scientific reports

Check for updates

OPEN

Development of a MVI associated HCC prognostic model through single cell transcriptomic analysis and 101 machine learning algorithms

Jiayi Zhang[1,3], Zheng Zhang[1,3], Chenqing Yang[2,3], Qingguang Liu[1]✉ & Tao Song[1]✉

Hepatocellular carcinoma (HCC) is an exceedingly aggressive form of cancer that often carries a poor prognosis, especially when it is complicated by the presence of microvascular invasion (MVI). Identifying patients at high risk of MVI is crucial for personalized treatment strategies. Utilizing the single-cell RNA-sequencing dataset (GSE242889) of HCC, we identified malignant cell subtypes associated with microvascular invasion (MVI), in conjunction with the TCGA dataset, selected a set of MVI-related genes (MRGs). We developed an optimal prognostic model comprising 11 genes (NOP16, YIPF1, HMMR, NDC80, DYNLL1, CDC34, NLN, KHDRBS3, MED8, SLC35G2, RAB3B) based on MVI-related signature genes by integrating single-cell transcriptomic analysis with 101 machine learning algorithms. This model is meticulously crafted to forecast the prognosis of individuals afflicted with hepatocellular carcinoma (HCC). Additionally, we affirmed the predictive precision and superiority of our model through a meta-analysis against existing HCC models. Furthermore, we explored the differences between high- and low-risk groups through mutation and immune infiltration analyses. Lastly, we investigated immunotherapy responses and drug sensitivities between risk groups, providing novel therapeutic insights for liver cancer.

Keywords Hepatocellular carcinoma, Microvascular infiltration, Prognostic prediction model, Machine learning

Liver cancer is the sixth most commonly diagnosed cancer and third most common cause of cancer-related deaths globally. Hepatocellular carcinoma (HCC), which accounts for more than 80% of liver cancers, is among the top three causes of cancer-related deaths in 46 countries and among the top five in 90 countries[1]. For patients with HCC who undergo curative resection, vascular invasion (VI) is a significant risk factor for early recurrence and poor prognosis. Microvascular infiltration (MVI), in particular, is an early indicator of vascular invasion and metastasis in HCC, and it is crucial for assessing the risk of recurrence and metastasis in HCC patients[2]. Hence, a thorough investigation into the mechanisms behind the occurrence of MVI is imperative. Such research will provide crucial insights that can inform and enhance clinical treatment strategies and patient management approaches for those with hepatocellular carcinoma.

Existing research indicates a strong correlation between the invasive and metastatic behaviors of hepatocellular carcinoma (HCC) and the surrounding tumor microenvironment (TME). This intricate relationship highlights the significance of the TME in the progression of HCC[3]. Currently, the complexity of the tumor microenvironment (TME) in hepatocellular carcinoma (HCC) is now well-established, with a diverse array of stromal cells, immune cells, and cancerous elements interacting to shape this critical landscape. The resulting variability within and between tumors creates substantial obstacles for the precise targeting of HCC, emphasizing the need to address these heterogeneities in order to refine and personalize therapeutic approaches[4]. Thus, our research focused on the characteristic of microvascular invasion (MVI), aiming to construct a prognostic model to predict patient

[1]Department of Hepatobiliary Surgery, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, Shaanxi Province, China. [2]Department of Gynaecology and Obstetrics Department, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, Shaanxi Province, China. [3]Jiayi Zhang, Zheng Zhang and Chenqing Yang contributed equally to this work. ✉email: qingguangliu@xjtu.edu.cn; 13572431619@163.com

survival and to compare the immune infiltration patterns in different high-risk groups, with the ultimate goal of informing clinical treatment strategies.

Over the past few years, significant technological progress has unveiled the complexity of gene expression within tumor tissues through the power of bulk transcriptomic sequencing[5]. High-throughput sequencing techniques, most notably RNA sequencing (RNA-seq), have revolutionized our understanding by allowing for the detailed examination of gene expression at a single-cell level through single-cell RNA sequencing (scRNA-seq)[6]. By peering into the intricacies of individual tumor cells, we can gain a clearer understanding of the subtle biological landscapes of cancer and develop targeted therapies that can address the disease's diversity[7].

Whole-genome expression profiling offers detailed insights into the diversity of diseases, which is invaluable for disease diagnosis, prediction of treatment response, and prognosis evaluation. Numerous studies have assessed the prognostic impact of array-based gene expression signatures derived from HCC tumors[8]. Previous studies often utilized traditional modeling approaches to discover genetic signatures that can forecast the likelihood of recurrence and/or death in patients with hepatocellular carcinoma (HCC)[9,10]. Yet, when compared with these conventional methods, machine learning algorithms have demonstrated enhanced data fitting capabilities and are progressively being integrated into biomedical research and applications[11]. The integration of diverse algorithms within the framework of stacking methodologies has proven to enhance predictive accuracy significantly, showcasing an impressive performance in forecasting outcomes[12]. In our study, we identified differentially expressed genes from genomic and single-cell sequencing data. Utilizing 101-combination machine learning algorithms, we constructed a prognostic model for liver cancer. To mitigate overfitting, we employed a leave-one-out cross-validation (LOOCV) framework. The resulting prognostic model not only demonstrated robust predictive power but also held significant implications for clinical decision-making.

Results

Differential gene identification

We initiated our analysis by preprocessing TCGA data, employing a Log2 (TPM + 1) transformation to normalize the data (Fig. 1A). Subsequently, we conducted a differential analysis between tumor samples from MVI-positive patients and their corresponding normal adjacent tissues. Utilizing the criteria of $|\log FC| \geq 1$ and $p < 0.05$, we identified 3183 upregulated genes and 653 downregulated genes. We then selected the upregulated genes (DEGs) specific to MVI tumor samples for further analysis.

We meticulously preprocessed the single-cell data from the GSE242889 dataset, applying appropriate quality thresholds to filter out and retain only high quality data (Supplementary Figure S1A, B). To mitigate batch effects, we employed the Harmony algorithm to integrate samples. Utilizing the "clustree" tool, we analyzed cell clustering at various resolutions and identified that at a resolution of 0.6, the cells were optimally segregated into 27 distinct clusters (Supplementary Figure S1C, D). In subsequent analyses, we disregarded less relevant clusters, focusing on the 0–20 clusters. We then applied the t-SNE algorithm to dimensionally reduce and cluster sample features, revealing the distribution of cells across different sample types. Figure 1B distinctly illustrates the clear segregation of cell clusters between tumor samples associated with microvascular invasion (MVI) and those not associated with MVI, particularly within clusters 7, 15, 18, and 20 of the t-SNE clustering (Supplementary Figure S1E). Based on this, we designated these clusters as malignant cell groups related to MVI. Following this, we meticulously annotated these cell groups using cell marker genes sourced from literature[13] and the Cellmarker database (Fig. 1C). Figure 1G depicts the expression patterns of these cell marker genes. With the FindAllMarkers tool, we identified differentially expressed genes across various cell groups, filtering them based on a criterion of $|\log FC| \geq 1$ and $p < 0.05$. Figure 1D presents the differential genes for malignant cells related to MVI, macrophages, hepatic stellate cells (HSC), and lymphocytes (B + T subgroups). Additionally, we calculated the proportions of various cell groups in different samples, further confirming the presence of the majority of MVI-related tumor cell groups in MVI-associated tumor samples (Fig. 1E, F). Ultimately, we extracted 4,119 highly expressed genes (MRMCGs) from the malignant cell groups related to MVI and intersected them with 3183 DEGs to derive the differentially expressed genes associated with MVI (MRGs) (Supplementary Figure S1F).

Signaling pathway analysis

In our analysis of the GSE242889 dataset using the "CellChat" R package for cell communication, we focused on the pathway enrichment between MVI-related malignant cells and other cellular subpopulations[14]. We identified significant interactions between MVI-associated malignant cells and hepatic stellate cells, particularly within the PTN signaling pathway, where these cells exert considerable influence (Fig. 2A–D). PTN is known to regulate angiogenesis by directly stimulating endothelial cells and indirectly by modulating the angiogenic effects of Vascular Endothelial Growth Factor A (VEGF-A)[15]. Extant literature documents the active role of the PTN pathway in angiogenesis in various cancers, including breast cnacer[16], small cell lung cancer[17], prostate cancer[18], glioblastoma[19], and colorectal cancer[20]. However, its role in liver cancer has been less explored. Our innovative single-cell data analysis reveals that MVI-related malignant cell subpopulations communicate significantly with hepatic stellate cells through autocrine/paracrine signaling within the PTN pathway, a feature absent in non-MVI-related malignant cell subpopulations. This suggests that MVI-related malignant cells can regulate tumor angiogenesis by interacting with stellate cells via the PTN pathway, offering new insights into the mechanisms of liver cancer and potential therapeutic avenues.Furthermore, our analysis of MVI related genes (MRGs) using KEGG and GSEA pathways identified significant enrichment in pathways related to DNA replication, base excision repair, cell cycle regulation, and immunity (Fig. 2E-I). These findings provide a direction for our further research endeavors.

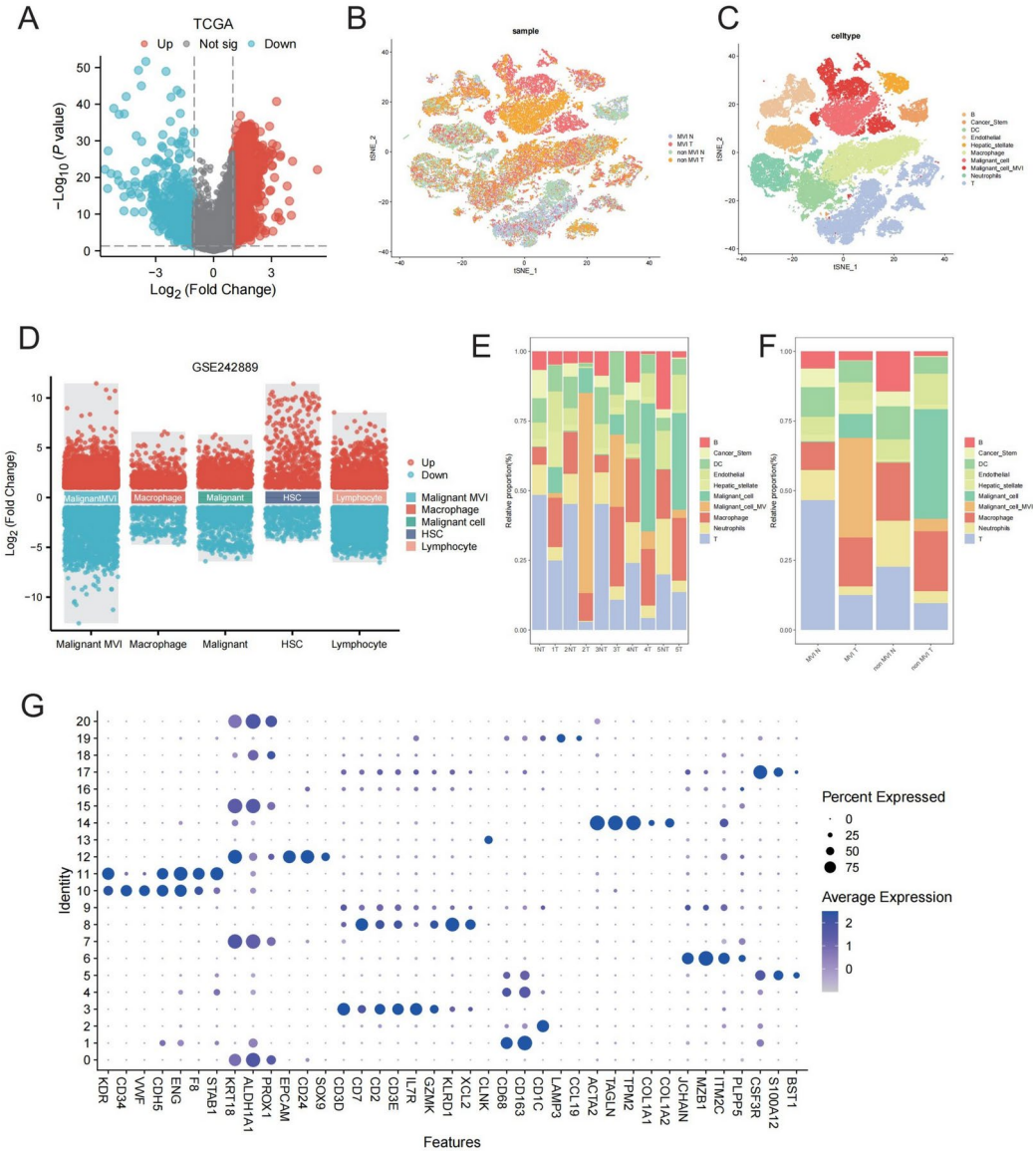

Fig. 1. (**A**) Volcano plot of differentially expressed genes in TCGA. (**B**) tSNE projection within each sample origin. (**C**) tSNE showing 10 cell types. (**D**) Volcano plot of differentially expressed genes in malignant cell related to MVI, macrophages, hepatic stellate cells (HSC), and lymphocytes versus other cells. (**E**) Proportion of each cell type in every sample included NT (non-tumour sample) and T (tumour sample). (**F**) Proportion of each cell type in MVI N (MVI + normal sample), MVI T (MVI + tumor sample), non-MVI T (MVI-normal sample), non-MVI N (MVI-tumor sample). (**G**) Dotplot of marker genes for ten major lineages.

Construction, validation, and evaluation of the prognostic model

We initially employed LASSO regression analysis to perform dimensionality reduction on 267 MRGs, culminating in the selection of 13 candidate genes for the construction of a prognostic model (Supplementary Figure S1G). We then combined 101 combinations of 10 machine learning algorithms to independently screen for key genes. The C-indices for these 101 models were calculated in the TCGA-LIHC training set and the ICGC external validation set. The optimal model, Stepcox(forward) + RSF, demonstrated an average C-index of 0.835 (Fig. 3A). This model encompasses 11 genes: NOP16, YIPF1, HMMR, NDC80, DYNLL1, CDC34, NLN, KHDRBS3, MED8, SLC35G2, and RAB3B. The corresponding weights of each gene within the model are detailed in Supplementary Table S4. We have conducted a comprehensive compilation of 38 published hepatocellular carcinoma (HCC) prognostic prediction models. Compared to the C-indices of 38 published HCC prognostic models, our selected risk model exhibited significant sensitivity and specificity in both the training and validation sets (Fig. 3B and Supplementary Table S2). We also assessed the stability of the prognostic predictions by comparing the AUC values at 1, 3, and 5 years (Fig. 3C). Based on the median risk score of the final model, we stratified the patients into high-risk and low-risk groups and conducted survival analyses in both the training and validation sets, revealing significant differences and indicating the poor prognosis associated with

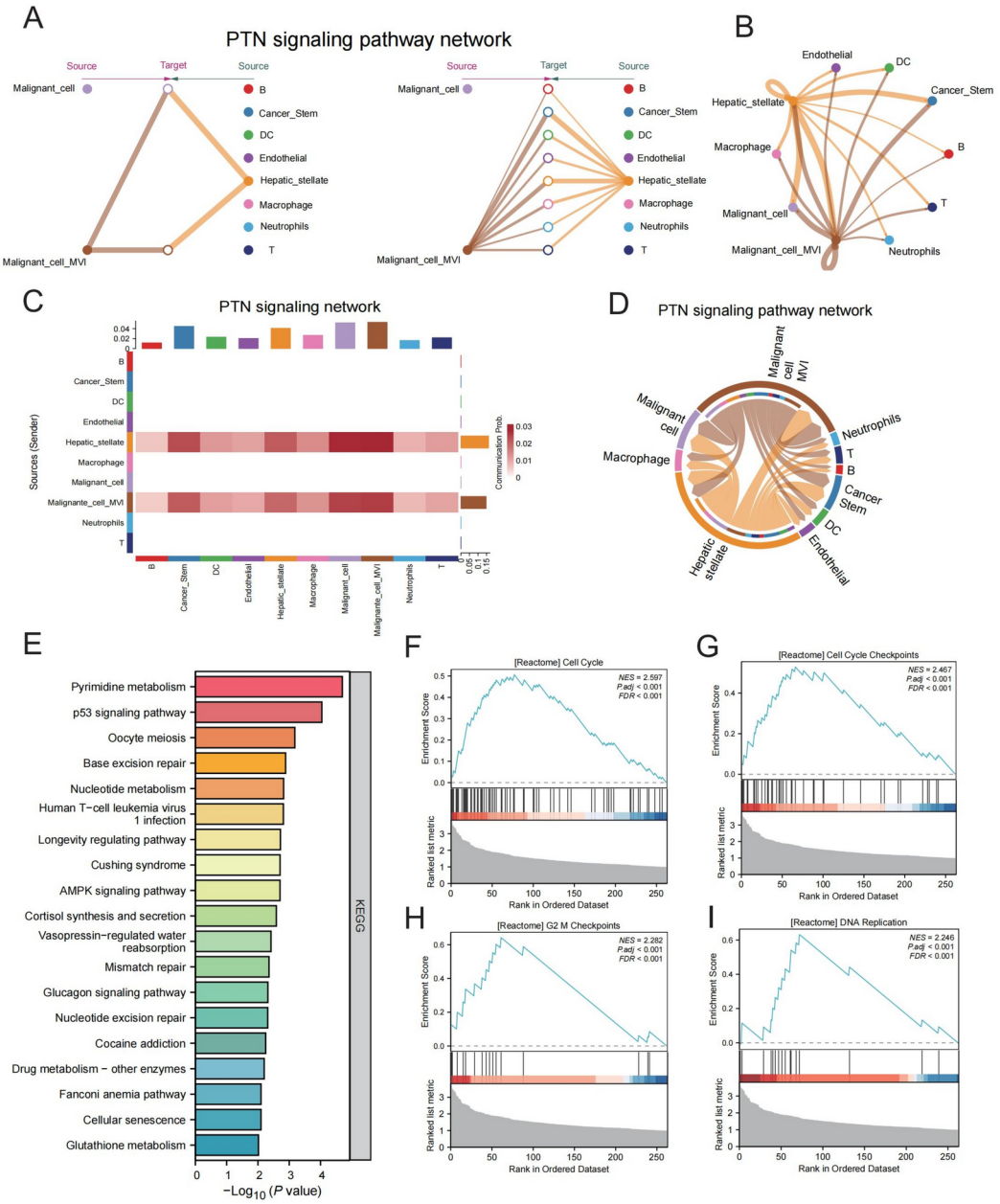

Fig. 2. (**A**) Predicted communications of the PTN signaling pathway between MVI related malignant cells and other cells. (**B**) Circular diagram illustrating the cell–cell communications of 10 types of cells in the PTN signaling pathway. (**C**,**D**) Heatmap and chord diagram of cell–cell communications in the PTN signaling pathway. (**E**) KEGG pathway analysis of MRGs. (**F**–**I**) GSEA pathway analysis of MRGs.

the high-risk group (Fig. 3D-E). Lastly, we performed a meta-analysis using the univariate regression results of the model genes in both the training and validation sets (Fig. 3F). The results indicated that the high-risk group's prognosis was significantly worse than that of the low-risk group, with a marked statistical difference. To further substantiate our model's credibility, we conducted the same analysis on 38 published models, and the results are shown in Fig. 3G, demonstrating that the prognostic risk identified by our constructed model was substantially higher than that of other models.

Gene mutation analysis

Previous studies have demonstrated that genetic mutations are often positively correlated with the malignancy of tumors and significantly impact prognosis[21]. Before, we enriched mutation-associated pathways through MRGs. To evaluate the correlation between the model's risk scores and genetic mutations, we conducted an analysis of the frequency of genetic mutations in both high- and low-risk groups from the TCGA cohort. The results were then presented visually using a waterfall plot (Fig. 4A). Furthermore, we examined the top 10 mutated genes in both risk groups and found that the types and frequencies of these genes varied between the groups,

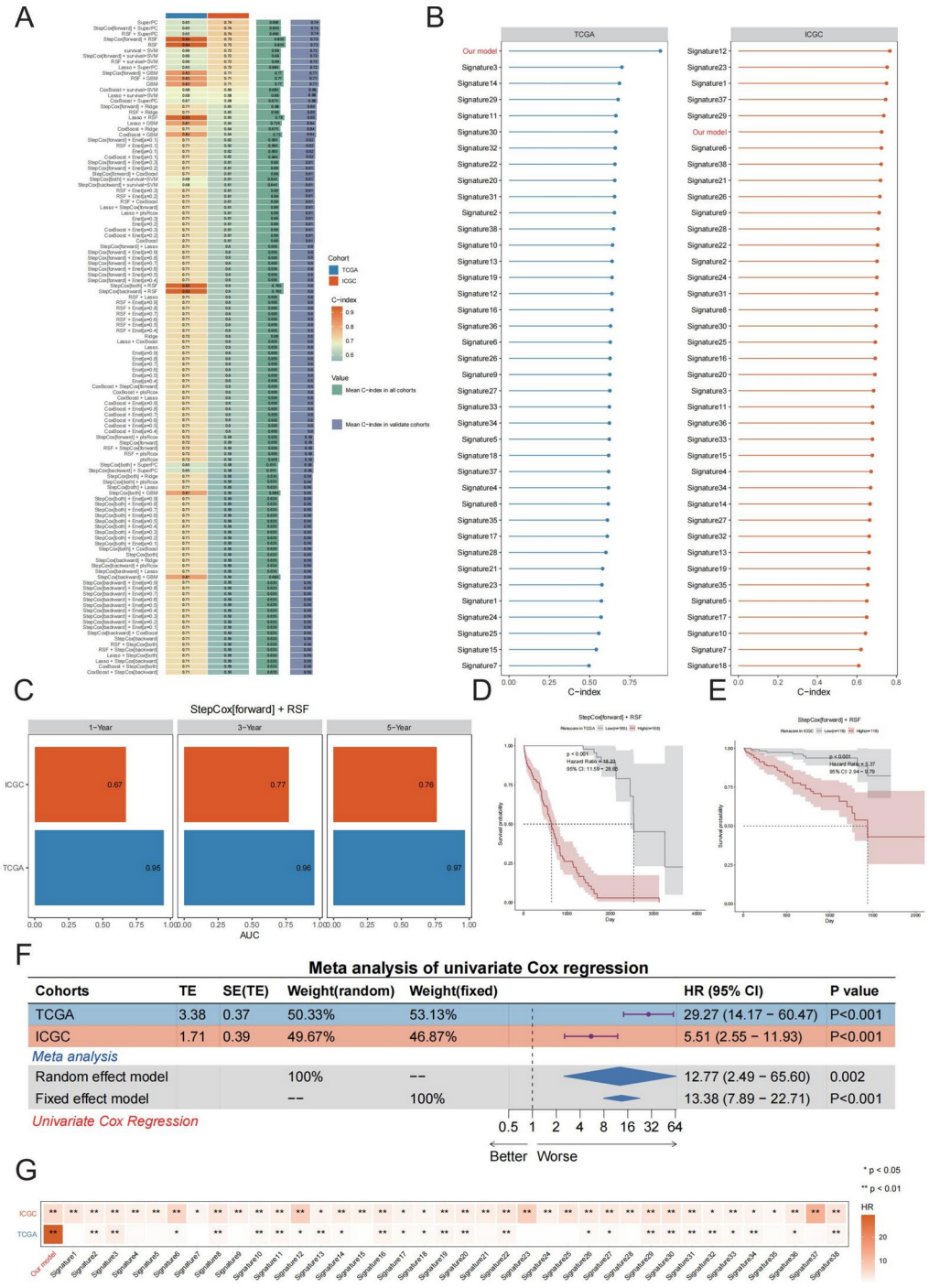

Fig. 3. (**A**) C-index of different models constructed using 101 machine-learning algorithm combinations in the TCGA training set and ICGC external validation set. (**B**) C-index comparison of the prognostic model based on model genes and 38 published HCC prognostic models in the TCGA-LIHC training set and ICGC external validation sets. (**C**) AUC values of TCGA training set and ICGC validation set at 1, 3, and 5 years. (**D**) KM analysis between high- and low-risk groups in TCGA datasets. (**E**) KM analysis between high- and low-risk groups in ICGC datasets. (**F**) Meta analysis of univariate cox regression in TCGA-LIHC training set and ICGC external validation set. (**G**) Meta analysis of univariate cox regression of the prognostic model based on model genes and 38 published HCC prognostic models in the TCGA-LIHC training set and ICGC validation sets.

indicating that distinct mutational landscapes indeed exist between the high- and low-risk groups (Fig. 4B, C). Lastly, we conducted an examination of the mutation types and single nucleotide variant (SNV) types within the TCGA cohort, as well as different risk groups. This comprehensive assessment allowed us to reveal the distinct mutational profiles associated with varying levels of risk. We observed that the high-risk group had a

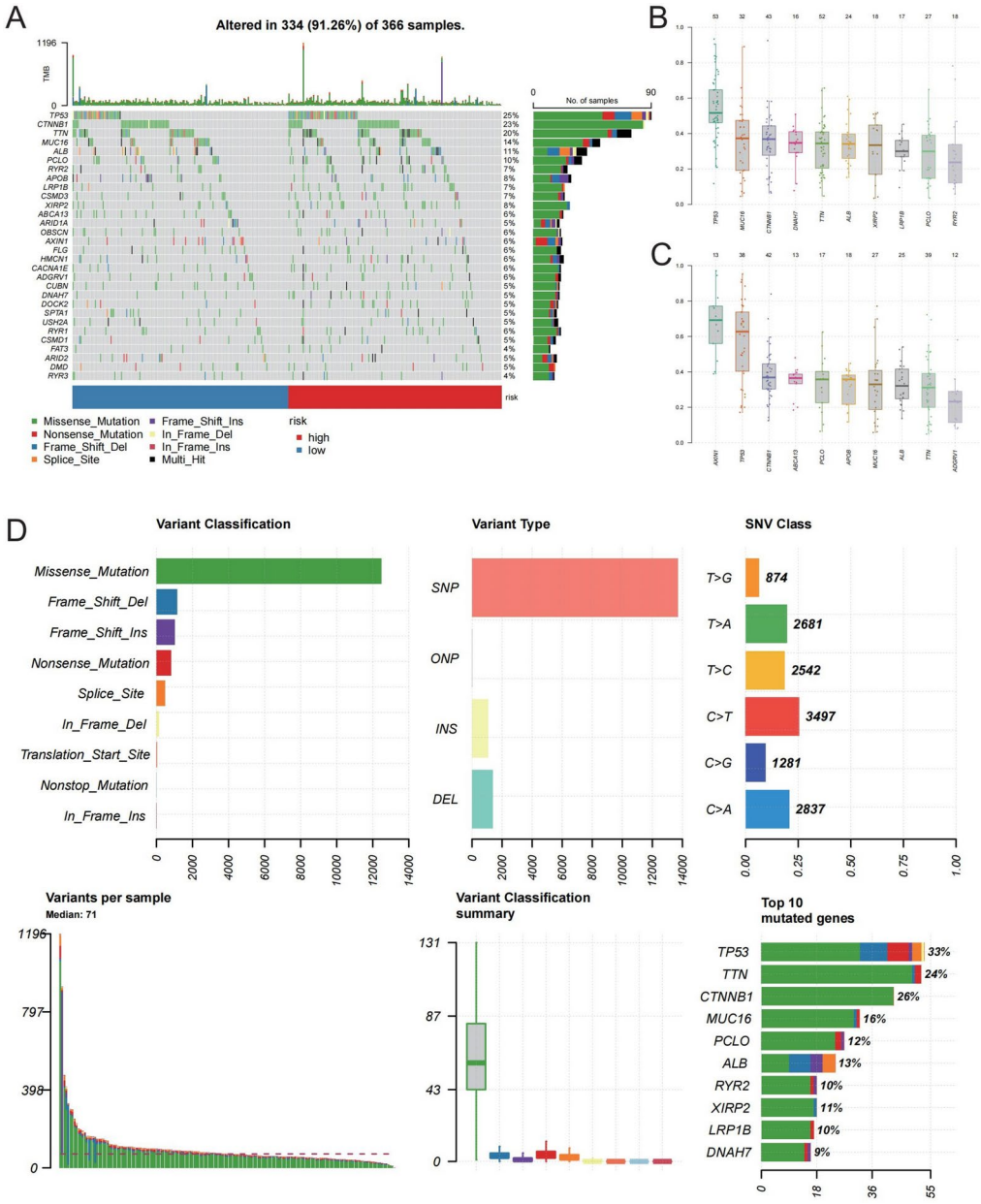

Fig. 4. (**A**) Waterfall diagram showed the top 30 genes with the highest mutation frequency in high and low risk group. (**B**) Top 10 mutated genes in the high-risk cohort. (**C**) Top 10 mutated genes in the low-risk cohort. (**D**) Mutation types and single nucleotide variant (SNV) types in the TCGA dataset.

higher number of mutations, a greater proportion of Frame-Shift Insertion Mutations, and a higher ratio of T-A mutations compared to the low-risk group (Fig. 4D and Supplementary Figure S1H-I).

TIDE score and immune checkpoint analysis

Previous studies have demonstrated that elevated TIDE scores correlate with an increased risk of tumor immune evasion and a diminished efficacy of immune checkpoint inhibitors. By leveraging the TIDE algorithm to analyze datasets from TCGA and ICGC, we observed significantly higher TIDE scores in the high-risk group as opposed to the low-risk group, with a pronounced statistical distinction (Fig. 5A, E). We also calculated scores for immune therapy (SPONDER), T cell exclusion (EXCLUSION), and myeloid-derived suppressor cells (MDSC) using the TIDE algorithm for both risk categories, uncovering substantial statistical differences (Fig. 5B-D, F–H). Moreover, our examination of immune checkpoint gene expression levels revealed that the high-risk group had significantly elevated levels, as depicted in Fig. 5I. Collectively, these results imply that individuals in the high-risk group are likely to experience more pronounced immune evasion, which could result in less favorable outcomes with immunotherapeutic approaches.

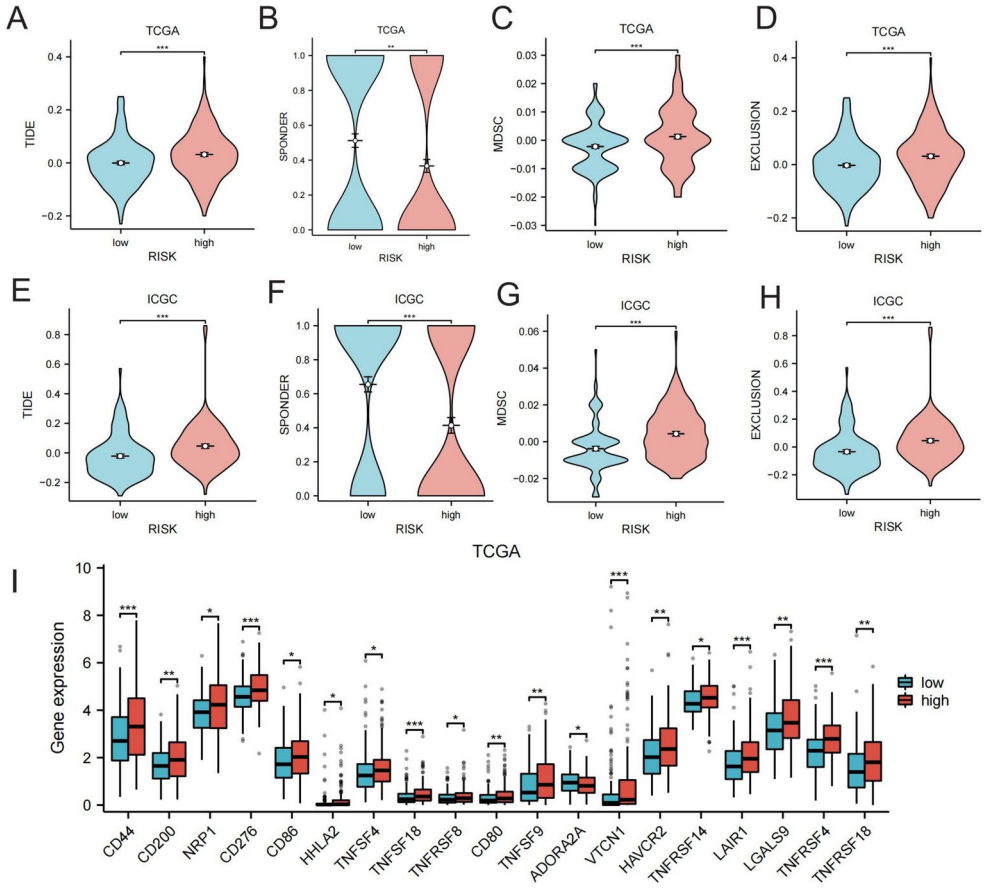

Fig. 5. (**A**–**D**) The TIDE scores, immune therapy scores (SPONDER), T cell exclusion scores (EXCLUSION), and myeloid-derived suppressor cell scores (MDSC) for the high-risk and low-risk groups within the TCGA dataset. (**E**–**H**) The TIDE scores, immune therapy scores (SPONDER), T cell exclusion scores (EXCLUSION), and myeloid-derived suppressor cell scores (MDSC) for the high-risk and low-risk groups within the ICGC dataset. (**I**) Differential expression of immune checkpoints in high-risk and low-risk groups within the TCGA dataset.

Model gene analysis of risk factors, co-expression, and correlation

In order to assess the dependability of our model, we analyzed individually the 11 genes that were instrumental in the development of the definitive risk model and determined that each of these genes possessed a notable ability to predict patient outcomes (Fig. 6A and Supplementary Figure S2A-K). Additionally, the expression of these genes shows significant statistical differences between liver cancer tumors and normal tissues, all of which suggest that the 11 genes play an important role in the development and progression of liver cancer (Supplementary Figure S3A, B, Figure S4A-K and Figure S5). We then examined the co-expression and correlation between these model genes and the mutation-associated genes TP53, CTNNB1, and DNAH7. Our findings revealed a significant link between the elevated expression of the model genes and the heightened expression of the mutated genes, thereby reinforcing the connection between genetic mutations and an elevated risk of poor prognosis in liver cancer (Fig. 6B-E). In further analysis, we explored the relationship between the model genes and immune checkpoint genes, identifying a pattern where increased expression of the model genes coincided with increased expression of immune checkpoints (Fig. 6F).

Analysis of immune infiltration

Moving forward, we sought to delve deeper into the disparities of immune cell infiltration between the high-risk and low-risk cohorts as stratified by our ultimate prognostic model. Utilizing the ssGSEA algorithm[22], we assessed the correlation between the model genes and a panel of 24 immune cell types[23], noting a consistent pattern of correlation among different model genes with respect to the same immune cells. Our analysis revealed an inverse relationship between the majority of model genes and anti-tumor immune cells (B cells, CD8 T cells, DC cells; NK cells, Th17 cells), while an opposite trend was observed with pro-tumor immune cells (Th2 cells) (Fig. 7A)[24,25]. Furthermore, through ssGSEA analysis of immune cell activation, it became evident that the low-risk group exhibited more favorable scores in the activation processes of the majority of immune cells (Fig. 7B). Consequently, we postulate that the low-risk group, as identified by our model, is characterized by a more potent anti-tumor immune milieu, in contrast to the high-risk group, which appears to harbor a more pronounced immune-evasion microenvironment.

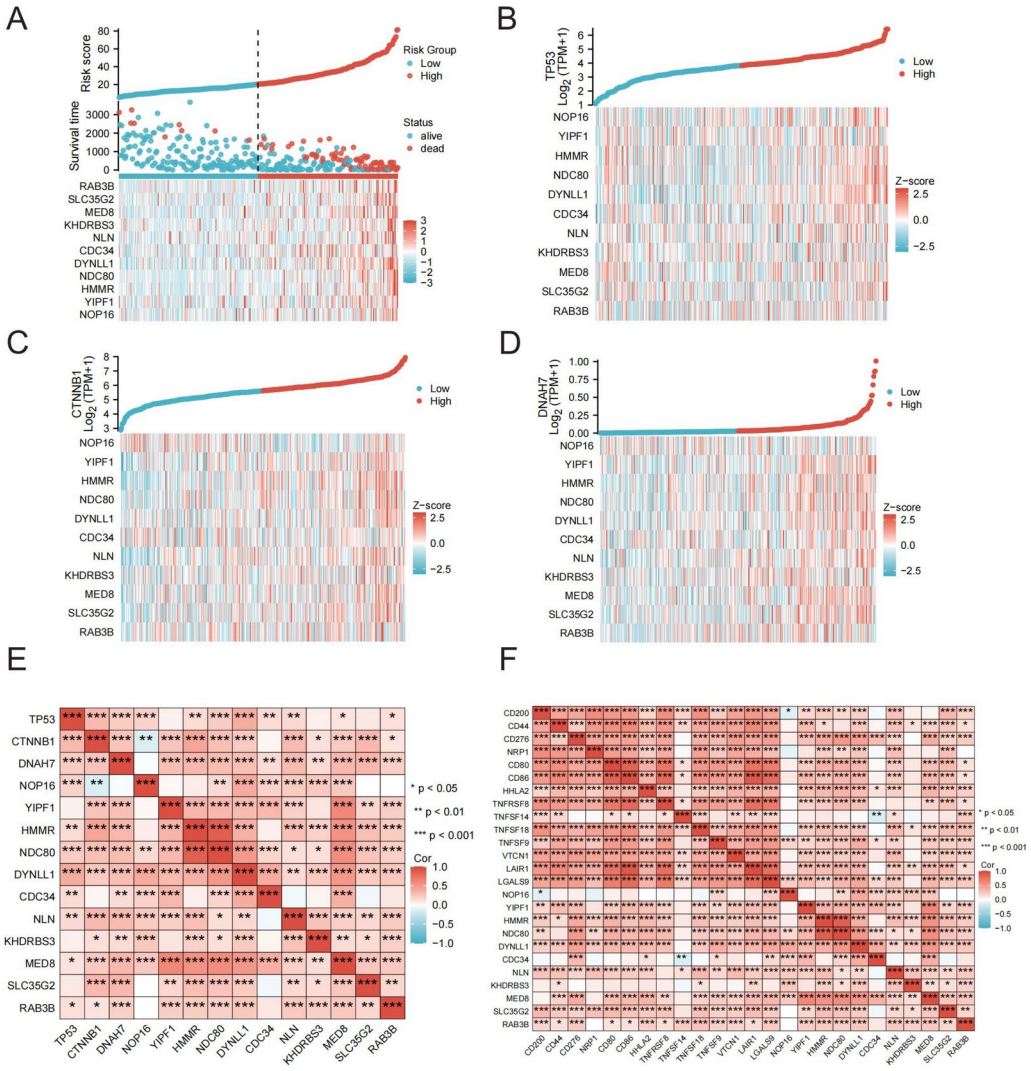

Fig. 6. (**A**) Risk factor plot of the 11 model genes. (**B**–**D**) Co-expression analysis between the model genes and the mutation-associated gene TP53, CTNNB1 and DNAH7. (**E**) Correlation analysis between the model genes and the mutation-associated genes. (**F**) Correlation analysis between the model genes and the immune checkpoint genes.

Continuing our investigation with the CIBERSORT algorithm, we determined the relative proportions of various immune cells within the high-risk and low-risk group (Fig. 8A, B). Notably, the high-risk group exhibited a pronounced increase in the proportion of tumor-associated macrophage M2 cells, in contrast to the low-risk group (Fig. 8C). Conversely, the representation of anti-tumor immune-related lymphocytes (encompassing naive B cells, memory B cells, plasma cells, CD8 T cells, and various subsets of CD4 T cells, as well as both resting and activated NK cells) was significantly reduced in the high-risk group (Fig. 8D). This observation serves to further substantiate the conclusions we have drawn regarding the immune microenvironmental disparities between the high-risk and low-risk groups.

Analysis of clinic and treatment

In our quest to apply the risk stratification from our model to clinical practice, we began by evaluating the variances in tumor stage distribution between the high-risk and low-risk groups (Fig. 8E). Following this, we determined the Immunotherapy Sensitivity Score (IPS) and the IC50 values for both risk groups in the TCGA-LIHC dataset. Notably, the low-risk group showed a superior response to immunotherapeutic interventions (Fig. 8F), whereas the high-risk group was more receptive to chemotherapy regimens (Fig. 8G–K), offering valuable insights for clinical treatment strategies. We hypothesize that these observations might be linked to the high-risk group's immune evasion mechanisms and an increased mutational frequency.

Discussion

The incidence rate of hepatocellular carcinoma is among the top ten globally, and it is still one of the malignant tumors with the highest mortality rates[26]. Predicting the survival of patients with liver cancer can promptly

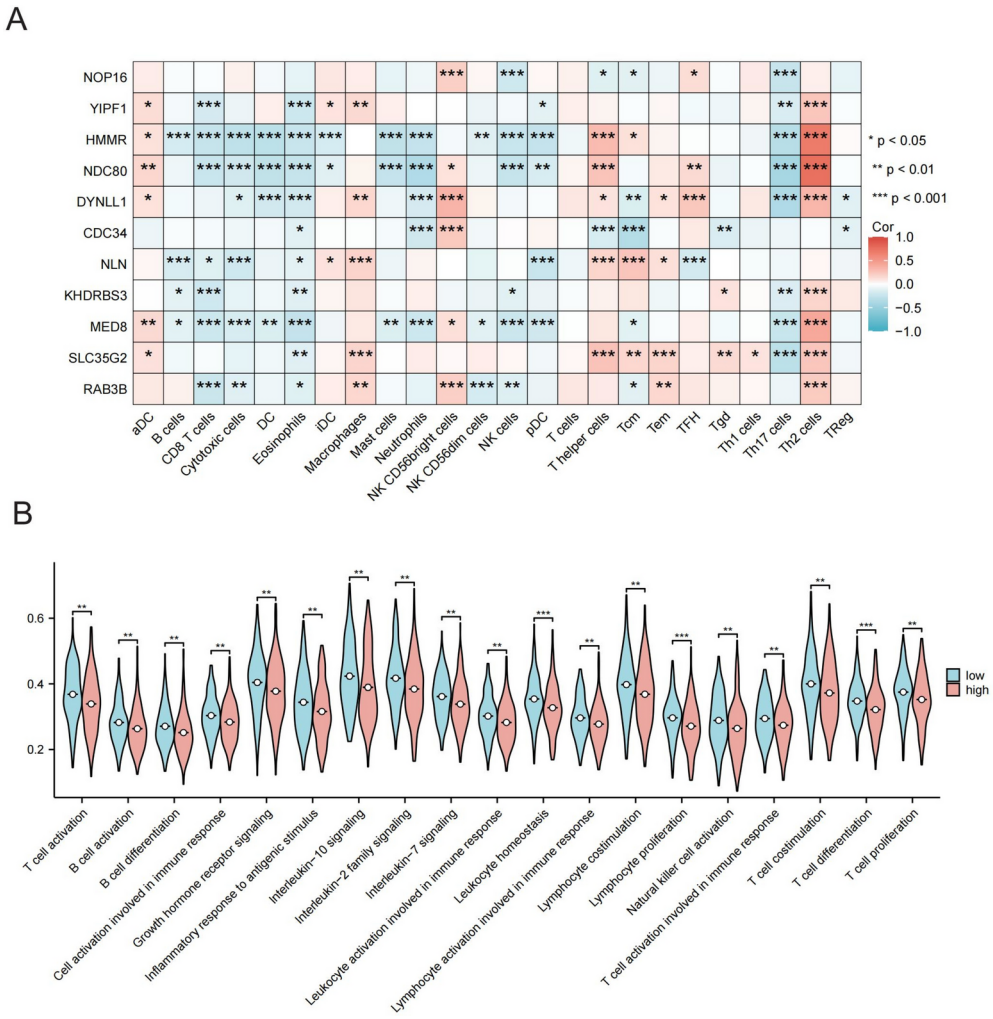

Fig. 7. (**A**) ssGSEA analysis of the correlation between model genes and 24 types of immune cells. (**B**) Differences in immune-related processes scores between high-risk and low-risk groups as analyzed by ssGSEA.

and effectively identify those at high risk of poor prognosis, allowing for early comprehensive monitoring. This is beneficial for guiding clinical decisions and thereby improving the overall prognosis of patients with liver cancer[27]. Previous research has often developed prognostic models for liver cancer that rely heavily on a single algorithm, primarily based on transcriptomic data[28]. These studies typically focus on categorizing patients into groups with differing prognoses and exploring the clinical characteristics of these groups. However, they frequently fall short of investigating the underlying reasons for these differences, which limits their practical application in clinical settings.

Microvascular invasion (MVI) has been established as an independent risk factor for recurrence and metastasis in patients with hepatocellular carcinoma (HCC), significantly correlating with their prognosis[29,30]. In recent years, an increasing number of studies have emphasized the pivotal role of MVI in guiding the treatment of liver cancer[31], with its prognostic predictive ability even surpassing the widely recognized Milan criteria[32]. Our research integrates single-cell data with transcriptomic data, focusing on the clinical characteristic of microvascular invasion, to identify MVI-associated genes. By employing 101 distinct machine learning methods, we have established a prognostic model for liver cancer and developed a scoring system to assess the risk of poor prognosis in patients with HCC.

In this study, we initially defined a malignancy-related cell subset associated with microvascular invasion (MVI) in single-cell data and identified 3,852 differentially expressed genes (MRMCGs) with high expression levels. Through cell–cell communication analysis, we discovered that this subset primarily interacts with the hepatic stellate cells through the PTN pathway, which is crucial for tumor angiogenesis. Subsequently, we conducted a differential analysis between tumor samples from MVI-positive patients in the TCGA dataset and all adjacent normal samples, identifying 2916 highly expressed genes (MRGs). We then found the intersection of MRMCGs and MRGs to obtain MVI-DEGs, which were used to construct a prognostic model. By employing KEGG and GSEA analyses, we revealed that these intersecting genes are predominantly enriched in pathways related to liver cancer metastasis (p53 signaling pathway)[33], genetic mutations, and immune responses (AMPK

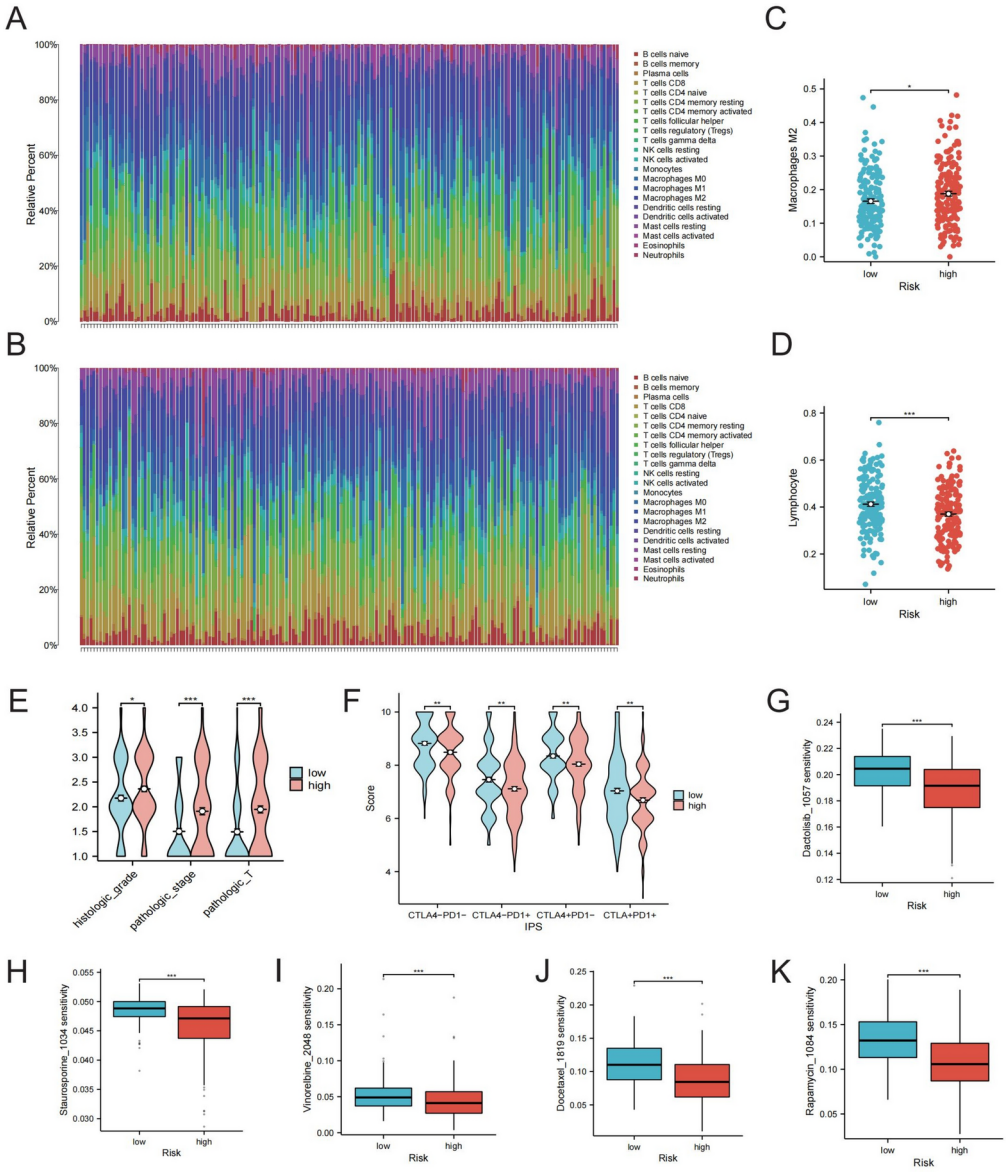

Fig. 8. (**A**) Barplot of immune cell proportions calculated by the CIBERSORT algorithm for the high-risk group in TCGA. (**B**) Barplot of immune cell proportions calculated by the CIBERSORT algorithm for the low-risk group in TCGA. (**C**) The difference in M2 cell proportions between high-risk and low-risk groups. (**D**) The difference in lymphocyte proportions between high-risk and low-risk groups. (**E**) Comparison of clinical stage differences between high-risk and low-risk groups. (**F**) Analysis of the difference in immunophenotype score (IPS) between high-risk and low-risk groups. (**G**–**K**) Drug sensitivity analysis between high-risk and low-risk groups.

signaling pathway)[34]. Accordingly, we subsequently focused our analysis on the mutational and immunological characteristics of the high- and low-risk groups stratified by the prognostic model.

Through the application of a diverse array of 101 machine learning strategies, we successfully identified an optimal analytical model for our prognostic study[35]. The refined selection process led us to pinpoint 11 key model genes—NOP16, YIPF1, RAB3B, SLC35G2, MED8, KHDRBS3, NLN, CDC34, DYNLL1, NDC80, and HMMR—that form the backbone of our predictive model. Existing literature has confirmed through experimental evidence that genes such as RAB3B[36], MED8[37], KHDRBS3[38], CDC34[39], DYNLL1[40], NDC80[41], and HMMR[42] play pivotal roles in the initiation and progression of liver cancer. Moreover, the significance of genes like SLC35G2[43] and NLN[44] in prognosis prediction has been extensively documented in the literature.

The efficacy of our model surpasses that of 38 previously published models, showcasing a marked improvement in predictive power. In addition to their collective impact, each of these model genes has been shown to independently predict the prognosis of liver cancer, providing a robust foundation for our model use as standalone prognostic indicators.

Mutational profiling has proven instrumental in elucidating the character and etiology of mutations present in different types of cancer, which in turn refines clinical decision-making[45]. High-frequency mutations, in particular, are significantly correlated with the propensity for tumor invasion, likelihood of recurrence, and potential for metastasis, making them critical factors in the overall prognosis for cancer patients[21]. Studies have shown that the development of cancer is intricately linked to its multifaceted tissue context, which it relies on for continuous growth, invasion, and spread[46]. Within the tumor microenvironment, a complex array of cells, including macrophages and fibroblasts associated with the tumor, along with immune cells, exert a substantial influence on the advancement, invasiveness, and dispersion of cancer through the secretion of a spectrum of cytokines and metabolic factors[47]. In summary, these differential characteristics can lead to distinct prognostic features. To elucidate the disparities within our model's stratification to inform treatment strategies, we sequentially undertook analyses focusing on mutations, immune profiles, and treatment-related factors. We observed a higher frequency of genetic mutations in the high-risk prognosis group, with a significant proportion of frameshift mutations and T-A transitions. Through immunological analyses, we found that the low-risk group exhibited stronger antitumor immunity and better response to immunotherapy. In contrast, the high-risk group showed more pronounced immune tolerance characteristics and performed poorly under immunotherapy. Finally, we performed drug sensitivity analysis by calculating the IC50 values, identifying five potential therapeutic drugs that may improve prognosis for patients in the high-risk group.

In summary, we have developed and validated a prognostic model for hepatocellular carcinoma patients by integrating various machine learning methods and calculating risk scores for different patients. This scoring system can serve as an independent prognostic indicator with good reliability and accuracy. However, our study inevitably has some limitations. Firstly, the classification of MVI-related malignant cells in the single-cell dataset is relatively coarse, based solely on the spatial distribution of different sample groups from t-SNE analysis results. Therefore, it may include some malignant cells that are not related to MVI. Secondly, our primary data sources are the TCGA and ICGC datasets. While the predictive models established based on these two datasets have demonstrated superior predictive capabilities, there is still a need to expand the sample size. Moreover, we lack further in vivo or in vitro functional experiments to explore the underlying molecular mechanisms of the model genes we have identified.

Materials and methods

Data sources

We accessed a comprehensive RNA-sequencing dataset, clinical information and somatic mutation data from the TCGA-LIHC, which obtained directly from the TCGA data repository. (http://tcga.cancer.gov/; November 19, 2023). We utilized single-cell sequencing data from the GSE242889 dataset, which is accessible within the GEO database (https://www.ncbi.nlm.nih.gov/geo/). We have been granted access to download data from the ICGC repository and have obtained the ICGC-LIRI-JP dataset, encompassing RNA sequencing profiles and associated clinical data (https://dcc.icgc.org/). The IPS analysis data was downloaded from the TCIA website (https://tcia.at/home).

Data processing

Within the TCGA-LIHC and ICGC-LIRI-JP dataset, the raw count data extracted from the project served as the foundation for our differential expression analysis. In contrast, for all other analytical pursuits, we relied on the Log2 (TPM + 1) transformed data, ensuring a more accurate and comprehensive exploration of the genetic expressions under study.We employed the "Seurat" R package[48] to process 10 × single-cell transcriptomic data. The quality control parameters for the scRNA-seq data were stringent and included the following filters: Quality control standards for scRNA-seq data included the following criteria: (1) nCount_RNA > 200; (2) nFeature_RNA > 3; (3) percent_mito < 25%; (4) percent_ribo > 3%; (5) percent_hb < 1%.These criteria ensured that the data met high-quality standards for subsequent analysis.

Identification of MVI-related genes in TCGA

We selected tumor samples from MVI-positive patients and normal adjacent tissues from the TCGA dataset for differential analysis. By applying the threshold criteria of log fold change|logFC|≥ 1 and p-value < 0.05, we identified 2916 genes (DEGs) associated with microvascular invasion (MVI).

Identification of MVI-related malignant cell genes in GSE242889

Post-quality control on the GSE242889 single-cell dataset, we deployed the Harmony function to correct for batch effects and applied t-SNE for dimensionality reduction, resulting in the classification of cells into 20 clusters at a resolution of 0.6.To annotate these clusters, we employed the "FindAllMarkers" function to identify signature genes, followed by manual annotation. We further differentiated MVI-associated malignant cell subpopulations from the overall malignant cells based on t-SNE plots stratified by sample types. MVI-related malignant cell subpopulations were identified from the malignant cell group through stratified t-SNE analysis. Subsequently, we performed differential expression analysis on the MVI-related malignant cell clusters against other clusters. Setting the logFC ≥ 1 and the adjusted p-value < 0.05, we successfully identified 3852 genes (MRMCGs) that were markedly differentially expressed in the MVI-related malignant cells. Reference cell markers were sourced from published articles[13] and the comprehensive CellMarker database (https://xteam.xbio.top/CellMarker/). The cell markers we have chosen are in Supplementary Table S1.

Cell–cell communication

We employed the "CellChat" R package to perform a analysis of intercellular interactions within single-cell transcriptomic data[14]. Our study aimed to determine the interactions between MVI-related malignant cells

and other cell types. Additionally, we sought to forecast the likelihood of intercellular communication and the potential pathways involved.

Analysis of function and pathway enrichment

We intersected the selected MVI-DEGs with MRMCGs to identify differentially expressed genes associated with MVI (MRGs). Subsequently, we conducted a KEGG pathway analysis for these genes and visually presented some of the top-ranking pathways[49]. Furthermore, we utilized GSEA (Gene Set Enrichment Analysis) to illustrate the enrichment results of several significant pathways[50,51].

Construction, validation, and evaluation of the prognostic model

We performed initial dimensionality reduction on the 267 selected MVI-DEGs using LASSO regression analysis, which led to the identification of 13 candidate genes. Drawing on the research of Liu and colleagues[35], we employed a variety of machine learning algorithms, comprising 101 different combinations across 10 methods, to construct a prognostic model. These methodologies included CoxBoost, Elastic Net (Enet), Generalized Boosted Regression Modeling (GBM), Least Absolute Shrinkage and Selection Operator (Lasso), Partial Least Squares Regression for Cox (plsRcox), Ridge Regression, Random Survival Forest (RSF), Stepwise Cox, Supervised Principal Component Analysis (SuperPC), and Support Vector Machine for Survival Analysis (survival-SVM). We applied these combinations to the 13 candidate genes and utilized Leave-One-Out Cross-Validation (LOOCV) to mitigate overfitting.

We employed the TCGA-LIHC dataset as our test set and validated our findings in the ICGC-LIRA-JP dataset. Based on the optimal algorithm determined to be Stepcox(forward) + RSF, we constructed our model. Subsequently, we gathered 38 published HCC prognostic models (Supplementary Table S2) and applied them to our TCGA-LIHC training set and the ICGC external validation set to calculate the C-index. This enabled a more nuanced comparative analysis of our refined model against the existing models within the scholarly domain.

Patients in both the TCGA-LIHC training cohort and the ICGC external validation set were stratified into high-risk and low-risk groups based on their median risk scores (Supplementary Table S3). We conducted survival analyses on the high- and low-risk groups derived from both the training and validation cohorts and plotted Kaplan–Meier curves, which consistently revealed significant statistical differences. Finally, we conducted univariate Cox regression analyses on our final model, followed by a meta-analysis of both the training and validation datasets. This comprehensive approach allowed us to rigorously assess the model's accuracy and its capacity for generalization across different data cohorts.

Gene mutation analysis

After downloading somatic mutation data, we utilized the "maftools" R package to visually present the distribution of mutated genes across high- and low-risk groups, identifying the top 10 most frequently mutated genes in each cohort[52]. We then conducted statistical analyses on mutation types and single nucleotide variant (SNV) mutation types, followed by visualization to reveal distinct mutational landscapes between the high- and low-risk groups.

TIDE score and immune checkpoint analysis

We employed the TIDE website (http://tide.dfci.harvard.edu/) to compute the TIDE scores for both high- and low-risk groups within the TCGA and ICGC cohorts, subsequently conducting a statistical analysis to discern significant differences[53]. Additionally, by analyzing the immune checkpoint genes (ICG), we further assessed the immunological disparities between the high and low risk groups, offering insights to inform potential therapeutic choices.

Analysis of the correlation between model genes and mutations, immunity

We performed co-expression and correlation analyses between key genes from our model and mutated genes, further confirming the higher mutational burden observed in the high-risk group. We performed a correlation analysis between the genes in our model and those of immune checkpoints, thereby elucidating a heightened correlation between the high-risk group and immune checkpoint genes.

Immune infiltration ssGSEA analysis

We employed single-sample gene set enrichment analysis (ssGSEA) to evaluate the correlation between our model genes and various immune cell populations. Additionally, we conducted ssGSEA on immune-related pathways and immune cell activation processes in both high- and low-risk groups, followed by statistical difference analysis. This approach further elucidated the disparities in the immune microenvironment between the high- and low-risk cohorts.

Clinical analysis, immunotherapy efficacy and drug sensitivity

By applying the CIBERSORT algorithm[54], we determined the relative frequencies of immune cells within high- and low-risk groups, illustrating these results through visualization. Focusing on the quantification of T cells and macrophages, we proceeded with a statistical analysis to underscore the variations in immune infiltration between the risk groups. To translate this model into clinical practice, we also conducted a statistical comparison of clinical staging between the high- and low-risk groups. Furthermore, we analyzed differences in immune treatment responses and chemotherapeutic drug sensitivities by "Oncopredict" R package between the groups, offering insights to guide therapeutic decisions.

Acquisition of tissue samples

HCC tissues and matched adjacent non-tumor tissues used in this study were collected from the First Affiliated Hospital of Xi'an Jiaotong University (Xi' an, China). We have obtained written informed consent from each patient[55].

Quantitative real-time polymerase chain reaction (qRT-PCR)

In accordance with the manufacturer's protocol for Trizol reagent (Invitrogen, Carlsbad, CA, USA), the total RNA was extracted from both cells and tissues. The RNA levels of model genes were determined using quantitative real-time PCR (qRT-PCR). The relative expression levels were calculated by the $2^{-\Delta\Delta Ct}$ method, with normalization of Ct values to GAPDH as an internal control. The primers used are shown in Supplementary Table S5.

Immunohistochemical analysis

The Human Protein Atlas (HPA) database (https://www.proteinatlas.org/) aims to map the distribution of all proteins in human cells, tissues, and organs by integrating multiple omics technologies, including antibody-based imaging, mass spectrometry-based proteomics, transcriptomics, and systems biology. We acquired immunohistochemical images of key model genes in both hepatocellular carcinoma and normal liver tissues from the Human Protein Atlas (HPA) database, followed by a comparative analysis of the differential protein expression patterns of these model genes between malignant and normal hepatic tissues.

Data availability

In this study, we analyzed publicly accessible datasets. The data utilized can be accessed through the following repositories: The Cancer Genome Atlas (TCGA) at http://www.cancer.gov/tcga, the International Cancer Genome Consortium (ICGC) at https://dcc.icgc.org/, the Gene Expression Omnibus (GEO) at https://www.ncbi.nlm.nih.gov/geo, and the Cancer Immunome Atlas (TCIA) at https://tcia.at/home. In accordance with the journal's guidelines, the datasets used in this research are available upon request from the corresponding author.

Received: 19 November 2024; Accepted: 20 February 2025
Published online: 07 March 2025

References

1. Singal, A. G., Kanwal, F. & Llovet, J. M. Global trends in hepatocellular carcinoma epidemiology: implications for screening, prevention and therapy. *Nat. Rev. Clin. Oncol.* **20**, 864–884. https://doi.org/10.1038/s41571-023-00825-3 (2023).
2. Lee, S. et al. Effect of microvascular invasion risk on early recurrence of hepatocellular carcinoma after surgery and radiofrequency ablation. *Ann. Surg.* **273**, 564–571 (2021).
3. Wu, Q., Zhou, L., Lv, D., Zhu, X. & Tang, H. Exosome-mediated communication in the tumor microenvironment contributes to hepatocellular carcinoma development and progression. *J. Hematol. Oncol.* **12**, 53. https://doi.org/10.1186/s13045-019-0739-0 (2019).
4. Li, X.-Y., Shen, Y., Zhang, L., Guo, X. & Wu, J. Understanding initiation and progression of hepatocellular carcinoma through single cell sequencing. *Biochim. Biophys. Acta BBA Rev. Cancer* **1877**, 188720. https://doi.org/10.1016/j.bbcan.2022.188720 (2022).
5. Hong, M. et al. RNA sequencing: new technologies and applications in cancer research. *J. Hematol. Oncol.* **13**, 166. https://doi.org/10.1186/s13045-020-01005-x (2020).
6. Lei, Y. et al. Applications of single-cell sequencing in cancer research: progress and perspectives. *J. Hematol. Oncol.* **14**, 91. https://doi.org/10.1186/s13045-021-01105-2 (2021).
7. Kuksin, M. et al. Applications of single-cell and bulk RNA sequencing in onco-immunology. *Eur. J. Cancer* **149**, 193–210. https://doi.org/10.1016/j.ejca.2021.03.005 (2021).
8. Cai, J. et al. Prognostic biomarker identification through integrating the gene signatures of hepatocellular carcinoma properties. *EBioMedicine* **19**, 18–30. https://doi.org/10.1016/j.ebiom.2017.04.014 (2017).
9. Tang, Y. et al. Identification and validation of a prognostic model based on three MVI-related genes in hepatocellular carcinoma. *Int. J. Biol. Sci.* **18**, 261–275. https://doi.org/10.7150/ijbs.66536 (2022).
10. Li, Y. & Zeng, X. A novel cuproptosis-related prognostic gene signature and validation of differential expression in hepatocellular carcinoma. *Front. Pharmacol.* **13**, 1081952. https://doi.org/10.3389/fphar.2022.1081952 (2022).
11. Swanson, K., Wu, E., Zhang, A., Alizadeh, A. A. & Zou, J. From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. *Cell* **186**, 1772–1791. https://doi.org/10.1016/j.cell.2023.01.035 (2023).
12. Gong, Q., Chen, X., Liu, F. & Cao, Y. Machine learning-based integration develops a neutrophil-derived signature for improving outcomes in hepatocellular carcinoma. *Front. Immunol.* **14**, 1216585. https://doi.org/10.3389/fimmu.2023.1216585 (2023).
13. Huang, H. et al. Multi-transcriptomics analysis of microvascular invasion-related malignant cells and development of a machine learning-based prognostic model in hepatocellular carcinoma. *Front. Immunol.* **15**, 1436131. https://doi.org/10.3389/fimmu.2024.1436131 (2024).
14. Jin, S. et al. Inference and analysis of cell-cell communication using Cell Chat. *Nat. Commun.* **12**, 1088. https://doi.org/10.1038/s41467-021-21246-9 (2021).
15. Papadimitriou, E. et al. Pleiotrophin and its receptor protein tyrosine phosphatase beta/zeta as regulators of angiogenesis and cancer. *Biochim. Biophys. Acta BBA Rev. Cancer* **252–265**, 2016. https://doi.org/10.1016/j.bbcan.2016.09.007 (1866).
16. Fang, W., Hartmann, N., Chow, D. T., Riegel, A. T. & Wellstein, A. Pleiotrophin stimulates fibroblasts and endothelial and epithelial cells and is expressed in human cancer. *J. Biol. Chem.* **267**, 25889–25897 (1992).
17. Jäger, R. et al. Differential expression and biological activity of the heparin-binding growth-associated molecule (HB-GAM) in lung cancer cell lines. *Int. J. Cancer* **73**, 537–543 (1997).
18. Diamantopoulou, Z., Kitsou, P., Menashi, S., Courty, J. & Katsoris, P. Loss of receptor protein tyrosine phosphatase β/ζ (RPTPβ/ζ) promotes prostate cancer metastasis. *J. Biol. Chem.* **287**, 40339–40349. https://doi.org/10.1074/jbc.M112.405852 (2012).
19. Lu, K. V. et al. Differential induction of glioblastoma migration and growth by two forms of pleiotrophin. *J. Biol. Chem.* **280**, 26953–26964. https://doi.org/10.1074/jbc.M502614200 (2005).
20. Souttou, B. et al. Relationship between serum concentrations of the growth factor pleiotrophin and pleiotrophin-positive tumors. *J. Natl. Cancer Inst.* **90**, 1468–1473. https://doi.org/10.1093/jnci/90.19.1468 (1998).
21. Chatsirisupachai, K., Lagger, C. & de Magalhães, J. P. Age-associated differences in the cancer molecular landscape. *Trends Cancer* **8**, 962–971. https://doi.org/10.1016/j.trecan.2022.06.007 (2022).

22. Yoshihara, K. et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612. https://doi.org/10.1038/ncomms3612 (2013).
23. Bindea, G. et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* **39**, 782–795. https://doi.org/10.1016/j.immuni.2013.10.003 (2013).
24. Li, H. X., Wang, S. Q., Lian, Z. X., Deng, S. L. & Yu, K. Relationship between tumor infiltrating immune cells and tumor metastasis and its prognostic value in cancer. *Cells* https://doi.org/10.3390/cells12010064 (2022).
25. Li, Z., Wu, T., Zheng, B. & Chen, L. Individualized precision treatment: Targeting TAM in HCC. *Cancer Lett.* **458**, 86–91. https://doi.org/10.1016/j.canlet.2019.05.019 (2019).
26. Forner, A., Reig, M. & Bruix, J. Hepatocellular carcinoma. *Lancet* **391**, 1301–1314. https://doi.org/10.1016/s0140-6736(18)30010-2 (2018).
27. Liu, J. et al. Prediction of liver cancer prognosis based on immune cell marker genes. *Front .Immunol.* **14**, 1147797. https://doi.org/10.3389/fimmu.2023.1147797 (2023).
28. Wang, T. et al. Disulfidptosis classification of hepatocellular carcinoma reveals correlation with clinical prognosis and immune profile. *Int. Immunopharmacol.* **120**, 110368. https://doi.org/10.1016/j.intimp.2023.110368 (2023).
29. Li, J. et al. Preoperative prediction and risk assessment of microvascular invasion in hepatocellular carcinoma. *Crit. Rev. Oncol. Hematol.* **190**, 104107. https://doi.org/10.1016/j.critrevonc.2023.104107 (2023).
30. Woo, H. Y. et al. Lung and lymph node metastases from hepatocellular carcinoma: Comparison of pathological aspects. *Liver Int. Off. J. Int. Assoc. Study Liver* **42**, 199–209. https://doi.org/10.1111/liv.15051 (2022).
31. Viganò, L. et al. Liver resection for colorectal metastases after chemotherapy: impact of chemotherapy-related liver injuries, pathological tumor response, and micrometastases on long-term survival. *Ann. Surg.* **258**, 731–740. https://doi.org/10.1097/SLA.0b013e3182a6183e (2013).
32. Lim, K. C. et al. Microvascular invasion is a better predictor of tumor recurrence and overall survival following surgical resection for hepatocellular carcinoma compared to the Milan criteria. *Ann. Surg.* **254**, 108–113. https://doi.org/10.1097/SLA.0b013e31821ad884 (2011).
33. Yuan, K. et al. Long noncoding RNA TLNC1 promotes the growth and metastasis of liver cancer via inhibition of p53 signaling. *Mol. Cancer* **21**, 105. https://doi.org/10.1186/s12943-022-01578-w (2022).
34. Ma, X. et al. NOD2 inhibits tumorigenesis and increases chemosensitivity of hepatocellular carcinoma by targeting AMPK pathway. *Cell Death Dis.* **11**, 174. https://doi.org/10.1038/s41419-020-2368-5 (2020).
35. Liu, H. et al. Mime: A flexible machine-learning framework to construct and visualize models for clinical characteristics prediction and feature selection. *Comput. Struct. Biotechnol. J.* **23**, 2798–2810. https://doi.org/10.1016/j.csbj.2024.06.035 (2024).
36. Tsunedomi, R. et al. Elevated expression of RAB3B plays important roles in chemoresistance and metastatic potential of hepatoma cells. *BMC Cancer* **22**, 260. https://doi.org/10.1186/s12885-022-09370-1 (2022).
37. Jin, X. et al. A predictive model for prognosis and therapeutic response in hepatocellular carcinoma based on a panel of three MED8-related immunomodulators. *Front. Oncol.* **12**, 868411. https://doi.org/10.3389/fonc.2022.868411 (2022).
38. Zhao, M. et al. KHDRBS3 accelerates glycolysis and promotes malignancy of hepatocellular carcinoma via upregulating 14–3-3ζ. *Cancer Cell Int.* **23**, 244. https://doi.org/10.1186/s12935-023-03085-4 (2023).
39. Tanaka, K. et al. Enhanced expression of mRNAs of antisecretory factor-1, gp96, DAD1 and CDC34 in human hepatocellular carcinomas. *Biochim. Biophys. Acta* **1536**, 1–12. https://doi.org/10.1016/s0925-4439(01)00026-6 (2001).
40. Liu, Y. et al. DYNLL1 accelerates cell cycle via ILF2/CDK4 axis to promote hepatocellular carcinoma development and palbociclib sensitivity. *Br. J. Cancer* **131**, 243–257. https://doi.org/10.1038/s41416-024-02719-2 (2024).
41. Ju, L. L. et al. Effect of NDC80 in human hepatocellular carcinoma. *World J. Gastroenterol.* **23**, 3675–3683. https://doi.org/10.3748/wjg.v23.i20.3675 (2017).
42. Wu, H. et al. HMMR triggers immune evasion of hepatocellular carcinoma by inactivation of phagocyte killing. *Sci. Adv.* https://doi.org/10.1126/sciadv.adl6083 (2024).
43. Fu, J. et al. A novel DNA methylation-driver gene signature for long-term survival prediction of hepatitis-positive hepatocellular carcinoma patients. *Cancer Med.* **11**, 4721–4735. https://doi.org/10.1002/cam4.4838 (2022).
44. Zhang, J. et al. Serum biomarker status with a distinctive pattern in prognosis of gastroenteropancreatic neuroendocrine carcinoma. *Neuroendocrinology* **112**, 733–743. https://doi.org/10.1159/000519948 (2022).
45. Harman, D. Mutation, cancer, and ageing. *Lancet* **1**, 200–201. https://doi.org/10.1016/s0140-6736(61)91371-x (1961).
46. Quail, D. F. & Joyce, J. A. Microenvironmental regulation of tumor progression and metastasis. *Nat. Med.* **19**, 1423–1437. https://doi.org/10.1038/nm.3394 (2013).
47. Anderson, N. M. & Simon, M. C. The tumor microenvironment. *Curr. Biol. CB* **30**, R921-r925. https://doi.org/10.1016/j.cub.2020.06.081 (2020).
48. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502. https://doi.org/10.1038/nbt.3192 (2015).
49. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353-d361. https://doi.org/10.1093/nar/gkw1092 (2017).
50. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740. https://doi.org/10.1093/bioinformatics/btr260 (2011).
51. 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.* **102**, 15545–15550. https://doi.org/10.1073/pnas.0506580102 (2005).
52. Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* **28**, 1747–1756. https://doi.org/10.1101/gr.239244.118 (2018).
53. Zhu, Y., Yao, S. & Chen, L. Cell surface signaling molecules in the control of immune responses: a tide model. *Immunity* **34**, 466–478. https://doi.org/10.1016/j.immuni.2011.04.008 (2011).
54. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457. https://doi.org/10.1038/nmeth.3337 (2015).
55. Wang, L. et al. Long non-coding RNA MAPKAPK5-AS1/PLAGL2/HIF-1α signaling loop promotes hepatocellular carcinoma progression. *J. Exp. Clin. Cancer Res. CR* **40**, 72. https://doi.org/10.1186/s13046-021-01868-z (2021).

Acknowledgements

We would like to express our gratitude to all the participants and researchers who have provided algorithmic and database project support for this study.

Author contributions

Conceptualization, data curation, formal analysis, investigation, methodology, resources, visualization, writing-original draft, writing-review and editing, J.Z., Z.Z and C.Y.; Conceptualization, methodology, project administration, resources, supervision, writing-review and editing, Q.L. and T.S.; All uthors have read and approved the final manuscript.

Funding
Natural Science Basic Research Program of Shaanxi (Program: 2023-JC-YB-808).

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-91475-1.

Correspondence and requests for materials should be addressed to Q.L. or T.S.