the TCGA cohort. Candidate prognostic genes were screened by univariate Cox and LASSO regression, and an EMT-based five-gene prognostic model was constructed and validated in TCGA and GSE14520 datasets using survival analysis, C-index, and hazard ratios (HRs). Functional enrichment analysis and immune cell infiltration assessment were performed to compare high- and low-risk groups. Single-cell analysis using GSE189903 was conducted to investigate the cellular distribution of SPP1. Immunohistochemistry (IHC) was performed on human HCC tissues and paired adjacent and normal liver tissues to preliminarily assess SPP1 protein expression.**Results:** The EMT-based five-gene prognostic model, consisting of MMP1, EZH2, SPP1, CCNA2, and MYCN, showed modest predictive performance in both training and validation cohorts. High-risk tumors were enriched in proliferation, cell cycle, and MYC target pathways, accompanied by immune microenvironment dysregulation. Single-cell analysis showed that SPP1 was expressed in multiple cell types, with relatively higher expression in myeloid cells (including macrophages, mast cells, and neutrophils) and detectable expression in hepatocarcinoma cells. Preliminary IHC analysis showed that SPP1 protein was elevated in HCC tissues compared with adjacent and normal liver tissues, with positive signals observed in hepatocellular carcinoma cells and some stromal areas, consistent with the single-cell finding.**Conclusion:** The EMT-based five-gene prognostic model showed modest predictive performance for survival risk in HCC patients. SPP1 was expressed in myeloid cells and hepatocarcinoma cells in single-cell data and was elevated in HCC tissues by IHC. However, the cellular source and functional role of SPP1 in HCC progression remain to be determined. These findings are descriptive and hypothesis-generating.

Keywords: Hepatocellular carcinoma (HCC); Epithelial‐mesenchymal transition

(EMT); Prognostic model; Single-cell RNA sequencing; SPP1; Immunohistochemistry

Introduction

Hepatocellular carcinoma (HCC) represents the predominant histological subtype of primary liver cancer and continues to impose a substantial global health burden. Liver cancer ranks among the most frequently diagnosed malignancies worldwide and remains a major contributor to cancer-related mortality, with increasing death rates reported in several regions [1 – 3]. Despite improvements in surveillance and therapeutic strategies, the prognosis of HCC patients remains poor, largely because of the high propensity for tumor invasion and distant metastasis. In clinical practice, a considerable proportion of patients are diagnosed at advanced stages, when curative treatment options are limited. Moreover, even after potentially curative therapies, the long-term outcome remains unsatisfactory due to high recurrence and metastasis rates, which can exceed 70% within five years [4 – 8]. Therefore, identifying reliable prognostic biomarkers and elucidating the molecular features of HCC progression are essential for improving patient prognosis and developing more effective therapeutic strategies.

The development of HCC is strongly associated with multiple etiological factors. Chronic infection with hepatitis B virus (HBV) has historically been considered the most important risk factor and is responsible for a substantial proportion of global HCC cases. However, with the implementation of large-scale vaccination programs and the increasing availability of antiviral therapies, the epidemiological landscape of HCC has gradually changed. In recent years, metabolic disorders, particularly those associated with non-alcoholic steatohepatitis (NASH) and metabolic dysfunction-associated fatty liver disease, have emerged as increasingly important contributors to hepatocarcinogenesis [9]. These distinct etiological backgrounds are

closely linked to differences in tumor biology, genetic alterations, and immune microenvironment characteristics, thereby complicating the development of personalized therapeutic strategies for HCC patients.

At present, surgical resection and liver transplantation remain the most effective curative treatment approaches for patients with early-stage HCC. In addition to surgical interventions, several locoregional therapies have been widely adopted in clinical practice. Radiofrequency ablation is commonly used for patients with small tumors who are not suitable candidates for surgery, while transarterial chemoembolization (TACE) has been considered the standard treatment for patients with intermediate-stage disease for more than two decades [10,11]. Although these conventional treatments have significantly improved clinical management, the high rates of tumor recurrence and metastasis remain major obstacles to achieving long-term survival benefits. In recent years, advances in molecular targeted therapies and immunotherapy have provided new opportunities for the treatment of HCC. Nevertheless, reliable molecular biomarkers for risk stratification and prognosis prediction are still needed to further improve therapeutic outcomes.

HCC typically develops in a background of chronic inflammation and is characterized by a complex and dynamic tumor microenvironment (TME). The TME consists of diverse cellular and molecular components, including various immune cell populations, cytokines, extracellular matrix components, and multiple signaling pathways that collectively influence tumor behavior [12 – 14]. Among these components, several immune cell types have been shown to contribute to immunosuppressive conditions within the TME. For example, regulatory T cells (Tregs) can impair the antitumor activity of cytotoxic T lymphocytes, while myeloid-derived suppressor cells (MDSCs) inhibit T-cell responses through the production of immunosuppressive factors such as transforming growth factor-β. In addition, tumor-associated macrophages (TAMs) and neutrophils have also been

implicated in promoting tumor growth, immune evasion, and metastatic progression. Increasing evidence suggests that the interactions among these immune cells may play a crucial role in shaping the immunosuppressive microenvironment of HCC, which in turn may facilitate tumor progression and metastasis [15,16]. Notably, the immunological characteristics of the TME have been reported to closely interact with epithelial–mesenchymal transition (EMT), a key biological process involved in cancer metastasis.

EMT is widely recognized as a fundamental mechanism underlying tumor invasion and metastatic dissemination [17–19]. During this process, epithelial tumor cells undergo a series of molecular and phenotypic changes that result in the loss of epithelial characteristics and the acquisition of mesenchymal features, thereby enhancing migratory and invasive capabilities [20,21]. Hallmark molecular events of EMT include the downregulation of epithelial markers such as E-cadherin and the upregulation of mesenchymal markers such as Vimentin. In addition, several transcription factors, including Snail, Slug, Twist1, ZEB1, and β-catenin, play essential roles in regulating EMT initiation and maintenance [22–28]. Previous studies have identified multiple EMT-related genes associated with HCC prognosis and tumor progression. However, most existing studies rely primarily on bulk transcriptomic data, and comprehensive analyses integrating single-cell transcriptomic information to investigate the cellular distribution and functional roles of EMT-related genes within the tumor microenvironment remain limited.

In the present study, we aimed to identify EMT-related prognostic genes in HCC by integrating multiple public transcriptomic datasets. Specifically, gene expression profiles from The Cancer Genome Atlas and Gene Expression Omnibus were analyzed using machine learning-based approaches to screen for candidate prognostic genes from an EMT-related gene set. An EMT-based five-gene prognostic model was constructed and validated, and functional enrichment and immune infiltration analyses

were performed to characterize the high- and low-risk groups. Furthermore, single-cell RNA sequencing data were employed to investigate the cellular expression pattern of SPP1, and immunohistochemistry (IHC) was performed to preliminarily assess SPP1 protein expression in HCC tissues, adjacent non-tumor tissues, and normal liver tissues. This work aimed to provide a descriptive foundation for future studies on SPP1 in HCC. No functional experiments were performed, and the mechanistic role of SPP1 remains to be determined.

Materials and Methods

Patients and Specimens

The specimens used in this study consisted of three pairs of hepatocellular carcinoma (HCC) tissues and their corresponding cancer-adjacent (CAP) and non-cancer (CAF) tissues, collected from patients who underwent liver transplantation at Guangxi Zhuang Autonomous Region People’s Hospital between 2013 and 2024. All specimens were stored in accordance with institutional regulations at the Biobank of Guangxi Zhuang Autonomous Region People’s Hospital (Biobank Approval No.: Guo Ke Yi Ban Shen Zi [2023] BC0039). The study protocol was approved by the Ethics Committee of Guangxi University of Chinese Medicine, and all patients provided written informed consent for the use of their tissues.

Data Sources

Transcriptomic data and corresponding clinical information of patients with HCC were retrieved from the The Cancer Genome Atlas through the Genomic Data Commons portal (https://portal.gdc.cancer.gov/). The TCGA-LIHC cohort included 369 tumor samples and 50 normal liver tissue samples, which were utilized for differential gene expression analysis. After removing cases lacking complete overall survival (OS) information, 363 HCC samples remained and were included in subsequent survival and prognostic analyses.

To further validate the findings, an independent dataset (GSE14520) was obtained

from the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/). This dataset contains 247 HCC samples and 241 normal liver tissue samples. Following the exclusion of samples with missing survival data, 221 HCC cases were retained for downstream analyses.

Single-cell RNA sequencing data used for investigating the cellular distribution of key genes were obtained from the GSE189903 dataset in the GEO database.

To identify epithelial‑mesenchymal transition (EMT)-associated genes potentially involved in HCC progression, an EMT-related gene list was downloaded from the dbEMT database. This dataset contains 1,184 genes reported to participate in EMT processes [29,30]. These genes were used as candidate features for subsequent analyses aimed at identifying EMT-related prognostic biomarkers in HCC.

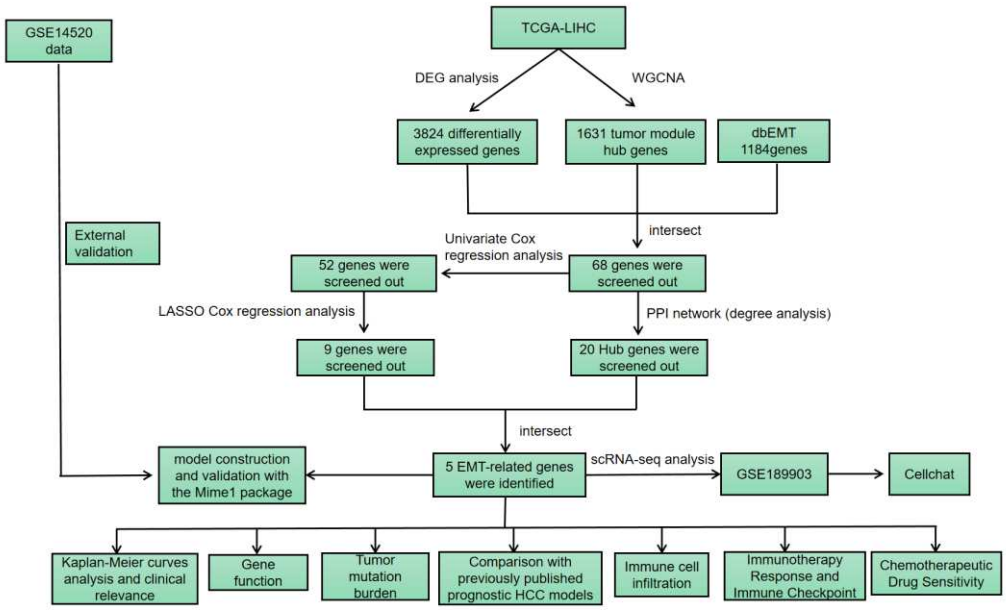


Figure 1. The workflow of this study includes the use of the following datasets and analytical methods: TCGA-LIHC (The Cancer Genome Atlas Liver Hepatocellular Carcinoma dataset), GSE14520, GSE189903 (GEO database datasets), MSigDB (Molecular Signatures Database), DEG (differentially expressed genes), LASSO (least absolute shrinkage and selection operator) regression analysis, and WGCNA (weighted gene co-expression network analysis for

identifying key gene modules associated with clinical traits).

Identification and Analysis of EMT-Related Genes

Differential gene expression analysis between hepatocellular carcinoma (HCC) tissues and normal liver samples in the The Cancer Genome Atlas cohort was conducted using the limma package in R. Genes with an absolute log2 fold change greater than 1 and a false discovery rate (FDR) below 0.05 were defined as differentially expressed genes (DEGs).

To further identify gene modules closely associated with tumor characteristics, Weighted Gene Co-expression Network Analysis (WGCNA) was performed [31]. Prior to network construction, the expression matrix was filtered to remove genes with low expression levels, and the top 30% of genes with the highest variance were retained for analysis. Pearson correlation coefficients were used to calculate pairwise gene correlations. A suitable soft-thresholding power (β) was selected to ensure scale-free network topology, after which a weighted adjacency matrix was generated and subsequently transformed into a topological overlap matrix (TOM). Gene modules were then identified through hierarchical clustering combined with the dynamic tree cut algorithm. Modules with highly similar expression profiles were merged based on module eigengene similarity. Correlation analysis between module eigengenes and clinical traits (tumor versus normal tissues) was performed to identify tumor-associated modules. Within these modules, gene significance (GS) and module membership (MM) were calculated, and genes meeting the criteria of $|GS| > 0.15$ and $|MM| > 0.4$ were considered candidate hub genes.

Next, the identified DEGs, candidate hub genes from WGCNA, and the epithelial–mesenchymal transition (EMT) gene set were intersected to obtain EMT-related genes that were aberrantly expressed in HCC. Based on the TCGA cohort, univariate Cox proportional hazards regression analysis was performed to evaluate the association between these genes and overall survival (OS). Genes with a P value less than 0.05

were regarded as prognostically relevant. To further reduce feature dimensionality and avoid model overfitting, least absolute shrinkage and selection operator (LASSO) regression analysis was subsequently applied, resulting in the identification of the final set of candidate prognostic genes.

To explore potential interactions among these genes, protein–protein interaction (PPI) analysis was performed using the STRING database (https://cn.string-db.org/) with a minimum interaction confidence score of 0.4. The resulting interaction network was visualized using Cytoscape software, and hub genes within the network were identified according to their degree values. The top 20 genes with the highest connectivity were selected and intersected with the genes obtained from the LASSO regression analysis to determine the final set of EMT-related genes associated with HCC.

Subsequently, predictive models were constructed using the R package "Mime1", which integrates ten commonly used machine learning algorithms [32]. Model performance was evaluated in both the training and validation cohorts by comparing the concordance index (C-index) and the area under the receiver operating characteristic curve (AUC) at 1, 3, and 5 years. The model demonstrating the most stable and superior predictive performance across datasets was selected as the optimal prognostic model. According to the median risk score calculated from the selected model, patients in the training cohort were stratified into high-risk and low-risk groups. Kaplan–Meier survival analysis was conducted using the survminer package to compare survival differences between groups, and time-dependent ROC curves were generated to evaluate the predictive accuracy of the model at different time points.

Finally, risk scores were calculated for patients in both the TCGA cohort and the GSE14520 validation dataset obtained from the Gene Expression Omnibus. Samples were classified into high-risk and low-risk groups according to the median risk score.

The distributions of risk scores, survival outcomes, and gene expression patterns were visualized and compared between groups. Principal component analysis (PCA) was performed to evaluate the separation of samples in multidimensional space. To further assess the reliability of the constructed model, its prognostic performance was compared with previously reported HCC prognostic models using hazard ratios (HRs) and concordance index values.

Nomogram Construction and Validation

To determine whether the constructed risk score model could independently predict the prognosis of patients with hepatocellular carcinoma (HCC), both univariate and multivariate Cox proportional hazards regression analyses were conducted using the survival package in R. Clinical variables together with the risk score grouping were analyzed in the training cohort derived from the The Cancer Genome Atlas and further validated in the external dataset GSE14520 from the Gene Expression Omnibus.

Based on the independent prognostic factors identified in the multivariate Cox analysis, a prognostic nomogram was established using the rms package to estimate the probability of overall survival (OS) at 1, 2, and 3 years. The predictive accuracy and stability of the nomogram were subsequently evaluated using calibration curves, which compared the predicted survival probabilities with the observed clinical outcomes.

Functional Enrichment Analysis

Patients were stratified into high-risk and low-risk groups according to the median risk score derived from the EMT-related prognostic model. Differential expression analysis between the two groups was conducted using the limma package in R to identify genes potentially involved in the biological differences associated with the risk signature.

To further explore the functional characteristics of these genes, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses were

performed using the clusterProfiler package. These analyses were used to investigate the potential biological processes, molecular functions, and cellular components associated with the differentially expressed genes.

In addition, gene set enrichment analysis (GSEA) was carried out to comprehensively evaluate pathway-level differences between the high-risk and low-risk groups. Gene sets were obtained from the Molecular Signatures Database (MsigDB, symbols.gmt v7.0), and the analyses were implemented using the clusterProfiler and enrichplot packages. This approach enabled the identification of biological pathways and molecular mechanisms potentially associated with EMT-related prognostic genes in hepatocellular carcinoma.

Gene Mutation Analysis

Somatic mutation data from hepatocellular carcinoma patients were obtained from the The Cancer Genome Atlas. The maftools package in R was used to analyze mutation patterns in patients classified into high-risk and low-risk groups according to the EMT-related risk model.

The analysis included the overall mutation burden, mutation classifications, and the frequency of commonly mutated genes in each group. Differences in mutation profiles between the two groups were compared to identify dominant mutation types, such as missense mutations and splice-site alterations, as well as genes with high mutation frequencies.

Furthermore, mutation patterns of key EMT-related prognostic genes included in the risk model were investigated. The relationship between mutation status and mRNA expression levels was also evaluated to explore the potential influence of gene mutations on EMT-related gene expression and clinical prognosis.

Immune Infiltration and Tumor Microenvironment Analysis

Information regarding immune cell infiltration in tumor samples was obtained from the The Cancer Genome Atlas. Based on the median risk score derived from the

EMT-related risk model, patients were categorized into high-risk and low-risk groups. To comprehensively evaluate immune cell composition in tumor samples, multiple computational algorithms were applied, including TIMER, CIBERSORT, CIBERSORT-ABS, QUANTISEQ, MCPCOUNTER, XCELL, and EPIC. These methods were used to estimate the relative abundance of various immune cell populations across samples, and the results were normalized to minimize potential biases among different datasets.

Subsequently, the immune infiltration matrices were arranged according to risk scores, and heatmaps were generated to visualize differences in immune cell distribution between the high-risk and low-risk groups across different estimation methods. In addition, the limma package was applied to compare the expression levels of commonly studied immune checkpoint genes between the two risk groups.

To further assess the potential response to immunotherapy, the Tumor Immune Dysfunction and Exclusion (TIDE) algorithm was employed to calculate TIDE scores for each patient sample. Differences in TIDE scores between the high-risk and low-risk groups were analyzed to estimate the likelihood of response to immune checkpoint blockade therapy, with lower scores indicating a higher probability of therapeutic response. TIDE data were obtained from the online resource (http://tide.dfci.harvard.edu/faq/).

Drug Sensitivity Analysis

To investigate potential therapeutic options for patients with different risk profiles, drug response prediction was performed using the oncoPredict package in R. This method estimates drug sensitivity by integrating gene expression profiles with pharmacogenomic reference datasets, enabling the prediction of individualized responses to anticancer agents.A panel of 198 commonly used drugs was included in the analysis. For each patient sample, the half-maximal inhibitory concentration (IC50) was estimated, representing the drug concentration required to suppress cell

proliferation by 50%. The predicted IC50 values were then compared between the high-risk and low-risk groups defined by the EMT-related prognostic model.Drugs showing significant differences in predicted IC50 values between the two groups were considered potentially differentially sensitive agents. These results may provide insights into personalized therapeutic strategies for hepatocellular carcinoma patients with distinct risk levels.

Single-Cell Sequencing Analysis

Raw single-cell sequencing data were imported in 10X Genomics HDF5 format, and the Seurat (v5) R package was used to create Seurat objects. Quality control (QC) was performed for each sample by calculating the proportion of mitochondrial genes (mt_percent) and hemoglobin genes (HB_percent). Low-quality cells were filtered out using the following criteria: detected genes < 300 or > 7000, UMI counts < 1000 or above the 97th percentile, mitochondrial gene proportion > 10%, and hemoglobin gene proportion > 3%. After QC, samples were merged and classified into tumor and normal groups according to their origin.

The expression matrix was then normalized using NormalizeData [33], and the top 2000 highly variable genes were selected with FindVariableFeatures. Data were scaled using ScaleData, with regression of the mitochondrial proportion effect. PCA was performed using RunPCA, and the top 30 principal components with the largest standard deviations were selected for downstream analyses. To reduce batch effects, Harmony (HarmonyIntegration) was applied for data integration.

For clustering, FindNeighbors and FindClusters were used to classify cell populations, with resolutions set between 0.1 and 1.0. Dimensionality reduction and visualization were performed using UMAP (RunUMAP) [34] to display cellular heterogeneity and population distribution. Cell types were annotated by collecting known marker genes and referencing relevant literature and publicly available single-cell liver cancer datasets. Bubble plots generated by Seurat were used to visualize the expression

patterns of marker genes across clusters, which guided the annotation of cell clusters. The integrated Seurat object was subsequently used for differential expression analysis and functional pathway enrichment analysis.

Cell–Cell Communication Analysis

Cell–cell communication is a crucial mechanism through which cells coordinate biological functions via ligand–receptor signaling. To explore interactions between different cell types, the CellChat package was used to construct cell communication networks. Based on known ligand–receptor complexes and single-cell expression data, CellChat computationally analyzes and reveals patterns and strengths of intercellular communication [35].

Immunohistochemistry (IHC) Analysis

Formalin-fixed, paraffin-embedded (FFPE) tissue specimens from pathologically confirmed hepatocellular carcinoma (HCC) patients were collected and categorized into three groups: cancer tissue (CA), cancer-adjacent tissue (CAP, pathologically confirmed to be tumor-free), and non-cancer tissue (CAF, obtained from normal liver tissue distant from the tumor or from healthy donors). Tissue sections (4 μm) were deparaffinized, rehydrated, and subjected to antigen retrieval in citrate buffer (pH 6.0). Endogenous peroxidase activity was blocked with 3% hydrogen peroxide. Sections were incubated overnight at 4° C with anti-SPP1 antibody (1:3000), followed by incubation with HRP-conjugated secondary antibody. Immunoreactivity was visualized using DAB and counterstained with hematoxylin.

SPP1 expression was quantified using ImageJ software. Three high-power fields were randomly selected per section to measure optical density, and the mean optical density (MOD) was calculated for each sample. Statistical analyses were performed using GraphPad Prism 9.5.1. One-way analysis of variance (ANOVA) followed by Tukey’s post hoc test was used for comparisons among the three groups. Data are presented as mean ± standard deviation (SD), and $P < 0.05$ was considered

statistically significant.

Statistical Analysis

All statistical analyses were performed using R software (version 4.4.2). Venn diagrams were generated using the online tool MicrobeX (https://www.bioinformatics.com.cn/)[36]. Data visualization was primarily conducted using R packages such as ggplot2, cowplot, and patchwork. Key R packages used in this study included limma (differential expression analysis), survival and survminer (survival analysis), caret (model training and cross-validation), Mime1 (integration of multiple machine learning methods for prognostic model construction), and clusterProfiler (functional enrichment analysis).

For continuous variables, comparisons between groups were conducted using t-tests or Wilcoxon rank-sum tests, while categorical variables were compared using chi-square tests or Fisher's exact tests. Survival analyses were performed using Kaplan–Meier curves and the log-rank test, and univariate and multivariate Cox regression analyses were applied to evaluate the independence of prognostic factors. Time-dependent ROC curves were used to assess model predictive accuracy, and the C-index was used to evaluate model consistency. Differences in immune infiltration and drug sensitivity were analyzed using the Wilcoxon test. Statistical significance was set at $p < 0.05$, with high significance defined as $p < 0.01$ and $p < 0.001$.

Results

Screening of Prognostic Genes for the EMT-Related Model

A total of 419 samples were obtained from the TCGA database, including 50 normal tissue samples and 369 HCC samples. Differential expression analysis identified 3,824 DEGs between tumor and normal tissues (Figure 2A). Subsequently, WGCNA was used to select 1,631 tumor-associated hub genes (Figures 2B and 2C). Meanwhile, 1,184 EMT-related genes were obtained from the dbEMT database. Intersection analysis of DEGs, WGCNA hub genes, and the EMT gene set yielded 68 candidate

genes (Figure 2D).

Univariate Cox regression and LASSO regression analyses were performed on these 68 genes (Figures 2E and 2F). Simultaneously, these 68 genes were imported into the STRING database to construct a protein–protein interaction (PPI) network, and the top 20 genes were selected based on their degree values (Figure 2G). Finally, the genes identified by LASSO regression were intersected with the PPI-selected hub genes, resulting in five key genes used to construct the machine learning-based prognostic model (Figure 2H).

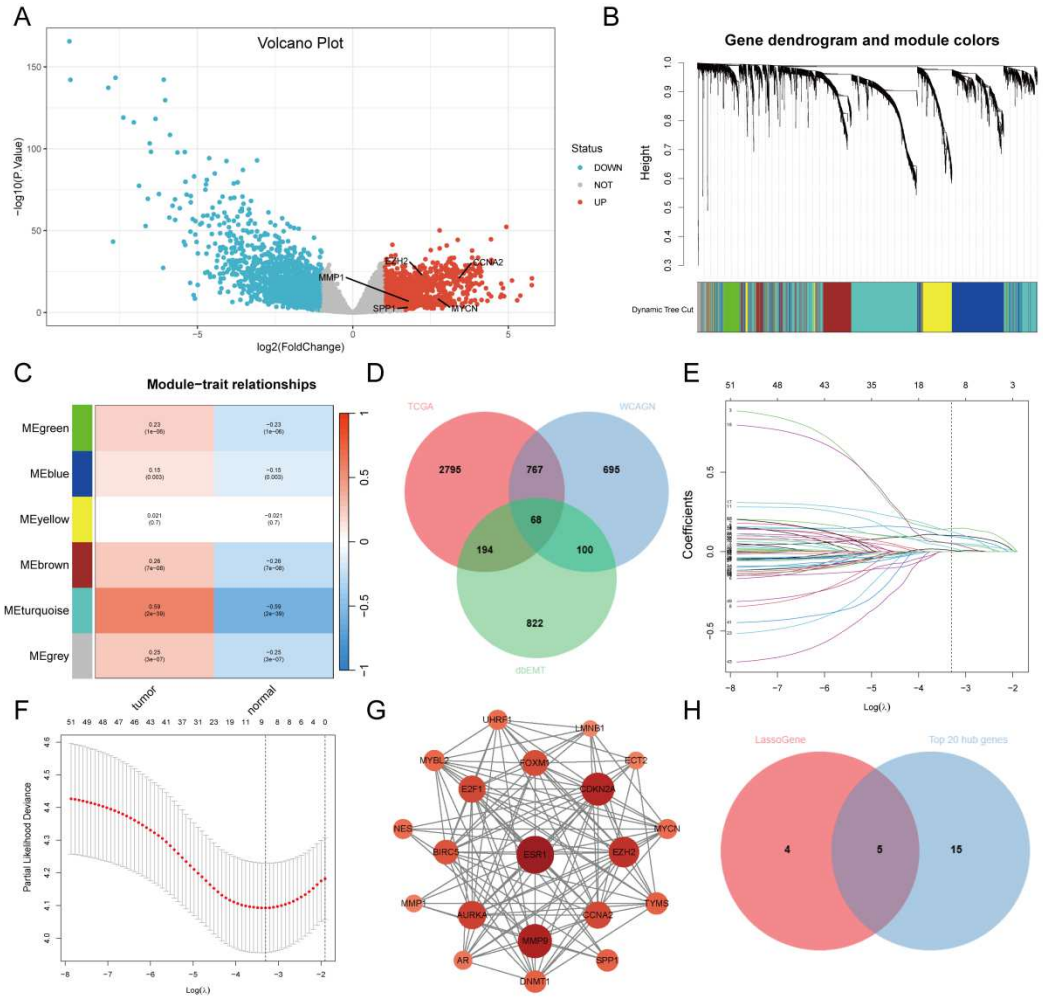


Figure 2. Workflow for screening EMT-related genes.

(A) Volcano plot of differentially expressed genes (DEGs) between TCGA HCC tissues and normal tissues.

(B) Gene dendrogram and module classification obtained from WGCNA analysis.

(C) Module – trait relationship heatmap showing correlations between WGCNA modules and clinical traits.

(D) Venn diagram displaying the intersection of DEGs, hub genes, and EMT-related genes from the dbEMT database.

(E) LASSO regression coefficient profiles of 51 EMT-related genes.

(F) Tuning plot for the LASSO regression model parameters.

(G) Top 20 hub genes in the PPI network ranked by degree values.

(H) Intersection of LASSO-selected genes and PPI hub genes yielding the final set of five prognostic genes used for model construction.

Construction and Validation of the EMT-Related Prognostic Gene Model

Using the Mime1 machine learning package, a total of 117 predictive models were fitted on the TCGA dataset, covering various algorithm combinations including Stepwise Cox, LASSO, and RSF. The C-index of each model was calculated in both the training and validation cohorts (Figure 3A). The results showed that the StepCox[forward] model achieved the highest C-index values in both the training and validation cohorts, with a relatively simple model structure, yielding average C-index values of 0.685 and 0.66, respectively (Figure 3A). In the TCGA training cohort and the independent validation cohort GSE14520, the model's C-index values were 0.71 and 0.66, respectively (Figure 3B).

Therefore, StepCox[forward] was selected as the optimal model, and its AUC values were extracted to demonstrate predictive performance (Figure 3B). The model's 1-, 3-, and 5-year prognostic predictive ability in the training and validation cohorts was further evaluated, and the corresponding ROC curves were plotted (Figures 3C–F). Patient risk scores were calculated based on the StepCox[forward] model regression coefficients using the following formula:

Risk score = 0.1432 × MMP1 + 0.1779 × EZH2 + 0.1130 × SPP1 + 0.0556 × CCNA2

+ 0.2306 × MYCN.

Patients in both the training and validation cohorts were classified into high- and low-risk groups according to the median risk score. Kaplan–Meier survival curves indicated that patients in the high-risk group had significantly lower overall survival (OS) (Figure 3G), and log-rank tests showed significant survival differences between the high- and low-risk groups, which was consistent in the validation cohort (Figure 3H).

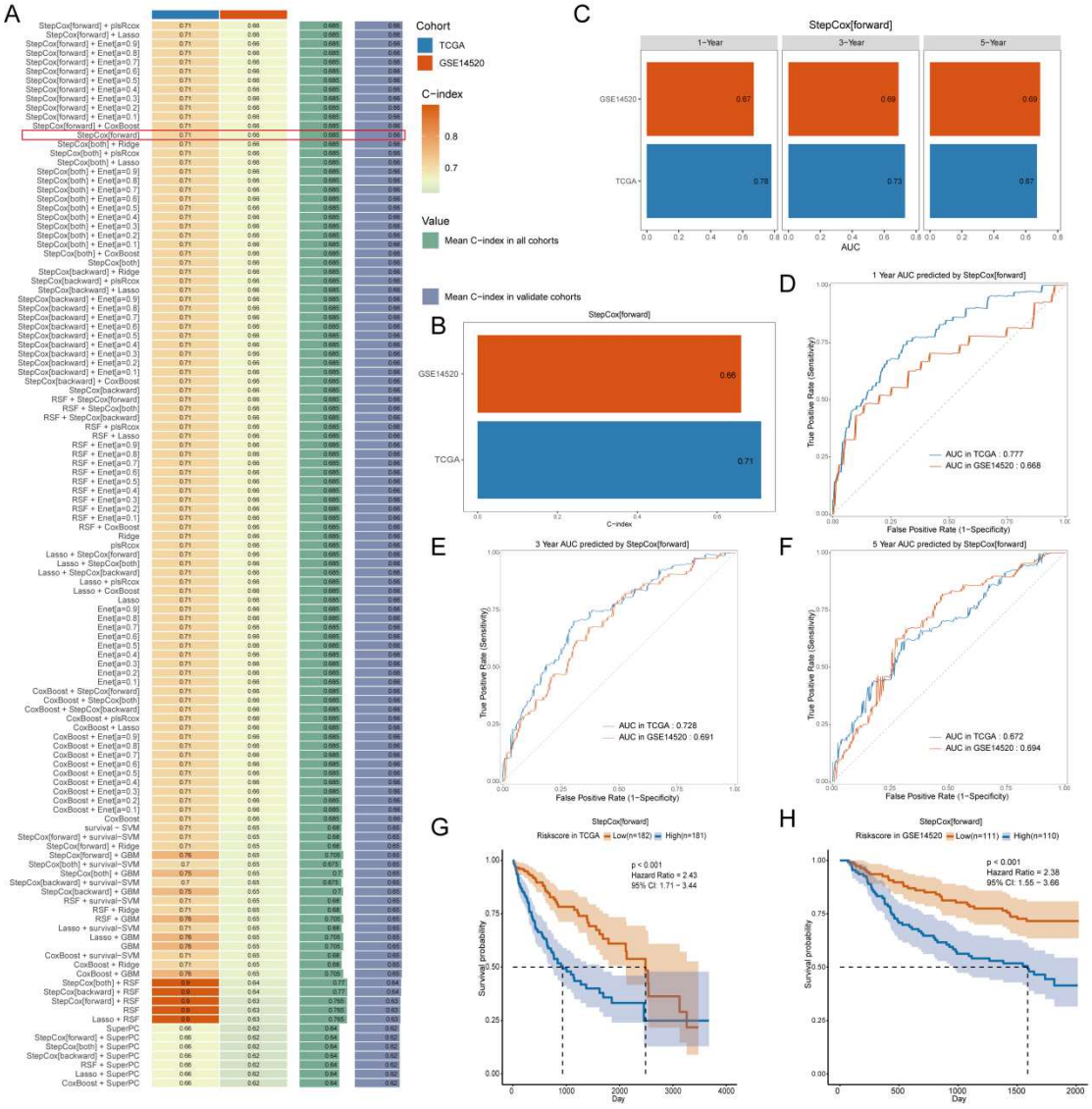


Figure 3. Model construction and validation.

(A) Predictive models constructed using multiple machine learning algorithms.

(B) C-index values of the StepCox[forward] model across different datasets.

(C) AUC values of the StepCox[forward] model for predicting 1-, 3-, and 5-year survival

outcomes in different datasets.

(D–F) ROC curves of the StepCox[forward] model for predicting 1-, 3-, and 5-year survival outcomes in each dataset.

(G, H) Kaplan–Meier survival curves showing differences in OS between high- and low-risk groups in the TCGA and GSE14520 datasets. Dashed lines indicate median OS. High: high-risk group defined based on EMT-related gene features. Low: low-risk group defined based on EMT-related gene risk scores.

Evaluation of the Risk Model

Patients in the TCGA and GSE14520 datasets were classified into high- and low-risk groups based on the median risk score. PCA was performed using the expression matrix of risk-related genes, and the results showed clear separation between high- and low-risk patients in multidimensional space (Figures 4A and 4B).

We further analyzed the distributions of gene expression levels, survival time, survival status, and risk scores. In the TCGA dataset, the mortality rate of patients in the high-risk group was significantly higher than that in the low-risk group (Figures 4C and 4E). A similar trend was observed in the GSE14520 dataset, where the high-risk group also exhibited markedly increased mortality (Figures 4D and 4F).

Overall, these results indicate that the risk model demonstrates good predictive performance and robustness across different datasets.

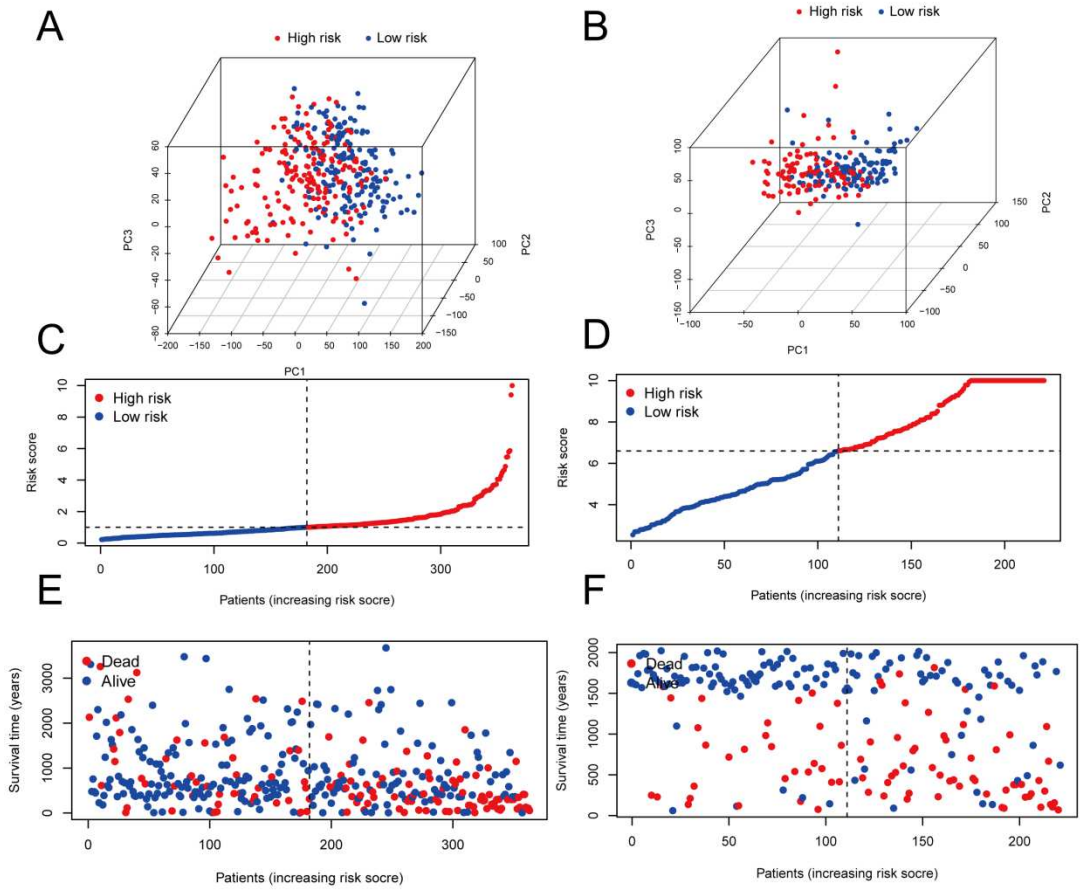


Figure 4. Performance of the prognostic model in risk subgroups of the TCGA and GSE14520 datasets.

(A, C, E) Three-dimensional PCA analysis, risk score distribution, and the relationship between survival time and survival status for risk subgroups in the TCGA dataset.

(B, D, F) Three-dimensional PCA analysis, risk score distribution, and the relationship between survival time and survival status for risk subgroups in the GSE14520 dataset.

Comparison with Published Prognostic Models

Through literature retrieval, we collected 43 recently published prognostic models for HCCand compiled the corresponding genes, regression coefficients (Coef), and PMID numbers for each model (Table S1). This summary provided a reference for evaluating our model and facilitated comparison with previous models.

The analysis showed that, compared with existing prognostic models, our constructed

model demonstrated good average C-index performance in both the training and validation cohorts (Figure 5A). In the heatmap, the HRs of individual genes were represented by different colors, with darker colors indicating higher HRs; statistical significance was indicated by symbols: ** for $p < 0.01$ and * for $p < 0.05$ (Figure 5B). Overall, the comprehensive analysis indicated that the model exhibited balanced and stable predictive performance in terms of both C-index and HR values, further validating its robustness and superiority in HCC prognosis prediction.

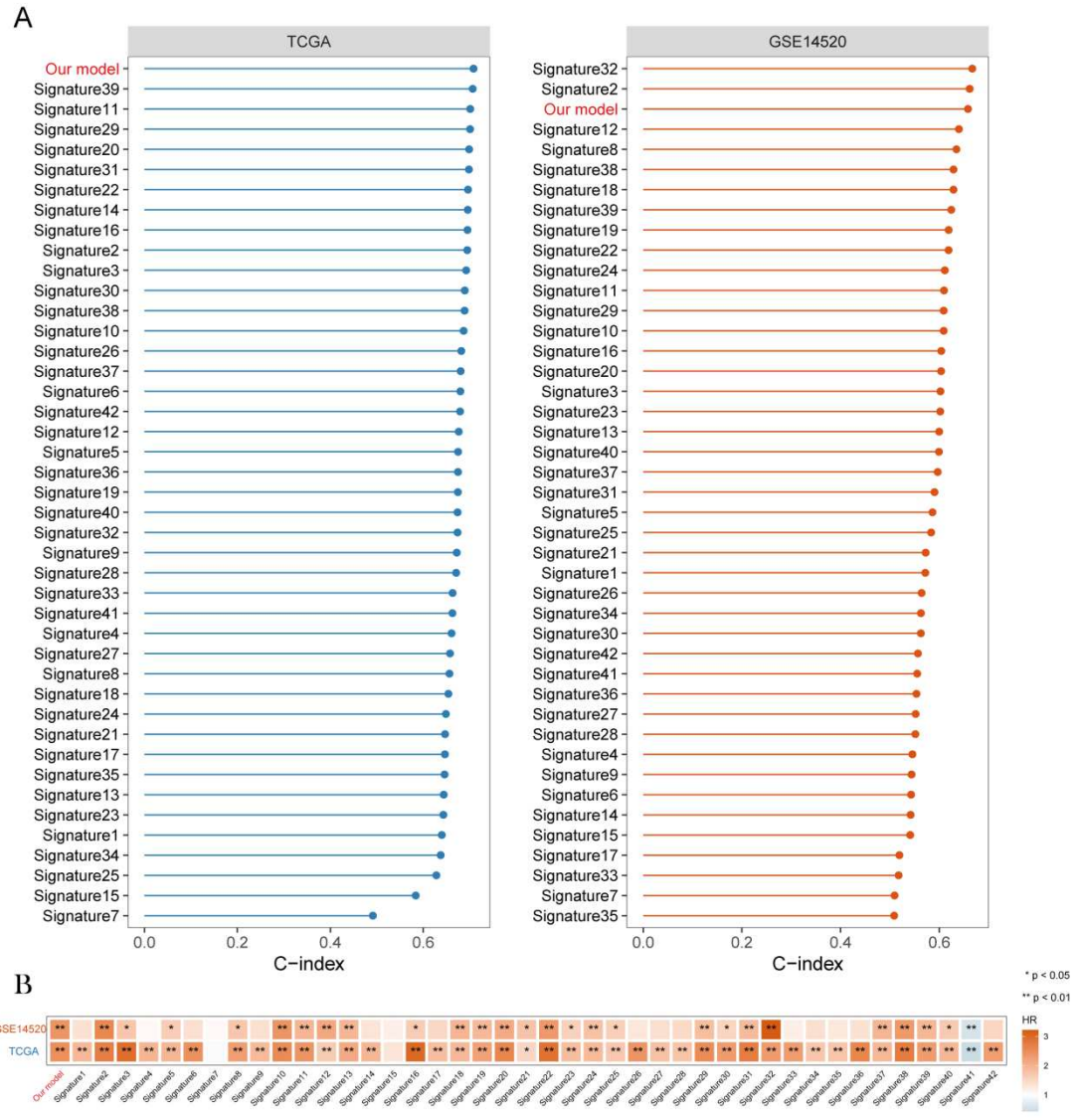


Figure 5. Comparison with published HCC-related prognostic gene models.

(A) Comparison of C-index values of different models across datasets.

(B) Heatmap of HRs of each model across datasets, with different colors representing

different HR values; darker colors indicate higher HRs. Statistical significance is indicated by symbols: * for $p < 0.05$ and ** for $p < 0.01$, representing statistical significance.

Construction and Validation of the Nomogram

To evaluate the independent prognostic value of the risk score and other clinical features in HCC, univariate and multivariate Cox regression analyses were performed in the TCGA cohort. Univariate analysis showed that the risk score, main tumor size, and ALT were significantly associated with patient prognosis ($p < 0.05$, Figure 6A). Multivariate analysis further indicated that only the risk score remained significant (Figure 6C), suggesting it can serve as an independent prognostic indicator for HCC.

Based on the significant factors from multivariate analysis in the TCGA cohort, a nomogram was constructed to predict 3-year overall survival (OS) (Figure 6E).

The same analyses were repeated in the external validation cohort GSE14520. Univariate analysis showed that TNM stage, alpha-fetoprotein levels, tumor size, cirrhosis status, and risk score were all significantly associated with prognosis ($p < 0.05$, Figure 6B); multivariate analysis identified the risk score and TNM stage as independent prognostic factors (Figure 6D). The nomogram constructed based on these significant factors (Figure 6F) further validated its accuracy and reliability in predicting prognosis for HCC patients.

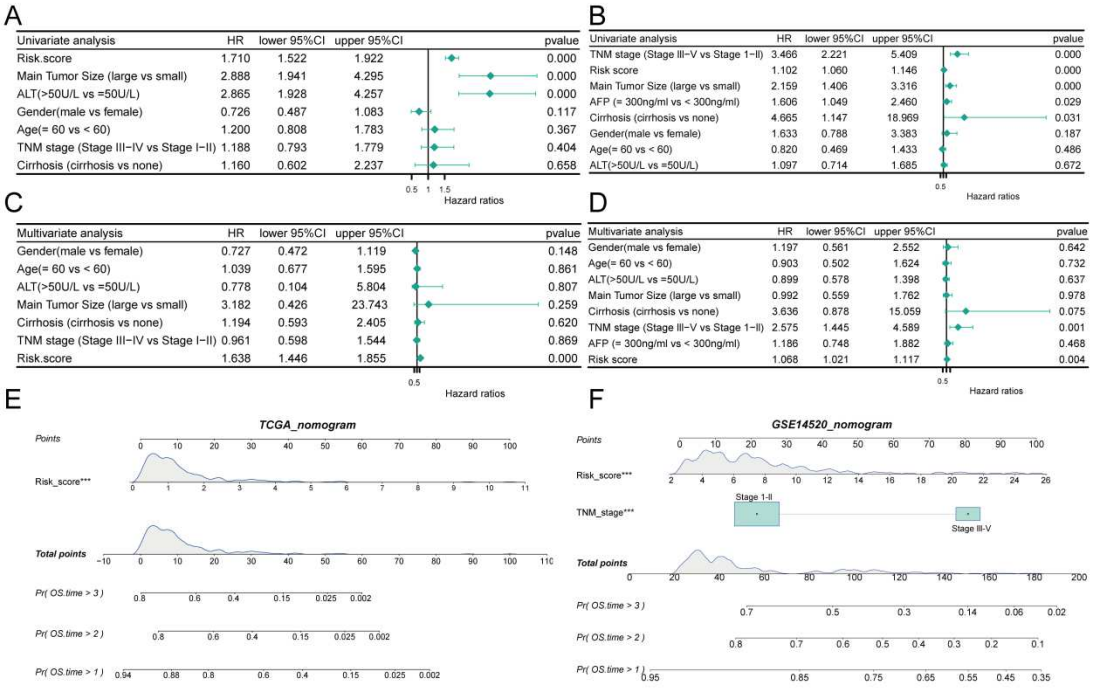


Figure 6. Construction and validation of the nomogram.

(A, C, E) Forest plots of univariate and multivariate Cox regression analyses for the risk score and clinical features in the TCGA dataset, and the nomogram constructed based on significant variables.

(B, D, F) Forest plots of univariate and multivariate Cox regression analyses for the risk score and clinical features in the GSE14520 dataset, and the nomogram constructed based on significant variables.

Functional Enrichment Analysis of DEGs between High- and Low-Risk Groups

To explore the potential biological functions of DEGs between high- and low-risk groups in the HCC prognostic model based on EMT-related genes, we performed KEGG and GO enrichment analyses, as well as GSEA.

KEGG pathway analysis showed that the DEGs were mainly enriched in metabolism and signal regulation-related pathways, including bile secretion, metabolism of xenobiotics by cytochrome P450, drug metabolism–cytochrome P450, retinol metabolism, steroid hormone biosynthesis, and chemical carcinogenesis–DNA

adducts, suggesting potential roles in drug metabolism, hormone synthesis, and carcinogenesis. Additionally, DEGs were involved in the PPAR signaling pathway, complement and coagulation cascades, and tyrosine metabolism, indicating possible participation in lipid metabolism regulation, immune responses, and amino acid metabolism (Figure 7A).

GO analysis revealed that, in terms of biological processes , DEGs were primarily involved in responses to exogenous stimuli, small molecule catabolic processes, and organic acid metabolism. Regarding cellular components , DEGs were significantly enriched in apical parts, apical plasma membrane, basement membrane, and collagen-containing extracellular matrix; for molecular functions , they were enriched in lipid transport, carboxylic acid transmembrane transport, iron ion binding, and oxidoreductase activity (Figure 7B). These findings suggest that DEGs may participate in metabolic regulation, transmembrane substance transport, and extracellular matrix remodeling.

GSEA analysis indicated that the high-risk group was significantly enriched in cell cycle and proliferation-related pathways, including E2F targets, G2/M checkpoint, mitotic spindle, and EMT pathways, with upregulation of MYC target genes, suggesting active tumor cell proliferation and invasive potential (Figure 7C). The low-risk group was mainly enriched in metabolism-related pathways such as lipogenesis, bile acid metabolism, fatty acid metabolism, and oxidative phosphorylation, indicating relatively active metabolic homeostasis (Figure 7D).

In immune-related GSEA, the high-risk group showed a state of immune overactivation, including enhanced B cell and $CD8^+$ T cell activity and imbalance between regulatory and conventional T cell functions (Figure 7E). The low-risk group was primarily enriched for gene sets related to resting or specific immune cell subsets, indicating relatively maintained immune homeostasis and lower activation levels (Figure 7F). Overall, the high-risk group exhibited characteristics of enhanced

proliferation and invasion, suppressed metabolic function, and immune overactivation, whereas the low-risk group maintained metabolic and immune homeostasis.

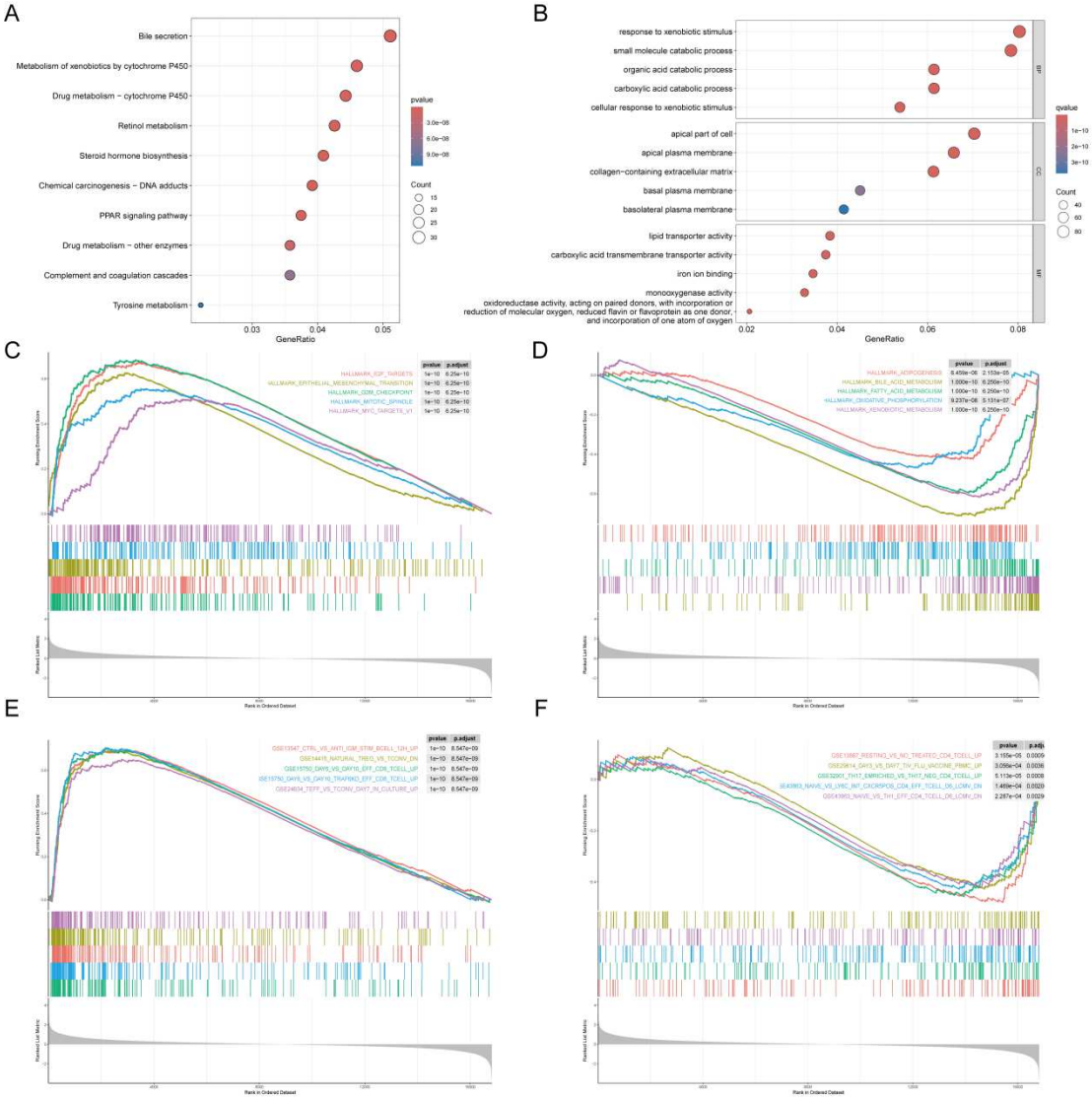


Figure 7. Functional enrichment analysis of differentially expressed genes (DEGs) between risk score subgroups.

(A) KEGG enrichment analysis of DEGs between high- and low-risk groups.

(B) GO enrichment analysis of DEGs between high- and low-risk groups.

(C, D) GSEA enrichment analysis of Hallmark gene sets.

(E, F) GSEA enrichment analysis of C7 gene sets.

Gene Mutation Analysis

To explore the mutation landscape, waterfall plots were generated for the top 20

frequently mutated genes in the TCGA dataset and in the high- and low-risk groups (Figures 8A, B), and the main mutation types were summarized. The results showed that missense mutations, single nucleotide polymorphisms , and C>T substitutions were the predominant mutation forms (Figures 8C, D).

Next, the co-occurrence and mutual exclusivity of the top 20 frequently mutated genes in the high- and low-risk groups were analyzed, and the median changes of various mutation types were calculated (Figures 8E, F). Green indicates gene co-occurrence, red indicates mutual exclusivity, and white indicates no significant association; statistical significance is marked with asterisks (*$p < 0.05$, **$p < 0.01$). In the high-risk group, TP53 and PCLO showed significant co-occurrence (green, $p < 0.05$), whereas in the low-risk group, CTNNB1 and ARID2 also displayed significant co-occurrence (green, $p < 0.01$). These findings suggest that high- and low-risk patients exhibit distinct mutation patterns in key genes, which may be associated with tumor biological behavior.

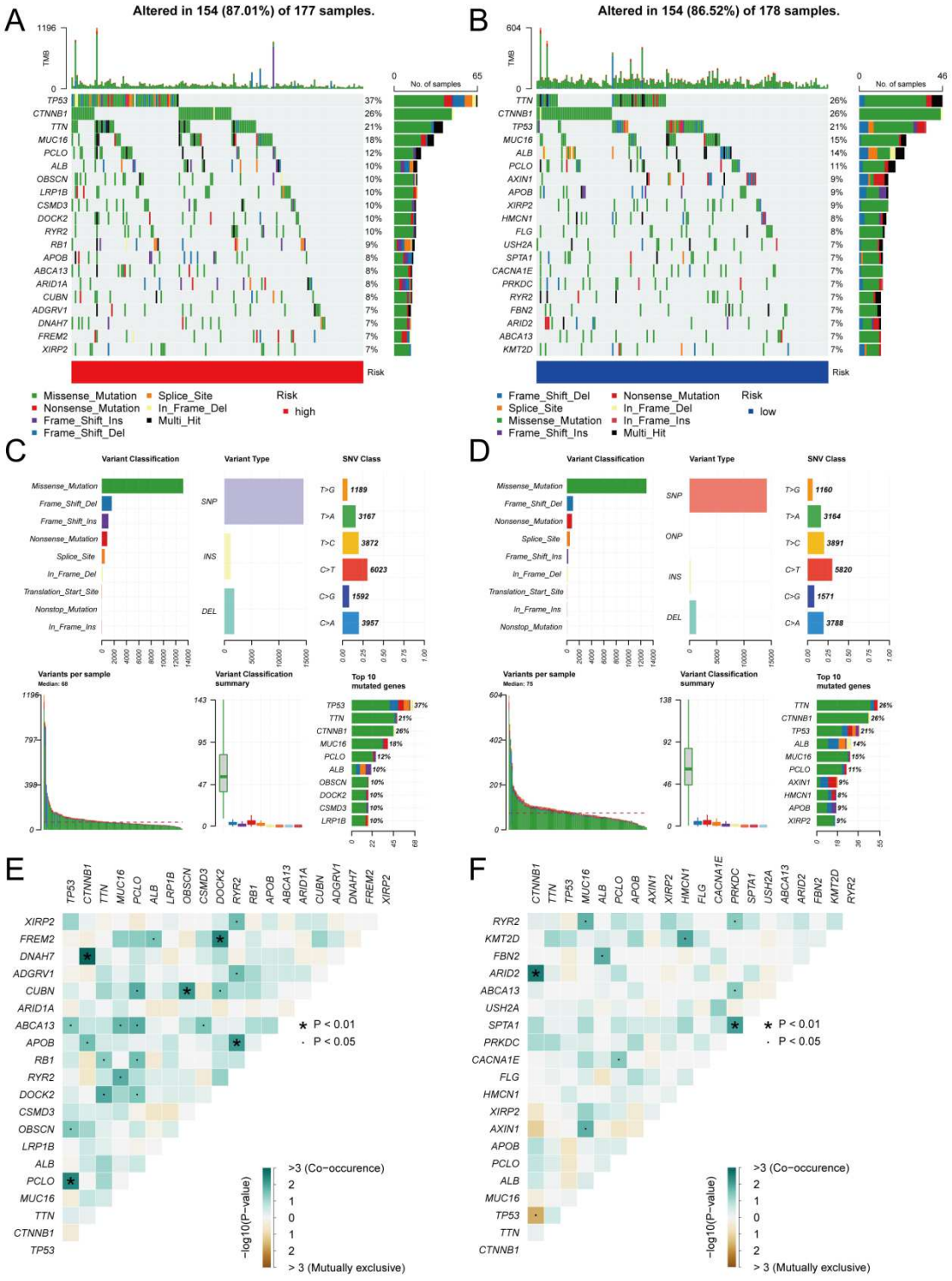


Figure 8. Tumor mutation analysis.

(A, B) Waterfall plots showing detailed mutations of the top 20 genes in the high-risk group (A) and low-risk group (B).

(C, D) Mutation profiles of different mutation types in the high-risk group (C) and low-risk group (D).

(E, F) Correlation heatmaps of the top 20 mutated genes in the high-risk group (E) and low-risk group (F).

Correlation Analysis between Risk Score and Tumor Immune Microenvironment

The heatmap showed that multiple immune cell infiltration levels were elevated in the high-risk group, with macrophages exhibiting the most pronounced increase (Figure 9A). To further explore the relationship between the risk score and immune cell infiltration, Spearman rank correlation coefficients (ρ) were used to assess the correlation between the risk score and 22 immune cell types, which is suitable for evaluating non-linear relationships.

The results demonstrated that M0 macrophages were significantly positively correlated with the risk score ($\rho = 0.36$, $p = 7.68\times10^{-13}$; Figure 9F), whereas the risk score was significantly negatively correlated with NK cells ($\rho = -0.23$, $p = 7.78\times10^{-6}$; Figure 9L) and resting mast cells ($\rho = -0.26$, $p = 2.56\times10^{-6}$; Figure 9N). The correlations of the remaining immune cells with the risk score were relatively weak (Figures9B–P).

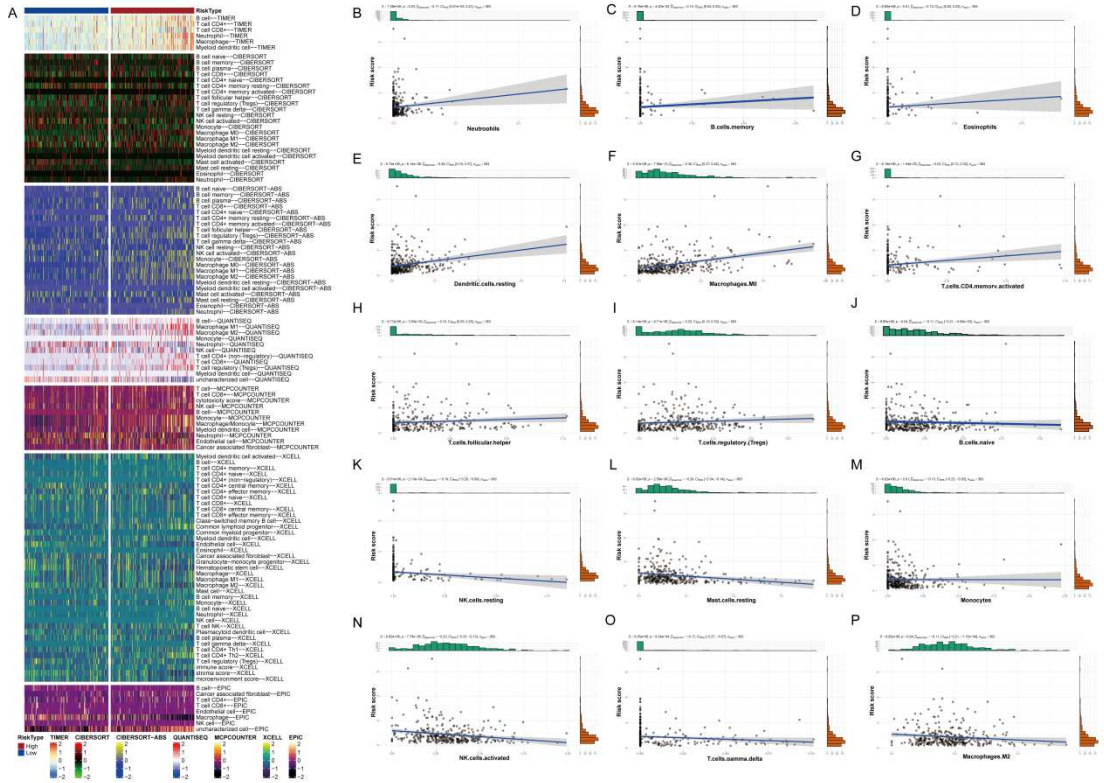


Figure 9. Correlation analysis of risk score with immune cell infiltration and tumor immune microenvironment features.

(A) Heatmap showing the infiltration levels of various immune cells across risk score-defined groups in the tumor microenvironment.

(B – P) Correlation plots between risk score and the abundance of different immune cell types.

Analysis of Risk Score with Immune Function, Immune Checkpoints, and Drug Sensitivity

Most immune function scores were significantly elevated in the high-risk group, suggesting a state of immune overactivation or inflammation. However, the low-risk group exhibited higher scores in Type I interferon response, Type II interferon response, and cytolytic activity, indicating potentially more effective anti-tumor immune activity (Figure 10A).

TIDE analysis showed that the high-risk group had significantly higher TIDE scores than the low-risk group (Figure 10B), suggesting stronger immune evasion in the high-risk patients.

Analysis of immune checkpoint gene expression identified 42 genes with significant differences between the high- and low-risk groups. Among them, IDO2 was upregulated in the low-risk group, while most other immune checkpoints were significantly upregulated in the high-risk group (Figure 10C).

Drug sensitivity analysis revealed significant differences in IC50 levels between the high- and low-risk groups (Figure 10D), with 111 drugs showing significant differences. Overall, the high-risk group exhibited lower IC50 values for most drugs, suggesting greater potential responsiveness to certain targeted therapies or chemotherapeutic agents.

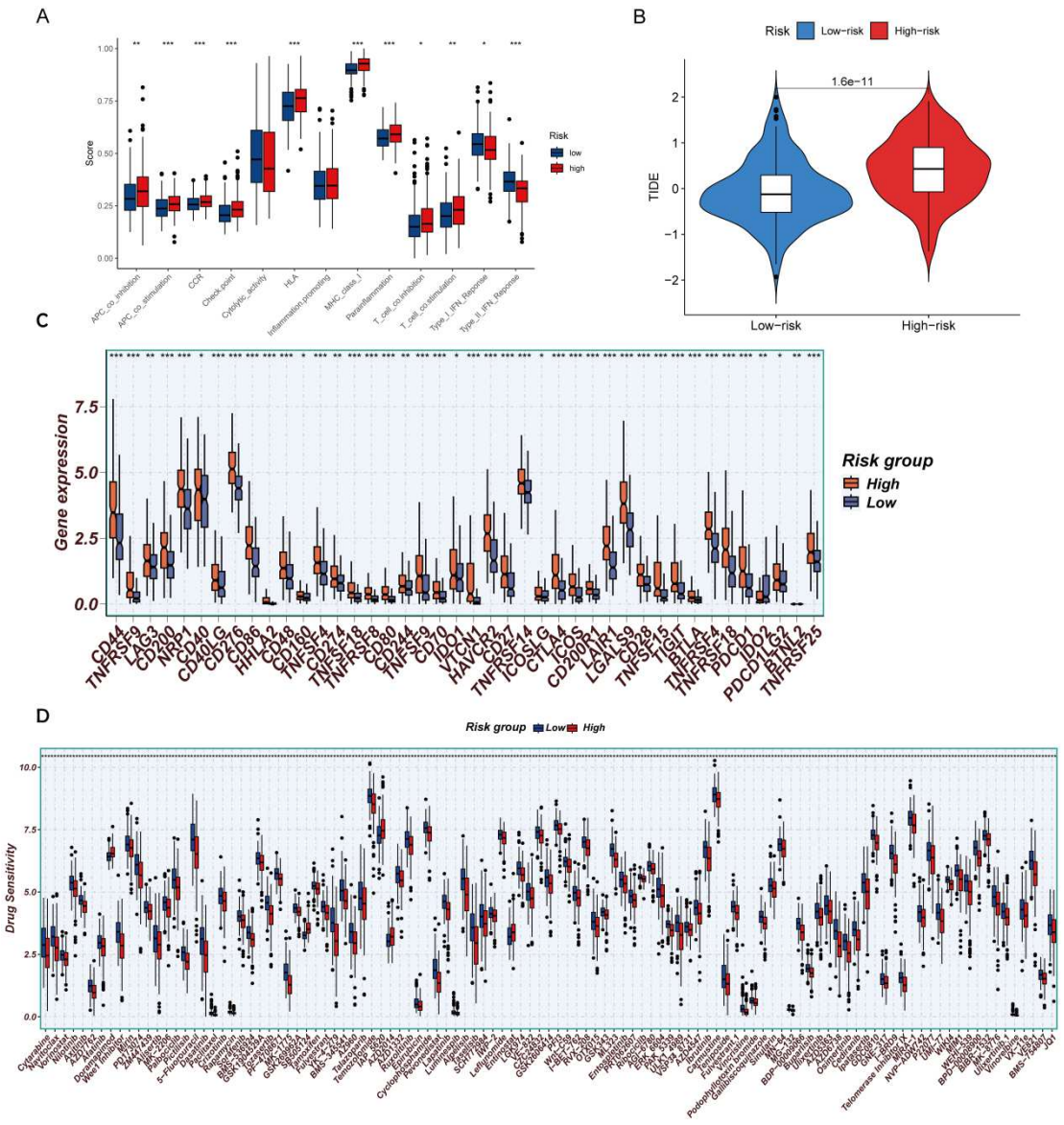


Figure 10. Association of Risk Score with Immune Function, Immune Evasion, Immune Checkpoint Expression, and Drug Sensitivity

(A) Comparison of immune function scores between different risk groups.

(B) Differences in TIDE scores between high- and low-risk groups, assessing potential immune evasion.

(C) Differential expression of immune checkpoint genes between high- and low-risk groups.

(D) Comparison of drug sensitivity (evaluated by IC50) to commonly used anticancer drugs between different risk groups.

Single-Cell Transcriptomic Analysis Reveals the Expression

Features of EMT-Related Prognostic Genes

At the single-cell transcriptome level, we further explored the cellular heterogeneity of the samples and the expression characteristics of EMT-related prognostic genes (Figure 11). UMAP clustering and cell annotation results indicated that the samples could be divided into multiple cell populations (Figure 11A). Differential expression analysis revealed significant transcriptional differences among these cell populations (Figure 11B). Subsequently, inferCNV analysis identified that cluster 11 exhibited tumor-associated copy number variation features (Figure 11C).

The bubble plot of canonical marker genes was highly consistent with known cell types, validating the accuracy of cell annotation (Figure 11D). In terms of spatial distribution, the proportion of different cell populations differed significantly between tumor core and border regions (Figure 11E).

Analysis of the expression patterns of the five EMT-related prognostic genes in the model showed heterogeneous distribution across different cell populations (Figures 11F, G). Among them, SPP1 was highly expressed in macrophages and also enriched to some extent in hepatocellular carcinoma cells and neutrophils, whereas the other genes did not show obvious cell type-specific expression.

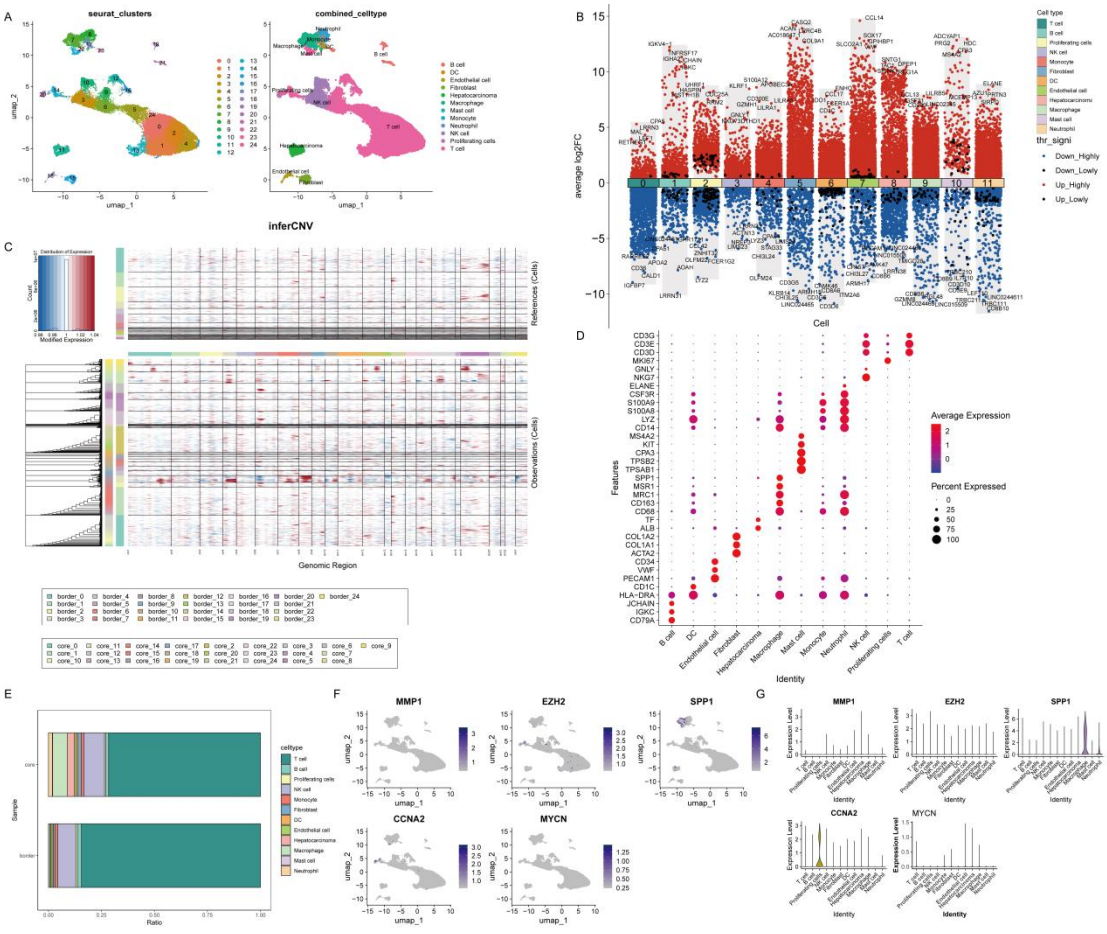


Figure 11. Single-cell transcriptomic analysis reveals tumor cell heterogeneity and expression features of EMT-related prognostic genes.

(A) UMAP plot of clustered and annotated cells showing the distribution of different cell types.

(B) Differential gene expression analysis among cell populations.

(C) Results of inferCNV analysis.

(D) Bubble plot of canonical marker genes for validating the accuracy of cell annotation.

(E) Comparison of the proportions of different cell types in the tumor core and border regions.

(F) UMAP expression distribution of the five EMT-related prognostic genes in single-cell data.

(G) Violin plot showing the expression of the five EMT-related prognostic genes across different cell populations.

CellChat Analysis

To explore potential cell‐cell communication involving SPP1, CellChat analysis was performed based on the single-cell data (Figure 12A, B). Exploratory analysis suggested interactions between macrophages and neutrophils (Figure 12C), and the SPP1 signaling pathway showed relatively high predicted communication activity (Figure 12D, E).

At the ligand‐receptor level, SPP1‐CD44 was among the interactions with relatively high predicted communication probability (Figure S1A). This predicted interaction was observed between several cell populations, including macrophages and neutrophils (Figure S1B). Bubble plot visualization showed that SPP1 was relatively highly expressed in macrophages, whereas CD44 was broadly expressed across multiple cell types, including neutrophils (Figure S1C).

CellChat sender‐receiver analysis indicated that macrophages were among the main predicted senders and mediators in the SPP1 signaling network, whereas neutrophils were among the predicted receivers (Figures S1D, E). These results are exploratory and were not functionally validated.

Group-wise comparison showed that predicted cell‐cell communication intensity was higher in the tumor core region than in the adjacent non-tumor region (Figure S2A‐C), with the SPP1 signaling pathway showing relatively higher predicted activity in the core region (Figure S2D). However, these findings are based on computational prediction and should be interpreted with caution.

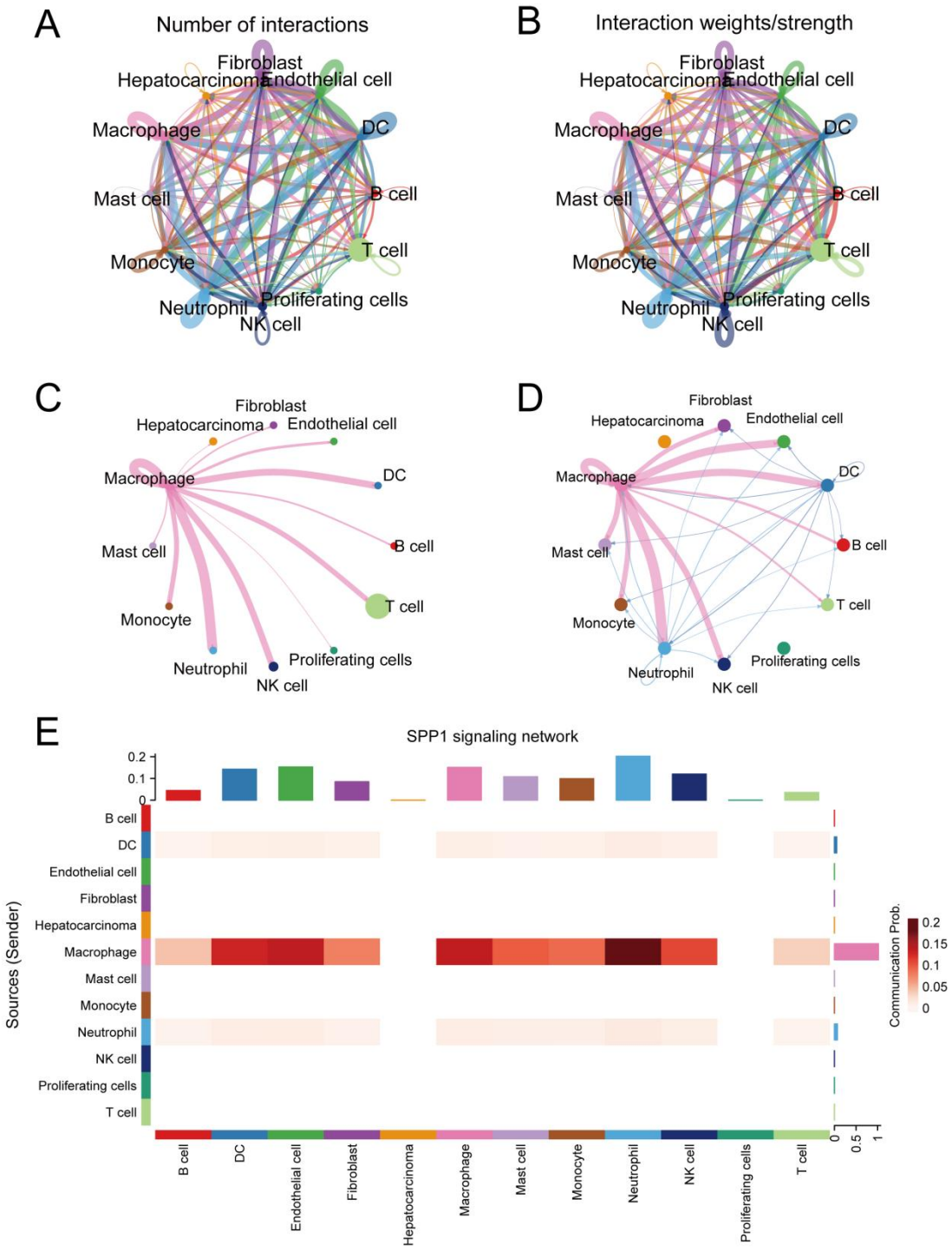


Figure 12. Exploratory CellChat analysis of cell－cell communication and SPP1－CD44 interactions.

(A, B) CellChat analysis workflow and overall cell–cell communication network based on single-cell transcriptomic data.

(C) Cell–cell communication relationships related to macrophages.

(D) Communication network of the SPP1 signaling pathway.

(E) Heatmap of the SPP1 signaling pathway.

Immunohistochemistry Results

Immunohistochemical analysis showed that SPP1 was markedly upregulated in hepatocellular carcinoma tissues (CA) compared with cancer-adjacent tissues (CAP) and non-cancer tissues (CAF), primarily localized in the cytoplasm of tumor cells and the intercellular matrix (Figure 13A). Quantitative analysis based on mean optical density (MOD) indicated that the average MOD in the CA group was significantly higher than that in the CAP and CAF groups ($P < 0.01$), whereas there was no significant difference between the CAP and CAF groups ($P > 0.05$) (Figure 13B).

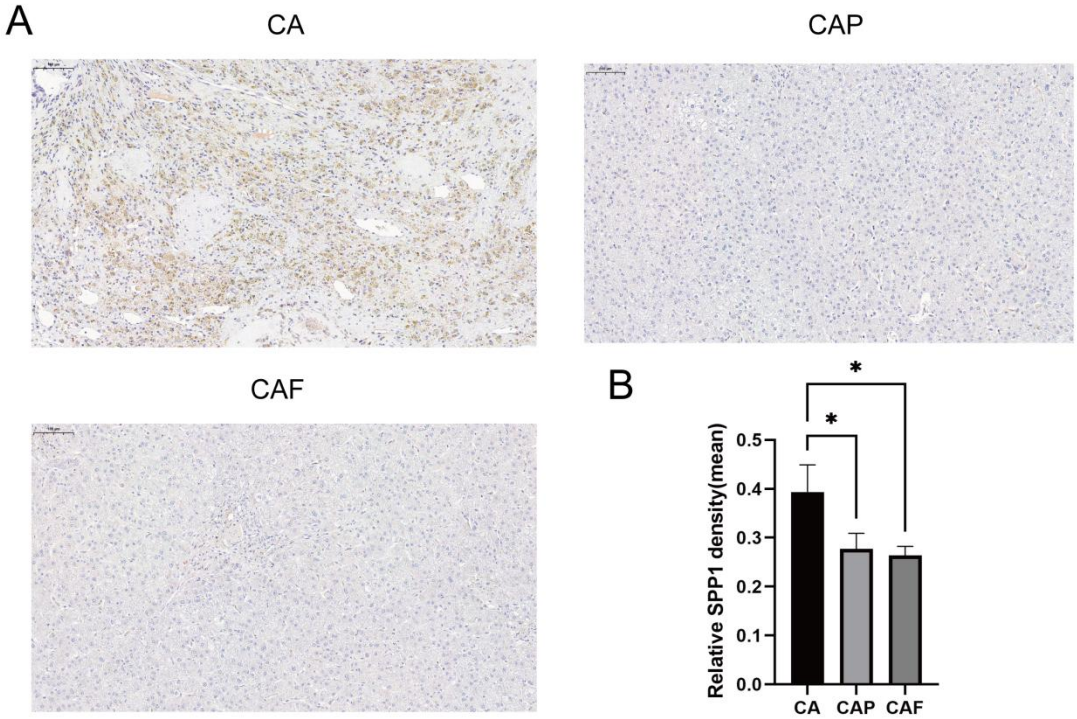


Figure 13. Immunohistochemical expression of SPP1 in liver tissues and quantitative analysis.

(A) Representative IHC images of CA (hepatocellular carcinoma), CAP (cancer-adjacent), and CAF (non-cancer) tissues (original magnification ×200). DAB staining (brown) indicates SPP1 expression.

(B) Quantitative analysis of SPP1 expression shown as mean ± SD (n = 3). Statistical comparisons were performed using one-way ANOVA followed by Tukey's post hoc test. ** indicates a significant difference compared with the CAP and CAF groups ($P < 0.01$), while no significant difference was observed between CAP and CAF ($P > 0.05$).

Discussion

In this study, we constructed an EMT-based five-gene prognostic model for hepatocellular carcinoma (HCC), comprising MMP1, EZH2, SPP1, CCNA2, and MYCN, and validated it in the TCGA training cohort and the GSE14520 validation cohort. The model showed modest predictive performance in terms of C-index and hazard ratios, and its performance was comparable to, but not clearly superior to, several previously published models. These findings suggest that the EMT-based five-gene signature may serve as a candidate tool for risk stratification, but its clinical utility requires further validation.

Functional enrichment and GSEA analyses revealed molecular differences between the high- and low-risk groups. The high-risk group was enriched in E2F target genes, G2/M checkpoint, MYC target genes, and EMT pathways, suggesting enhanced proliferative capacity and cell cycle dysregulation. The low-risk group showed higher activity in pathways related to lipogenesis, bile acid metabolism, fatty acid metabolism, oxidative phosphorylation, and xenobiotic metabolism, suggesting relatively preserved metabolic homeostasis. These observations are consistent with the notion that proliferative and metabolic programs are differentially regulated in HCC subgroups.

At the level of the immune microenvironment, the high-risk group generally exhibited increased infiltration of various immune cells, with M0 macrophages showing a significant positive correlation with risk scores, whereas NK cells and resting mast cells were negatively correlated. The high-risk group also displayed significantly higher TIDE scores, indicating a stronger potential for immune evasion, and immune checkpoint genes were overall upregulated in this group. In contrast, the low-risk group scored higher in Type I/II interferon responses and cytolytic activity, suggesting retention of more effective anti-tumor immune activity. Drug sensitivity analysis indicated that the high-risk group was more responsive to certain targeted or chemotherapeutic agents, providing potential guidance for individualized therapy.

These findings are consistent with previous reports that highly proliferative HCC is often accompanied by immune microenvironment disturbances [37,38].

Among the five genes in the model, matrix metalloproteinase-1 (MMP1) degrades the extracellular matrix and has been associated with tumor invasion and poor prognosis in HCC [39,40]. EZH2, a histone methyltransferase, is involved in tumor initiation and progression and has been linked to poor prognosis in HCC [41]. Cyclin A2 (CCNA2) regulates the G1/S phase transition of the cell cycle and has been identified as a key gene associated with poor prognosis in HCC [42]. MYCN is a member of the Myc family, and its amplification has been reported in various tumors; silencing MYCN inhibits liver cancer cell proliferation and induces cell death [43,44]. These genes collectively reflect proliferative and cell cycle-related programs as well as EMT-related features.

SPP1 (also known as osteopontin) is a secreted glycoprotein involved in cell adhesion, migration, and immune regulation. In HCC, high SPP1 expression has been associated with tumor progression and poor prognosis [45,46]. In our single-cell analysis, SPP1 was expressed in multiple cell types, with relatively higher expression in myeloid cells, including macrophages, mast cells, and neutrophils, and detectable expression in hepatocarcinoma cells. Other cell types, such as T cells, B cells, endothelial cells, and fibroblasts, showed relatively low SPP1 expression. In a preliminary IHC analysis, SPP1 protein was elevated in HCC tissues compared with adjacent and normal liver tissues, with positive signals observed in hepatocellular carcinoma cells and some stromal areas. The presence of SPP1 in tumor cells was consistent with the single-cell finding that hepatocarcinoma cells express SPP1. Because SPP1 is a secreted glycoprotein, IHC detects both intracellular and deposited protein, and the cellular source cannot be definitively assigned without double staining.

Exploratory CellChat analysis suggested potential SPP1 – CD44 communication between cell populations; however, these results were not functionally validated and

are therefore not further discussed as a mechanistic axis.

This study has several limitations. First, it primarily relies on public databases such as TCGA and GEO, lacking validation in large-scale prospective clinical cohorts; therefore, the clinical applicability of the model remains to be further confirmed. Second, no functional experiments were performed, and the role of SPP1 in HCC progression remains unknown. Third, the IHC cohort was small (n = 3), and double staining was not performed to confirm the cellular source of SPP1. Fourth, the prognostic model was constructed through multiple selection steps and may be overfitted; external validation and prospective studies are needed. Finally, this study did not fully account for the spatial heterogeneity of the tumor microenvironment or the potential roles of other immune cell subsets.

In summary, the EMT-based five-gene prognostic model constructed in this study showed modest predictive performance and was associated with proliferation and immune microenvironment features. Single-cell and IHC analyses showed that SPP1 is expressed in myeloid cells and hepatocarcinoma cells, and elevated in HCC tissues. These findings are descriptive and hypothesis-generating, and further functional studies are required to determine the biological role of SPP1 in HCC.

Conclusion

In this study, we constructed an EMT-based five-gene prognostic model for HCC, comprising MMP1, EZH2, SPP1, CCNA2, and MYCN, and validated it in the TCGA and GSE14520 cohorts. The model showed modest predictive performance for overall survival. Functional enrichment and GSEA analyses indicated that the high-risk group exhibited enhanced proliferative activity and cell cycle dysregulation, accompanied by immune microenvironment alterations.

Single-cell analysis showed that SPP1 was expressed in myeloid cells and hepatocarcinoma cells, and IHC showed elevated SPP1 protein in HCC tissues with positivity in tumor cells and stroma. Exploratory CellChat analysis suggested

potential SPP1 – CD44 communication, but this was not functionally validated. However, the IHC cohort was small, double staining was not performed, and no functional experiments were conducted. The cellular source and functional role of SPP1 in HCC progression remain to be determined.

Overall, this study provides a descriptive foundation for future studies on SPP1 in HCC. The findings are hypothesis-generating and do not establish a causal mechanism.

Abbreviation Section

Hepatocellular carcinoma	HCC
The Cancer Genome Atlas	TCGA
hepatitis B virus	HBV
tumor-associated macrophages	TAMs
Weighted Gene Co-expression Network Analysis	WGCNA
Protein – protein interaction	PPI
transarterial chemoembolization	TACE
Differentially expressed genes	DEGs
immune checkpoint inhibitors	ICIs
Tumor Immune Dysfunction and Exclusion	TIDE
Epithelial – mesenchymal transition	EMT
Overall survival	OS
Concordance Index	C-index

Area Under the Curve	AUC
Hazard ratio	HR
Principal Component Analysis	PCA
Gene Ontology	GO
Kyoto Encyclopedia of Genes and Genomes	KEGG
Gene Set Enrichment Analysis	GSEA
Molecular Signatures Database	MsigDB
Quality control	QC
Search Tool for the Retrieval of Interacting Genes	STRING
Tumor immune microenvironment	TME
Half maximal inhibitory concentration	IC50
Coefficients	Coef
pre-metastatic niche	PMN

Declarations

Ethical approval and Accordance Statement： All procedures involving human tissue samples were conducted in accordance with the ethical standards of the Institutional Ethics Committee of Guangxi University of Chinese Medicine (Approval No. GXTCMU-EC KS20250000-525) and with the 1964 Helsinki Declaration and its later amendments. All methods were carried out in accordance with relevant

guidelines and regulations.

Consent to participate： Written informed consent was obtained from all participants prior to sample collection.

Consent to publish： All authors consent to the publication of this manuscript.

Accordance Statement：All experimental procedures, data collection, and analyses were performed in accordance with relevant institutional, national, and international guidelines and regulations.

Availability of data and material: The datasets analyzed in this study were obtained from publicly available repositories. The primary transcriptomic and clinical data were downloaded from The Cancer Genome Atlas (TCGA-LIHC, https://www.cancer.gov/tcga) and the Gene Expression Omnibus (GEO) database, including GSE14520(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE14520) and GSE189903 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE189903).

The epithelial‑mesenchymal transition (EMT)-related gene list was obtained from the dbEMT database (http://dbemt.bioinfo-minzhao.org/). Protein‑protein interaction (PPI) data were retrieved from the STRING database (https://string-db.org/). Gene sets for functional enrichment analysis were downloaded from the Molecular Signatures Database (MSigDB, https://www.gsea-msigdb.org/gsea/msigdb). The Tumor Immune Dysfunction and Exclusion (TIDE) database (http://tide.dfci.harvard.edu/) was used to predict immunotherapy response.

All processed data generated during this study are included in this published article

and its supplementary materials. Further inquiries can be directed to the corresponding author.

Competing interests: The authors declare no conflict of interest.

Funding: This research was funded by the National Natural Science Foundation of China grant number [No. 82460957], the Guangxi Natural Science Foundation Project [No. 2024GXNSFDA010005], and Innovation Project of

Guangxi Graduate Education of GXUCM [No. YCBXJ2025014].

Authors' contributions: Xinyuan Chen and Lu Lu contributed equally to this work. Data curation, Yu Zhang and Yuanqin Du; Software, Jiazhi Li; Validation, Yanhua Lai; Visualization, Chengting Wu; Methodology, investigation, Writing—original draft, Xinyuan Chen and Lu Lu; Writing—review & editing, Jingjing Huang and Hongna Huang. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

This study was greatly supported by the research team of Professor Jingjing Huang, to whom we express our sincere gratitude. We also thank The Cancer Genome Atlas (TCGA) and the GEO database for providing public data resources that laid an important foundation for this work. Special thanks go to Professor Jingjing Huang for her dedicated guidance in statistical analysis and bioinformatics. We are also grateful to the Guangxi Key Laboratory of Translational Medicine for Treating High-Incidence Infectious Diseases with Integrative Medicine for their technical assistance, and to the reviewers for their valuable comments and suggestions, which significantly improved the quality of this manuscript.

References:

1. Rumgay H, Arnold M, Ferlay J, Lesi O, Cabasag CJ, Vignat J, Laversanne M, McGlynn KA, Soerjomataram I. Global burden of primary liver cancer in 2020 and predictions to 2040. J Hepatol. 2022

Dec;77(6):1598-1606. doi: 10.1016/j.jhep.2022.08.021. Epub 2022 Oct 5. PMID: 36208844; PMCID: PMC9670241.

2. Llovet JM, Kelley RK, Villanueva A, Singal AG, Pikarsky E, Roayaie S, Lencioni R, Koike K, Zucman-Rossi J, Finn RS. Hepatocellular carcinoma. Nat Rev Dis Primers. 2021 Jan 21;7(1):6. doi: 10.1038/s41572-020-00240-3. Erratum in: Nat Rev Dis Primers. 2024 Feb 12;10(1):10. doi: 10.1038/s41572-024-00500-6. PMID: 33479224.

3. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA Cancer J Clin. 2024 Jan-Feb;74(1):12-49. doi: 10.3322/caac.21820. Epub 2024 Jan 17. Erratum in: CA Cancer J Clin. 2024 Mar-Apr;74(2):203. doi: 10.3322/caac.21830. PMID: 38230766.

4. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021 May;71(3):209-249. doi: 10.3322/caac.21660. Epub 2021 Feb 4. PMID: 33538338.

5. Schlachterman A, Craft WW Jr, Hilgenfeldt E, Mitra A, Cabrera R. Current and future treatments for hepatocellular carcinoma. World J Gastroenterol. 2015 Jul 28;21(28):8478-91. doi: 10.3748/wjg.v21.i28.8478. PMID: 26229392; PMCID: PMC4515831.

6. Calderaro J, Petitprez F, Becht E, Laurent A, Hirsch TZ, Rousseau B, Luciani A, Amaddeo G, Derman J, Charpy C, Zucman-Rossi J, Fridman WH, Sautès-Fridman C. Intra-tumoral tertiary lymphoid structures are associated with a low risk of early recurrence of hepatocellular carcinoma. J Hepatol. 2019 Jan;70(1):58-65. doi: 10.1016/j.jhep.2018.09.003. Epub 2018 Sep 11. PMID: 30213589.

7. Cho JY, Han HS, Choi Y, Yoon YS, Kim S, Choi JK, Jang JS, Kwon SU, Kim H. Association of Remnant Liver Ischemia With Early Recurrence and Poor Survival After Liver Resection in Patients With Hepatocellular Carcinoma. JAMA Surg. 2017 Apr 1;152(4):386-392. doi: 10.1001/jamasurg.2016.5040. PMID: 28052154; PMCID: PMC5470428.

8. Yoh T, Seo S, Taura K, Iguchi K, Ogiso S, Fukumitsu K, Ishii T, Kaido T, Uemoto S. Surgery for Recurrent Hepatocellular Carcinoma: Achieving Long-term Survival. Ann Surg. 2021 Apr 1;273(4):792-799. doi: 10.1097/SLA.0000000000003358. PMID: 31058698.

9. Anstee QM, Reeves HL, Kotsiliti E, Govaere O, Heikenwalder M. From NASH to HCC: current concepts and future challenges. Nat Rev Gastroenterol Hepatol. 2019 Jul;16(7):411-428. doi: 10.1038/s41575-019-0145-7. PMID: 31028350.

10. Nevola R, Ruocco R, Criscuolo L, Villani A, Alfano M, Beccia D, Imbriani S, Claar E, Cozzolino D, Sasso FC, Marrone A, Adinolfi LE, Rinaldi L. Predictors of early and late hepatocellular carcinoma recurrence. World J Gastroenterol. 2023 Feb 28;29(8):1243-1260. doi: 10.3748/wjg.v29.i8.1243. PMID: 36925456; PMCID: PMC10011963.

11. Raoul JL, Forner A, Bolondi L, Cheung TT, Kloeckner R, de Baere T. Updated use of TACE for hepatocellular carcinoma treatment: How and when to use it based on clinical evidence. Cancer Treat Rev. 2019 Jan;72:28-36. doi: 10.1016/j.ctrv.2018.11.002. Epub 2018 Nov 12. PMID: 30447470.

12. Sachdeva M, Chawla YK, Arora SK. Immunology of hepatocellular carcinoma. World J Hepatol. 2015 Aug 18;7(17):2080-90. doi: 10.4254/wjh.v7.i17.2080. PMID: 26301050; PMCID: PMC4539401.

13. Fu J, Xu D, Liu Z, Shi M, Zhao P, Fu B, Zhang Z, Yang H, Zhang H, Zhou C, Yao J, Jin L, Wang H, Yang Y, Fu YX, Wang FS. Increased regulatory T cells correlate with CD8 T-cell impairment and poor survival in hepatocellular carcinoma patients. Gastroenterology. 2007 Jun;132(7):2328-39. doi: 10.1053/j.gastro.2007.03.102. Epub 2007 Apr 14. PMID: 17570208.

14. Gabrilovich DI, Nagaraj S. Myeloid-derived suppressor cells as regulators of the immune system. Nat Rev Immunol. 2009 Mar;9(3):162-74. doi: 10.1038/nri2506. PMID: 19197294; PMCID: PMC2828349.

15. Yin Y, Feng W, Chen J, Chen X, Wang G, Wang S, Xu X, Nie Y, Fan D, Wu K, Xia L. Immunosuppressive tumor microenvironment in the progression, metastasis, and therapy of hepatocellular carcinoma: from bench to bedside. Exp Hematol Oncol. 2024 Aug 1;13(1):72. doi: 10.1186/s40164-024-00539-x. PMID: 39085965; PMCID: PMC11292955.

16. Pan J, Zhang M, Rao D, Ma J, Shen X, Qin H, Gan K, Lin J, Huang Y, Sang C, Zhang J, Ma J, Wu Y, Tang Z, Gao D, Gao Q, Yang L, Fan J. CAD manipulates tumor intrinsic DHO/UBE4B/NF-κB pathway

and fuels macrophage cross-talk, promoting HCC metastasis. Hepatology. 2025 Mar 12. doi: 10.1097/HEP.0000000000001304. Epub ahead of print. PMID: 40073276.

17. Brabletz T, Kalluri R, Nieto MA, Weinberg RA. EMT in cancer. Nat Rev Cancer. 2018 Feb;18(2):128-134. doi: 10.1038/nrc.2017.118. Epub 2018 Jan 12. PMID: 29326430.

18. Nieto MA, Huang RY, Jackson RA, Thiery JP. EMT: 2016. Cell. 2016 Jun 30;166(1):21-45. doi: 10.1016/j.cell.2016.06.028. PMID: 27368099.

19. Saitoh M. Involvement of partial EMT in cancer progression. J Biochem. 2018 Oct 1;164(4):257-264. doi: 10.1093/jb/mvy047. PMID: 29726955.

20. Diepenbruck M, Christofori G. Epithelial-mesenchymal transition (EMT) and metastasis: yes, no, maybe? Curr Opin Cell Biol. 2016 Dec;43:7-13. doi: 10.1016/j.ceb.2016.06.002. Epub 2016 Jun 29. PMID: 27371787.

21. Yeung KT, Yang J. Epithelial-mesenchymal transition in tumor metastasis. Mol Oncol. 2017 Jan;11(1):28-39. doi: 10.1002/1878-0261.12017. Epub 2016 Dec 9. PMID: 28085222; PMCID: PMC5242415.

22. Bracken CP, Goodall GJ. The many regulators of epithelial-mesenchymal transition. Nat Rev Mol Cell Biol. 2022 Feb;23(2):89-90. doi: 10.1038/s41580-021-00442-x. PMID: 34887545.

23. Dongre A, Weinberg RA. New insights into the mechanisms of epithelial-mesenchymal transition and implications for cancer. Nat Rev Mol Cell Biol. 2019 Feb;20(2):69-84. doi: 10.1038/s41580-018-0080-4. PMID: 30459476.

24. Fung SW, Cheung PF, Yip CW, Ng LW, Cheung TT, Chong CC, Lee C, Lai PB, Chan AW, Tsao GS, Wong CH, Chan SL, Lo KW, Cheung ST. The ATP-binding cassette transporter ABCF1 is a hepatic oncofetal protein that promotes chemoresistance, EMT and cancer stemness in hepatocellular carcinoma. Cancer Lett. 2019 Aug 10;457:98-109. doi: 10.1016/j.canlet.2019.05.010. Epub 2019 May 14. PMID: 31100412.

25. Han J, Meng J, Chen S, Wang X, Yin S, Zhang Q, Liu H, Qin R, Li Z, Zhong W, Zhang C, Zhang H, Tang Y, Lin T, Gao W, Zhang X, Yang L, Liu Y, Zhou HG, Sun T, Yang C. YY1 Complex Promotes Quaking Expression via Super-Enhancer Binding during EMT of Hepatocellular Carcinoma. Cancer Res. 2019 Apr 1;79(7):1451-1464. doi: 10.1158/0008-5472.CAN-18-2238. Epub 2019 Feb 13. Erratum in: Cancer Res. 2022 Dec 16;82(24):4694. doi: 10.1158/0008-5472.CAN-22-3444. PMID: 30760518.

26. Mittal V. Epithelial Mesenchymal Transition in Tumor Metastasis. Annu Rev Pathol. 2018 Jan 24;13:395-412. doi: 10.1146/annurev-pathol-020117-043854. PMID: 29414248.

27. Shibue T, Weinberg RA. EMT, CSCs, and drug resistance: the mechanistic link and clinical implications. Nat Rev Clin Oncol. 2017 Oct;14(10):611-629. doi: 10.1038/nrclinonc.2017.44. Epub 2017 Apr 11. PMID: 28397828; PMCID: PMC5720366.

28. Wang H, Xu H, Ma F, Zhan M, Yang X, Hua S, Li W, Li Y, Lu L. Zinc finger protein 703 induces EMT and sorafenib resistance in hepatocellular carcinoma by transactivating CLDN4 expression. Cell Death Dis. 2020 Apr 8;11(4):225. doi: 10.1038/s41419-020-2422-3. PMID: 32269215; PMCID: PMC7142083.

29. Zhao M, Liu Y, Zheng C, Qu H. dbEMT 2.0: An updated database for epithelial-mesenchymal transition genes with experimentally verified information and precalculated regulation information for cancer metastasis. J Genet Genomics. 2019 Dec 20;46(12):595-597. doi: 10.1016/j.jgg.2019.11.010. Epub 2019 Dec 21. PMID: 31941584.

30. Zhao M, Kong L, Liu Y, Qu H. dbEMT: an epithelial-mesenchymal transition associated gene resource. Sci Rep. 2015 Jun 23;5:11459. doi: 10.1038/srep11459. PMID: 26099468; PMCID: PMC4477208.

31. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008 Dec 29;9:559. doi: 10.1186/1471-2105-9-559. PMID: 19114008; PMCID: PMC2631488.

32. Liu H, Zhang W, Zhang Y, Adegboro AA, Fasoranti DO, Dai L, Pan Z, Liu H, Xiong Y, Li W, Peng K, Wanggou S, Li X. Mime: A flexible machine-learning framework to construct and visualize models for clinical characteristics prediction and feature selection. Comput Struct Biotechnol J. 2024 Jun 29;23:2798-2810. doi: 10.1016/j.csbj.2024.06.035. PMID: 39055398; PMCID: PMC11269309.

33. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018 Jun;36(5):411-420. doi: 10.1038/nbt.4096. Epub 2018 Apr 2. PMID: 29608179; PMCID: PMC6700744.

34. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, Ginhoux F, Newell EW. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2018 Dec 3. doi: 10.1038/nbt.4314. Epub ahead of print. PMID: 30531897.

35. Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan CH, Myung P, Plikus MV, Nie Q. Inference and analysis of cell-cell communication using CellChat. Nat Commun. 2021 Feb 17;12(1):1088. doi: 10.1038/s41467-021-21246-9. PMID: 33597522; PMCID: PMC7889871.

36. Bardou P, Mariette J, Escudié F, Djemiel C, Klopp C. jvenn: an interactive Venn diagram viewer. BMC Bioinformatics. 2014 Aug 29;15(1):293. doi: 10.1186/1471-2105-15-293. PMID: 25176396; PMCID: PMC4261873.

37. Xia Y, Brown ZJ, Huang H, Tsung A. Metabolic reprogramming of immune cells: Shaping the tumor microenvironment in hepatocellular carcinoma. Cancer Med. 2021 Sep;10(18):6374-6383. doi: 10.1002/cam4.4177. Epub 2021 Aug 13. PMID: 34390203; PMCID: PMC8446566.

38. Chen C, Wang Z, Ding Y, Qin Y. Tumor microenvironment-mediated immune evasion in hepatocellular carcinoma. Front Immunol. 2023 Feb 10;14:1133308. doi: 10.3389/fimmu.2023.1133308. PMID: 36845131; PMCID: PMC9950271.

39. Scheau C, Badarau IA, Costache R, Caruntu C, Mihai GL, Didilescu AC, Constantin C, Neagu M. The Role of Matrix Metalloproteinases in the Epithelial-Mesenchymal Transition of Hepatocellular Carcinoma. Anal Cell Pathol (Amst). 2019 Nov 26;2019:9423907. doi: 10.1155/2019/9423907. PMID: 31886121; PMCID: PMC6899323.

40. Dai L, Mugaanyi J, Cai X, Dong M, Lu C, Lu C. Comprehensive bioinformatic analysis of MMP1 in hepatocellular carcinoma and establishment of relevant prognostic model. Sci Rep. 2022 Aug 10;12(1):13639. doi: 10.1038/s41598-022-17954-x. PMID: 35948625; PMCID: PMC9365786.

41. Bai B, Liu Y, Fu XM, Qin HY, Li GK, Wang HC, Sun SL. Dysregulation of EZH2/miR-138-5p Axis Contributes to Radiosensitivity in Hepatocellular Carcinoma Cell by Downregulating Hypoxia-Inducible Factor 1 Alpha (HIF-1α). Oxid Med Cell Longev. 2022 Aug 29;2022:7608712. doi: 10.1155/2022/7608712. PMID: 36071871; PMCID: PMC9444475.

42. Wang T, Li W, Wu Y, You L, Zheng C, Zhang J, Qu L, Sun X. Construction of a prognostic model based on disulfidptosis-related genes and identification of CCNA2 as a novel biomarker for hepatocellular carcinoma. Biol Direct. 2024 Dec 19;19(1):128. doi: 10.1186/s13062-024-00569-9. PMID: 39695705; PMCID: PMC11656849.

43. Qin XY, Suzuki H, Honda M, Okada H, Kaneko S, Inoue I, Ebisui E, Hashimoto K, Carninci P, Kanki K, Tatsukawa H, Ishibashi N, Masaki T, Matsuura T, Kagechika H, Toriguchi K, Hatano E, Shirakami Y, Shiota G, Shimizu M, Moriwaki H, Kojima S. Prevention of hepatocellular carcinoma by targeting MYCN-positive liver cancer stem cells with acyclic retinoid. Proc Natl Acad Sci U S A. 2018 May 8;115(19):4969-4974. doi: 10.1073/pnas.1802279115. Epub 2018 Apr 23. PMID: 29686061; PMCID: PMC5949003.

44. Ambrosio S, Amente S, Saccà CD, Capasso M, Calogero RA, Lania L, Majello B. LSD1 mediates MYCN control of epithelial-mesenchymal transition through silencing of metastatic suppressor NDRG1 gene. Oncotarget. 2017 Jan 17;8(3):3854-3869. doi: 10.18632/oncotarget.12924. PMID: 27894074; PMCID: PMC5354800.

45. Dong Q, Zhu X, Dai C, Zhang X, Gao X, Wei J, Sheng Y, Zheng Y, Yu J, Xie L, Qin Y, Qiao P, Zhou C, Yu X, Jia H, Ren N, Zhou H, Ye Q, Qin L. Osteopontin promotes epithelial-mesenchymal transition of hepatocellular carcinoma through regulating vimentin. Oncotarget. 2016 Mar 15;7(11):12997-3012. doi: 10.18632/oncotarget.7016. PMID: 26824421; PMCID: PMC4914337.

46. Liu J, Zhang Y, Ran W, Yang L, Zhang W, Liu Z. Mechanisms of tumor heterogeneity in TACE-resistant liver cancer: Insights from single-cell and whole-exome sequencing. Hepatol Commun. 2025 Jun 9;9(7):e0705. doi: 10.1097/HC9.0000000000000705. PMID: 40489756; PMCID: PMC12153283.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.zip