Identification of key molecules in the formation of portal vein tumor thrombus in hepatocellular carcinoma based on single cell transcriptomics and *in vitro* experiments

Man Zhang[1]^, Chenglei Su[1], Xiaoyu Liu[1]^, Shuqun Hu[1,2], Xianliang Yan[1,3]

[1]Department of Emergency Medicine, The Affiliated Hospital of Xuzhou Medical University, Xuzhou, China; [2]Laboratory of Emergency Medicine, Second Clinical Medical College of Xuzhou Medical University, Xuzhou, China; [3]Department of Emergency Medicine, Suining People's Hospital, Xuzhou, China

Contributions: (I) Conception and design: M Zhang, X Yan; (II) Administrative support: X Yan, S Hu; (III) Provision of study materials or patients: M Zhang, X Yan; (IV) Collection and assembly of data: M Zhang, S Hu, C Su, X Liu; (V) Data analysis and interpretation: M Zhang, X Yan, S Hu, C Su, X Liu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Xianliang Yan, MD, PhD. Department of Emergency Medicine, The Affiliated Hospital of Xuzhou Medical University, No. 99 Huaihai West Road, Quanshan District, Xuzhou 221000, China; Department of Emergency Medicine, Suining People's Hospital, Xuzhou 221000, China. Email: 100002018002@xzhmu.edu.cn; Shuqun Hu, PhD. Department of Emergency Medicine, The Affiliated Hospital of Xuzhou Medical University, No. 99 Huaihai West Road, Quanshan District, Xuzhou 221000, China; Laboratory of Emergency Medicine, Second Clinical Medical College of Xuzhou Medical University, Xuzhou 221002, China. Email: hushuqun88@xzhmu.edu.cn.

Background: The presence of portal vein tumor thrombus (PVTT) is a significant indicator of advanced-stage hepatocellular carcinoma (HCC). Unfortunately, the prediction of PVTT occurrence remains challenging, and there is a lack of comprehensive research exploring the underlying mechanisms of PVTT formation and its association with immune infiltration.

Methods: Our approach involved analyzing single-cell sequencing data, applying high dimensional weighted gene co-expression network analysis (hdWGCNA), and identifying key genes associated with PVTT development. Furthermore, we constructed competing endogenous RNA (ceRNA) networks and employed weighted gene co-expression network analysis (WGCNA), as well as three machine-learning techniques, to identify the upstream regulatory microRNAs (miRNAs) and long non-coding RNAs (lncRNAs) of the crucial mRNAs. We employed fuzzy clustering of time series gene expression data (Mfuzz), gene set variation analysis (GSVA), and cell communication analysis to uncover significant signaling pathways involved in the activation of these important mRNAs during PVTT development. In addition, we conducted immune infiltration analysis, survival typing, and drug sensitivity analysis using The Cancer Genome Atlas (TCGA) cohort to gain insights into the two patient groups under study.

Results: Through the implementation of hdWGCNA, we identified 110 genes that was closely associated with PVTT. Among these genes, *TMEM165* emerged as a crucial candidate, and we further investigated its significance using COX regression analysis. Furthermore, through machine learning techniques and survival analysis, we successfully identified the upstream regulatory miRNA (*hsa-miR-148a*) and lncRNA (*LINC00909*) that targeted *TMEM165*. These findings shed light on the complex regulatory network surrounding *TMEM165* in the context of PVTT. Moreover, we conducted CIBERSORT analysis, which unveiled correlations between *TMEM165* and immune infiltration in HCC patients. Specifically, *TMEM165* exhibited associations with various immune cell populations, including memory B cells and CD8[+] T cells. Additionally, we observed implications for immune function, particularly in relation to immune checkpoints, within the context of HCC.

Conclusions: The regulatory axis involving *TMEM165*, *hsa-miR-148a*, and *LINC00909* emerges as a

---

^ ORCID: Man Zhang, 0009-0007-8361-3740; Xiaoyu Liu, 0009-0003-3855-7533.

crucial determinant in the development of PVTT in HCC patients, and it holds significant implications for prognosis. Furthermore, alterations in the *TMEM165/hsa-miR-148a/LINC00909* regulatory axis exhibit a strong correlation with immune infiltration within the HCC tumor microenvironment, leading to immune dysfunction and potential failure of immunotherapy.

Introduction

Primary liver cancer, with hepatocellular carcinoma (HCC) and intrahepatic cholangiocarcinoma as the main subtypes, is a prevalent gastrointestinal malignancy worldwide. HCC, responsible for 75–85% of primary liver cancer cases, is associated with a poor prognosis and is a leading cause of cancer-related deaths globally. In patients with cirrhosis and HCC, the development of portal vein tumor thrombus (PVTT) is a common occurrence (1-3). Although

---

Highlight box

Key findings
- We identified 110 genes that were closely associated with portal vein tumor thrombus (PVTT). Among these genes, *TMEM165* emerged as a crucial candidate, and we further investigated its significance using COX regression analysis. Furthermore, through machine learning techniques and survival analysis, we successfully identified the upstream regulatory microRNA (*hsa-miR-148a*) and long non-coding RNA (*LINC00909*) that targeted *TMEM165*. These findings shed light on the complex regulatory network surrounding *TMEM165* in the context of PVTT. Moreover, we conducted CIBERSORT analysis, which unveiled correlations between *TMEM165* and immune infiltration in hepatocellular carcinoma (HCC) patients.

What is known and what is new?
- PVTT is one of the late-stage indicators in HCC patients, and there is a clear awareness that this malignant tumor behavior can result in devastating consequences for individuals with HCC.
- We identified the *TMEM165/hsa-miR-148a/LINC00909* axis as a crucial pathway contributing to the formation of PVTT in HCC patients.

What is the implication, and what should change now?
- We provide a new perspective on the formation of PVTT in HCC patients. Based on this, people can explore the deeper level of PVTT formation.

---

advances in imaging technology have improved early detection of HCC, a considerable proportion of patients (12.5–39.7%) still present with portal vein invasion that goes undetected. PVTT is considered a distinct form of hematogenous metastasis of HCC, characterized by the invasive growth of HCC cells within the portal vein. This results in widespread tumor dissemination throughout the liver, elevated portal vein pressure, rupture of esophageal varices, reduced portal vein flow, ascites, jaundice, hepatic encephalopathy, and liver failure. Consequently, the prognosis for HCC patients deteriorates significantly upon the development of PVTT, with a median survival time of only 2.7–4 months in the absence of appropriate treatment. Although transcatheter arterial chemoembolization (TACE) has demonstrated some success in selected patients, the median survival period remains limited to 3.8–9.5 months. Monotherapy targeting PVTT in HCC patients has shown minimal improvement in prognosis, with response rates below 20% (4-7). Previous studies have implicated various molecules, such as chemokine receptors CXCR4, KiSS-1, matrix metalloproteinase-9, protein disulfide isomerase A6, and apolipoprotein AI, are involved in the development of PVTT (8-10). However, a comprehensive understanding of the precise underlying mechanisms of this lethal tumor thrombus is still lacking. Therefore, unraveling the molecular basis of PVTT formation is crucial for both predicting and devising effective treatments for PVTT in HCC patients.

MicroRNAs (miRNAs), a class of highly conserved and tissue-specific non-protein-coding short RNAs, exert control over gene expression by recognizing and binding to homologous sequences, thereby interfering with transcription, translation, or epigenetic processes. These small molecules have a wide range of target genes and influence virtually all genetic processes, including cell

cycle checkpoints, cell proliferation, and apoptosis (11-13). Initially, the post-transcriptional control of gene expression was believed to occur through the binding of miRNAs to the miRNA response element (MRE) on target mRNA, leading to translational inhibition or mRNA degradation. However, as researchers delve into in-depth transcriptome research, they discovered the presence of MRE not only exist in mRNA but also in other RNA types such as lncRNA, pseudogene, and circRNA. This implies that the same miRNA can interact with multiple types of RNA, resulting in a competitive relationship among different RNA molecules that bind to the same miRNA (14-16). These findings have led to the proposal of the competing endogenous RNA (ceRNA) hypothesis, which represents a novel regulatory mechanism in gene expression. CeRNA operates in conjunction with the miRNA regulatory network, expanding the regulatory network to encompass a broader range of genes and RNA types. Investigating gene function and regulation at a deeper level, including the involvement of ceRNA in both healthy and pathological conditions, is crucial for understanding various biological processes, such as cellular development and the molecular mechanisms underlying diseases (17,18).

The continuous progress in bioinformatics technologies and machine learning algorithms has enabled more precise evaluation of transcriptomics data. Taking advantage of these advancements, this study integrated diverse bioinformatics analysis techniques and machine learning algorithms. By analyzing transcriptome data from HCC, we seek to identify significant genes and ceRNA regulatory networks involved in the development of PVTT. Through this comprehensive approach, we aim to unravel the precise mechanisms underlying the formation of PVTT in HCC patients and provide novel insights for future studies in this field. We present this article in accordance with the TRIPOD reporting checklist (available at https://tcr.amegroups.com/article/view/10.21037/tcr-23-1589/rc).

Methods

Downloading and preliminary collation of single cell sequencing data

This study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). We obtained single-cell sequencing data (registration number GSE149614) from the Gene Expression Omnibus (GEO) database. The dataset consisted of one metastatic lymph node tissue, two PVTT tissues, eight normal tissues, and ten primary HCC tissues, totaling 21 tissue sequencing data. The initial processing of the upstream data was performed using Cell Ranger software (version 2.2.0). For subsequent analysis, we focused on two PVTT tissues and ten primary tumor tissues. The Seurat package (version 4.1.1) was employed for single-cell data processing. To ensure data quality and eliminate low-quality cells, we implemented several quality control measures. The following criteria were used: (I) cells expressing fewer than 500 or more than 6,000 genes were excluded; (II) the unique molecular identifier (UMI) count value of each cell sequencing had to exceed 1,000, and the top 3% of cells with the highest UMI count were removed; (III) the mitochondrial gene expression in each cell was required to account for less than 35% of the total gene expression, and the top 2% of cells with the highest mitochondrial gene expression were eliminated; (IV) the percentage of rRNA expression across all genes was determined, and the cells with the lowest and highest top 1% values were excluded. Since our study involved combining single-cell sequencing data from multiple samples, it was essential to account for batch effects. Cells from different chips, sequencing channels, or time points were classified into distinct groups. Batch effects can arise due to variations in experimental conditions, potentially impacting transcriptome measurements and cellular transcriptional changes. To mitigate batch effects and ensure robust downstream analysis, we utilized the harmony method implemented in the "harmony" package (version 0.1.0) to integrate and remove batch effects from the 12 samples. Given the inherent variability in each operational step, even when sequencing the same cell twice, differences in the counting depth may occur. To address this technical variation and prevent downstream analysis errors, we employed the Normalization function. This procedure adjusted the count data, enabling the comparison of relative gene expression abundance among cells.

Subgroup clustering and cell annotation

After the initial processing, the single-cell data still retained a high dimensionality. To reduce computational costs, minimize noise, and improve data visualization, we employed the FindVariableFeatures function to select 2,000 genes known as highly variable genes. These genes provided information on the data's variability and served as features for downstream analysis. To conform the gene expression z-scores to a Gaussian distribution, we

applied the ScaleData function. Next, we employed the principal component analysis (PCA) algorithm, a linear dimensionality reduction method, to map the expression matrix to a low-dimensional space. We sought to identify the optimal low dimension that could capture the biological morphology represented by the cell expression profiles while preserving all the data's information. For further dimensionality reduction, we utilized the uniform manifold approximation and projection (UMAP) method, a nonlinear dimensionality reduction technique. This step allowed us to map the multidimensional data to a two-dimensional space suitable for observation. To establish connections between units based on their shared overlap (Jaccard similarity) in the immediate neighborhood, we utilized the FindNeighbors function. This approach involved building a K-nearest neighbor (KNN) graph using Euclidean distance in the PCA space. The edge weights between units were refined according to their shared overlap.

To enhance modular functionality, we utilized the FindClusters function with a resolution parameter set to 0.5. This step facilitated the aggregation of cells into distinct clusters based on their similarities and differences, enabling further analysis and interpretation. To ensure reliable and accurate cell annotation, we combined results from both automatic and manual annotation approaches. For automatic annotation, we primarily relied on the "singleR" package (version 1.8.1), which predicted the potential cell types of each cell by comparing them to a reference transcriptome data set of pure cell types. Manual annotation was based on the results of differential analysis, which aimed to identify genes that were differentially expressed between subgroups and all other subgroups. We used the FindAllMarkers function with a filtering criterion of P value less than 0.05 for this analysis. To validate and supplement the automatic annotation results from the "singleR" package, we employed additional resources. These included the CellMarker database (19), the BMC Genome Biology online database, and an extensive literature search. By combining information from these sources, we obtained comprehensive annotation results for each cell cluster. To distinguish between benign and malignant cells, we calculated the copy number distribution of individual cells using the state-of-the-art Copykat technique. The subclonal structure was determined by combining the Bayesian method with hierarchical clustering. Additionally, we utilized the Gaussian mixture model (GMM) to calculate the variance of each cell population. High-confidence diploid cells were identified as the cell population with

the minimum estimated variance, based on stringent categorization criteria. Finally, hierarchical clustering of single-cell copy number data was performed to achieve the greatest separation between diploid normal cells and aneuploid tumor cells, providing insights into the genomic alterations associated with the tumor cells.

Weighted gene co-expression network analysis (WGCNA) of single cell sequencing

We implemented WGCNA for single-cell data using the "hdWGCNA" package (version 0.2.2) developed by Morabito *et al.* (20,21). This advanced package allowed us to construct a co-expression network across multi-scale cells and spatial hierarchies, identifying robust modules of interconnected genes and enabling WGCNA of single-cell sequencing data. To perform WGCNA, we first set up a Seurat object. We then used the KNN algorithm to identify similar cell groups that needed to be aggregated using hdWGCNA. The average or sum expression of these cells was calculated to generate a low sparse meta cell gene expression matrix. Next, we defined the cell type consisting of malignant cells using the SetDatExpr function and created an expression matrix. To determine the appropriate soft power threshold for building the co-expression network, we conducted parameter scans using the TestSoftPowers function. We evaluated the architecture of the resultant networks for different power values and selected the soft power threshold that maintained a robust gene-gene adjacency matrix while eliminating weak links. In this study, we chose the minimum soft power threshold of 0.8 or higher based on the scale-free topology model. Using the ConstructNetwork function, we built the co-expression network below the selected soft threshold. To identify module feature genes, we employed the ModuleEigengenes tool, which performed PCA on a subset of the gene expression matrix for each module. This allowed us to obtain the module feature genes [module eigengene (ME)] present in various modules. Additionally, we calculated the central gene feature score for each module using the Seurat or UCell algorithms with the help of the ModuleExprScore function. To visualize the association between modules, we utilized the ModuleCorrelogram tool, which represented the relationships among modules based on their hME (hub gene), ME, or hub gene scores. To identify essential modules, we employed three techniques. First, we conducted correlation analysis on the modules and identified the most crucial modules using the GetModuleTraitCorrelation

method. The PlotModuleTraitCorrelation function was used to visualize the results as a heatmap. Second, we compared the correlation between module membership and sample information, specifically differentiating primary tissue from PVTT tissue. Lastly, we employed random forest (RF) to assess the significance of all modules and "sample" data, ranking them based on the IncNodePurity value. The module with the highest-ranking value was considered the most relevant to PVTT.

Screening of PVTT-related genes (PRGs) based on The Cancer Genome Atlas (TCGA) cohort

For further screening of module genes, we utilized TCGA cohort. Specifically, we accessed the HCC data [TCGA-Liver Hepatocellular Carcinoma (TCGA-LIHC)] from the TCGA database, which provided Fragments Per Kilobase of Exon Model Per Million Mapped Fragments (FPKM) data for the TCGA cohort patients. We also processed the clinical data of the patients appropriately. To perform the analysis, we extracted the survival data from the TCGA cohort, including survival time and month. We excluded patients who were younger than 18 years old or had a survival time of fewer than 30 days. Next, we obtained the gene expression matrix of the TCGA cohort module and conducted univariate COX regression analysis on all variables. The genes that showed an association with prognosis in the univariate COX regression analysis were included in the subsequent multivariate COX regression study. In the multivariate COX regression analysis, we focused on the genes that were previously identified as significant predictors of prognosis. Among these genes, we specifically selected those that were associated with PVTT, forming a set of PRGs.

Construction of ceRNA network based on online database

Based on an online database, we constructed a ceRNA network. To identify potential interactions between the PRGs and miRNAs, we utilized the TargetScan database (accessed on 23 November 2022). Additionally, we searched the StarBase database (accessed on 23 November 2022) to identify likely target long non-coding RNAs (lncRNAs) of the miRNAs. To visualize the mRNA-miRNA-lncRNA ceRNA network, we employed Cytoscape v3.9.1, which allowed us to depict the interactions and relationships among these components.

Screening of PVTT-related miRNA and lncRNA based on WGCNA

To validate the predictions made by the TargetScan and StarBase databases, we divided the TCGA cohort into two groups based on the median expression value of the PRGs. The high PRG expression group and the low PRG expression group were compared to assess their association with the development of PVTT. The high-risk group demonstrated a significantly higher propensity for PVTT compared to the low-risk group. To identify miRNAs and lncRNAs that are significantly associated with PVTT, we performed WGCNA on the miRNA expression matrix and lncRNA expression matrix of the two patient groups. The WGCNA analysis was conducted using the "WGCNA" package (version 1.71). Initially, we calculated the median absolute deviation (MAD) for each gene and eliminated the 50% of genes with the smallest MAD. The remaining differentially expressed genes (DEGs) were used to construct a scale-free co-expression network. The adjacency degree was computed using a soft threshold power (β) derived from co-expression similarity, and the adjacency was transformed into a topological overlap matrix (TOM) to calculate gene dissimilarity and connectivity. Subsequently, we applied a dynamic tree-cutting function and hierarchical clustering to detect modules based on average linkage hierarchical clustering and TOM-based dissimilarity. Genes with similar expression profiles were grouped into gene modules. To merge modules for further analysis, we computed the dissimilarity of module feature genes and determined the appropriate cutting line in the module tree. To identify the most important regulatory axis in the ceRNA network, we employed various machine learning methods after the WGCNA screening. The following machine learning algorithms were used: (I) least absolute shrinkage and selection operator (LASSO): Based on the "glmnet" package (version 4.1.4), LASSO performed variable selection and regularization while fitting the generalized linear model; (II) support vector machine recursive feature elimination (SVM-RFE): using the "e1071" package (version 1.7.11), SVM-RFE employed a sequential backward selection approach to extract features from two different types of data based on the maximum margin principle; (III) RF: RF estimated the average contribution of each feature to each decision tree in the RF using the "randomForest" package (version 4.7.1.1). The features were ranked based on their contributions for further analysis.

Mfuzz pattern expression clustering and enrichment analysis

For gene expression or protein expression profile data processing, we utilized the "Mfuzz" package (version 2.54.0), which offers a clustering approach. This approach, known as fuzzy c-means (FCM) clustering, allowed us to cluster transcriptome and proteome data with both time series and non-time series features. It enabled the grouping of genes or proteins based on their similar expression patterns. To evaluate the disparities among different expression pattern clusters, we calculated scores using single-sample gene set enrichment analysis (ssGSEA) and compared them. Subsequently, we employed Pearson correlation analysis to determine the correlation between each clustering module and the PRGs. By examining the correlation coefficient and P value, we identified the gene module that exhibited the closest association with PRG. To gain insights into the gene functions of the identified modules, we performed Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses on the retrieved gene modules. This analysis helped us understand the biological processes and pathways associated with the identified modules. Additionally, we conducted gene set variation analysis (GSVA) on the TCGA cohort using PRG as a reference. We compared the results of the Mfuzz clustering module's enrichment analysis with the findings from the GSVA to validate the reliability of the enrichment analysis results. In summary, the Mfuzz package, through FCM clustering, ssGSEA scoring, and correlation analysis, allowed us to identify gene modules associated with PRG. Subsequent functional analysis provided insights into the biological functions of these modules, and the validation process enhanced the reliability of the enrichment analysis results.

Cell communication analysis

By employing cell communication analysis, we were able to identify aberrantly activated signaling pathways within the tumor microenvironment of primary tumors and PVTT. Cell communication refers to the transfer of information between cells, leading to a corresponding cellular response. In animals and plants, the primary mechanism of intercellular communication involves chemical signal molecules. Cell-cell communication mediated by ligand-receptor complexes plays a crucial role in coordinating various biological processes, such as development, differentiation, and inflammation. To infer and analyze intercellular interaction networks, we utilized the "CellChat" package (version 1.1.3). Our approach involved several steps. First, we identified ligands or receptors that were overexpressed in a given cell population. We then mapped gene expression data onto a protein-protein interaction (PPI) network. If either the ligand or receptor was found to be overexpressed, we detected the interaction between the overexpressed ligand and its receptor. Subsequently, we inferred the communication probability at the signaling pathway level by calculating the communication probability for all ligand-receptor interactions associated with each signaling pathway. This allowed us to quantify the likelihood of communication occurring via specific signaling pathways. To establish a cell communication network, we calculated the aggregation communication network between cells by considering the number of links or the aggregate communication probability. This network captured the overall intercellular communication patterns. Moreover, we constructed a cell communication network at the level of cell-specific ligand-receptor interactions and signaling pathways, providing insights into the precise interactions and communication events occurring among cells. In summary, our utilization of the "CellChat" package enabled the inference and analysis of intercellular interaction networks. This approach facilitated the identification of overexpressed ligands and receptors, as well as the evaluation of communication probabilities at the signaling pathway level. The resulting cell communication network provided a comprehensive view of intercellular communication within the tumor microenvironment.

Drug sensitivity analysis of patients in different risk groups

Using the previous grouping of patients, we proceeded to analyze the sensitivity of these patients to different medications. The assessment of drug sensitivity was conducted using the pRRophetic package (version 0.5) developed by Geeleher et al. in 2014 (22). The pRRophetic algorithm employs a ridge regression model that utilizes the gene expression profiles from the Genomics of Drug Sensitivity in Cancer (GDSC) cell line and the TCGA gene expression profiles. This model predicts the half-inhibitory concentration (IC50), which represents the drug concentration required to induce a 50% reduction in cell viability or cause 50% apoptotic cells. In our analysis, we selected clinical therapeutic medications for the treatment of HCC based on a combination of clinical experience and prior research from randomized controlled trials (RCTs).

This approach allowed us to consider established knowledge in the field and evidence from rigorous scientific studies when making decisions regarding medication choices for HCC treatment. By leveraging the pRRophetic package and considering both gene expression profiles and clinical knowledge, we aimed to gain insights into the medication sensitivity of patients within different risk groups. This analysis provided valuable information for understanding potential treatment responses and guiding personalized therapeutic strategies for HCC patients.

CIBERSORT identified immune infiltration patterns in HCC patients in different modes

The expression matrix of human immune cell subtypes was subjected to deconvolution using the CIBERSORT method. This method utilizes linear support vector regression to estimate the proportions of immune cell subtypes present in the expression matrix. The default gene expression feature set, LM22, consisting of 22 immune cell subtypes, was used as a reference dataset. By deconvolving the expression matrix, we were able to determine the levels of immune cell infiltration in the samples. To examine the differences in immune cell infiltration between the two patient groups, we applied the concept of differential analysis. By comparing the immune cell profiles of the two groups, we assessed the variations in immune function within the immunological milieu of the different patient groups. Additionally, we investigated the expression patterns of common immunological checkpoints in patients with different immune cell infiltration patterns. To forecast the responsiveness of patients with different immune profiles to immunotherapy, we utilized the Tumor Immune Dysfunction and Exclusion (TIDE) online database. TIDE provided a predictive framework for assessing the likelihood of response to immunotherapy based on the specific immune characteristics observed in patients. By integrating these analyses, we aimed to gain insights into the immune landscape and functionality in the two patient groups, as well as to provide predictions regarding the potential response to immunotherapy. This information has the potential to guide treatment decisions and improve patient outcomes in the context of immunotherapy for cancer.

Identification of different survival pattern subtypes of HCC patients based on regulatory axis

We stratified HCC patients into different survival subgroups based on the expression levels of mRNA, miRNA, and lncRNA. The median expression value was used as a threshold, where expression levels above the median were classified as high expression and levels below the median were classified as low expression. All HCC cases were assigned to one of the eight survival patterns, as listed below:

mRNAhigh/miRNAhigh/lncRNAhigh (mode 1);
mRNAhigh/miRNAlow/lncRNAhigh (mode 2);
mRNAlow/miRNAhigh/lncRNAhigh (mode 3);
mRNAlow/miRNAlow/lncRNAhigh (mode 4);
mRNAhigh/miRNAhigh/lncRNAlow (mode 5);
mRNAhigh/miRNAlow/lncRNAlow (mode 6);
mRNAlow/miRNAhigh/lncRNAlow (mode 7);
mRNAlow/miRNAlow/lncRNAlow (mode 8).

These eight categories represent different combinations of high and low expression for mRNA, miRNA, and lncRNA. To investigate the survival outcomes of patients with different HCC profiles, we utilized the Kaplan-Meier (KM) method to analyze the survival curves for each of the eight patient groups. This analysis allows us to assess the association between RNA expression patterns and patient survival, providing valuable insights into the prognosis and potential treatment strategies for different subgroups of HCC patients.

Online websites

❖ GEO: https://www.ncbi.nlm.nih.gov/geo/;
❖ CellMarker: http://xteam.xbio.top/CellMarker/;
❖ BMC Genome Biology: https://genomebiology.biomedcentral.com/;
❖ TCGA: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga;
❖ Targetscan: https://www.targetscan.org/vert_80/;
❖ Starbase: https://starbase.sysu.edu.cn/;
❖ CIBERSORTx: https://cibersortx.stanford.edu/;
❖ TIDE: http://tide.dfci.harvard.edu/.

Statistical analysis

All data processing in this study was conducted using R (version 4.1.3). Differential analysis between all experimental and control groups was performed using the Wilcoxon rank-sum test. Pearson correlation method was employed for correlation analysis on public databases. The significance level for all analyses was set at $P<0.05$, and Benjamini-Hochberg correction method was applied to

adjust P values for multiple comparisons.

Results

GSE149614 dataset revealed cellular components in primary and PVTT tissues of HCC patients

The flowchart outlining the study's process is presented in Figure S1. At the post-processing stage, single-cell sequencing samples from 12 HCC patients were processed and analyzed. The raw data consisted of 25,208 genes and 40,385 cells. After quality control, a Seurat object containing 36,241 cells and 25,208 genes was obtained (Figure S2A). For further investigation, the first 30 principal components were selected after performing PCA (Figure S2B). By reducing the dimensions of all cells using UMAP, 22 cell clusters were identified (*Figure 1A*). Among these clusters, 4,692 cells originated from PVTT tissues, while 31,549 cells originated from original tumor tissues (*Figure 1B*). Different cell clusters exhibited varying levels of gene expression. Clusters 1, 6, and 9 demonstrated higher gene expression levels compared to clusters 5, 8, and 11, while clusters 1, 6, and 11 displayed lower levels (*Figure 1C*). Based on the stromal cell marker gene "*MME*", the immune cell marker gene "*PTPRC*", and the tumor cell marker gene "*AFP*", the total cell clusters were roughly categorized (*Figure 1D*). Various annotation methods were employed to further identify the cell types within each cluster, leading to the preliminary identification of ten cell types (*Figure 1E*). The marker genes for CD4+ T cells were "*CD3D*" and "*CD4*", for CD8+ T cells were "*CD7*" and "*CD8A*", for regulatory T cells was "*FOXP3*", for plasma cells were "*IGHG1*" and "*CD79A*", for monocytes were "*CD14*" and "*CD68*", for Kupffer cells were "*CD68*" and "*VSIG4*", for stellate cells was "*ACTA2*", for hepatocytes were "*ARG1*" and "*ALB*", for bile duct cells was "*KRT19*", and for endothelial cells were "*CD34*" and "*BTNL9*". Aneuploid cells identified by the Copykat algorithm were similar to the cell clusters identified by the tumor marker gene "*AFP*" and were derived from hepatocytes (*Figure 1F*, Figure S2C). Detailed proportions of different clusters and cells across tissues and samples are presented in *Figure 1G-1L*.

Single-cell WGCNA considered module 23 to be most relevant for PVTT formation in HCC patients

We selected the optimal soft threshold of "9" to construct a single-cell co-expression network using hdWGCNA

(*Figure 2A*). By clustering the single-cell data, we specifically focused on tumor cell clusters (clusters 1, 4, 9, 10, 12, 14, 18, 19, and 21) identified by the Copykat technique for subsequent network analysis. Our goal was to identify important modules within HCC cells that contribute to the development of PVTT. After merging related modules, we obtained a total of 24 modules that played a crucial role in PVTT formation (*Figure 2B,2C*). The correlation between the modules was visualized using a Pearson analysis (*Figure 2D*). The enrichment analysis of module genes and module core genes within cell clusters is presented in *Figure 2E,2F*.

To determine the most relevant module for HCC cell types, we employed the author's method to assess the inter-module connections and selected one module for further investigation. Module 23 exhibited the highest correlation coefficient (cor =0.64) and was deemed the most pertinent module for HCC cell typology (*Figure 3A*). Subsequently, we evaluated the correlation between module 23 and the cell tissue sources (primary tumor and PVTT) using the conventional WGCNA module screening method. The analysis revealed that module 23 exhibited the strongest association with the cell tissue sources (*Figure 3B*). To validate these findings, we employed RF analysis. The metrics MeanDecreaseAccuracy and MeanDecreaseGini, which measure the importance of variables in RF models, consistently indicated that module 23 played a crucial role in identifying the cell tissue sources (*Figure 3C*). Thus, we concluded that module 23, identified through hdWGCNA, was the most significant module. The genes within this module likely influenced the aggressiveness of HCC primary tumors towards PVTT (*Figure 3D,3E*).

Further identification of TMEM165 as a PRG in HCC based on TCGA cohort

Module 23 consisted of 110 genes. To identify key core genes, we further validated these 110 genes in the TCGA cohort. Among them, ten genes displayed altered expression patterns in HCC tissues compared to normal liver tissues (*Figure 4A,4B*). Given the strong relationship between PVTT development and the prognosis of HCC patients, we focused on identifying prognostic genes among these ten genes. Through univariate COX regression analysis, four genes were initially identified as being associated with the prognosis of HCC patients: *TMEM165* (P<0.001), *RRAGD* (P=0.006), *FADS3* (P=0.01), and *MYPOP* (P=0.005) (*Figure*
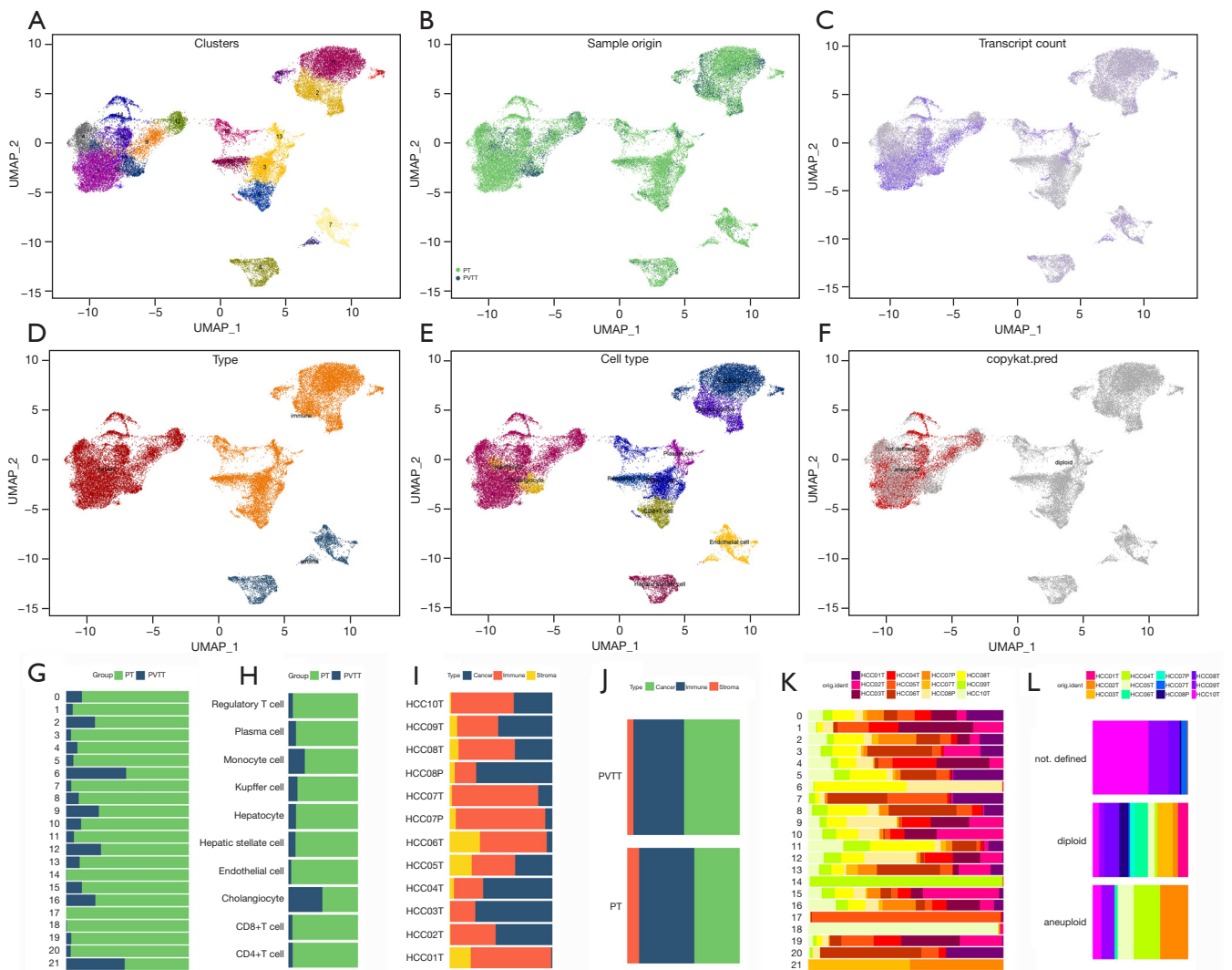
Figure 1 Results of single cell sequencing data analysis in HCC tissues. (A) Twenty-two cell clusters were divided after quality control and dimensionality reduction (PCA, UMAP). (B) Tissue types of all cell sources. (C) Gene expression per cell, the darker the color, the higher the cell expression. (D) Preliminary annotation of cell types in the tumor microenvironment based on general markers of immune cells, stromal cells, and tumor cells. (E) The annotation results of 22 cell clusters by multiple methods. (F) Copykat method for the identification of malignant cells and non-malignant cells. Red represents aneuploid cells, namely malignant cells. (G-L) The stacked bar graph shows the proportions of different cell compartments from different tissue/sample sources. (G) The proportion of 22 cell clusters in PT and PVTT tissues. (H) The proportion of 10 cell types in PT and PVTT tissues. (I) The proportion of three microenvironment components in 12 tissue samples. (J) The proportion of three microenvironment components in PT and PVTT tissues. (K) The proportion of 22 cell clusters in 12 tissue samples. (L) Proportion of aneuploid cells identified by Copykat in 12 tissue samples. UMAP, uniform manifold approximation and projection; PT, primary tumor; PVTT, portal vein tumor thrombus; HCC, hepatocellular carcinoma; PCA, principal component analysis.
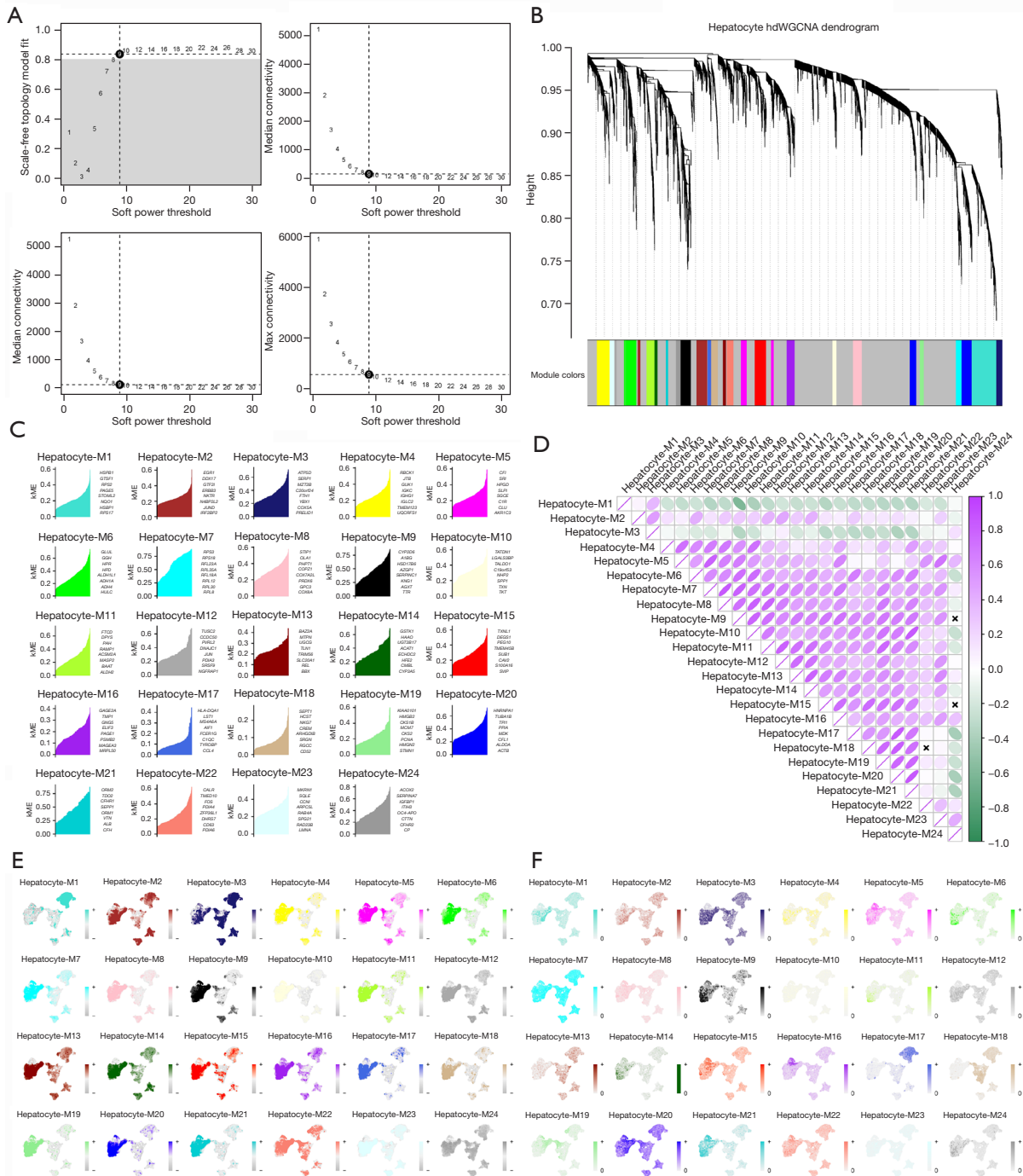
Figure 2 Co-expression network for single-cell sequencing using hdWGCNA. (A) Select the scale-free topology model to fit the lowest soft power threshold greater than or equal to 0.8, so that the constructed network is more consistent with the scale-free topology. (B) A co-expression network is constructed based on the optimal soft threshold, and a gene clustering tree is drawn after genes are divided into different modules. The upper part is the hierarchical clustering tree of genes, and the lower part is gene module, namely network module. (C) Calculate the feature-based gene connectivity (kME) of each gene in the co-expression network analysis to determine the highly connected genes (hub genes) in each module. (D) Correlation heat map between modules based on Pearson correlation analysis. X: no correlation. (E) Calculating gene scores for each module gene based on the UCell algorithm. (F) Gene scores are calculated for the central genes of each module based on the UCell algorithm. hdWGCNA, high dimensional weighted gene co-expression network analysis; kME, module membership.
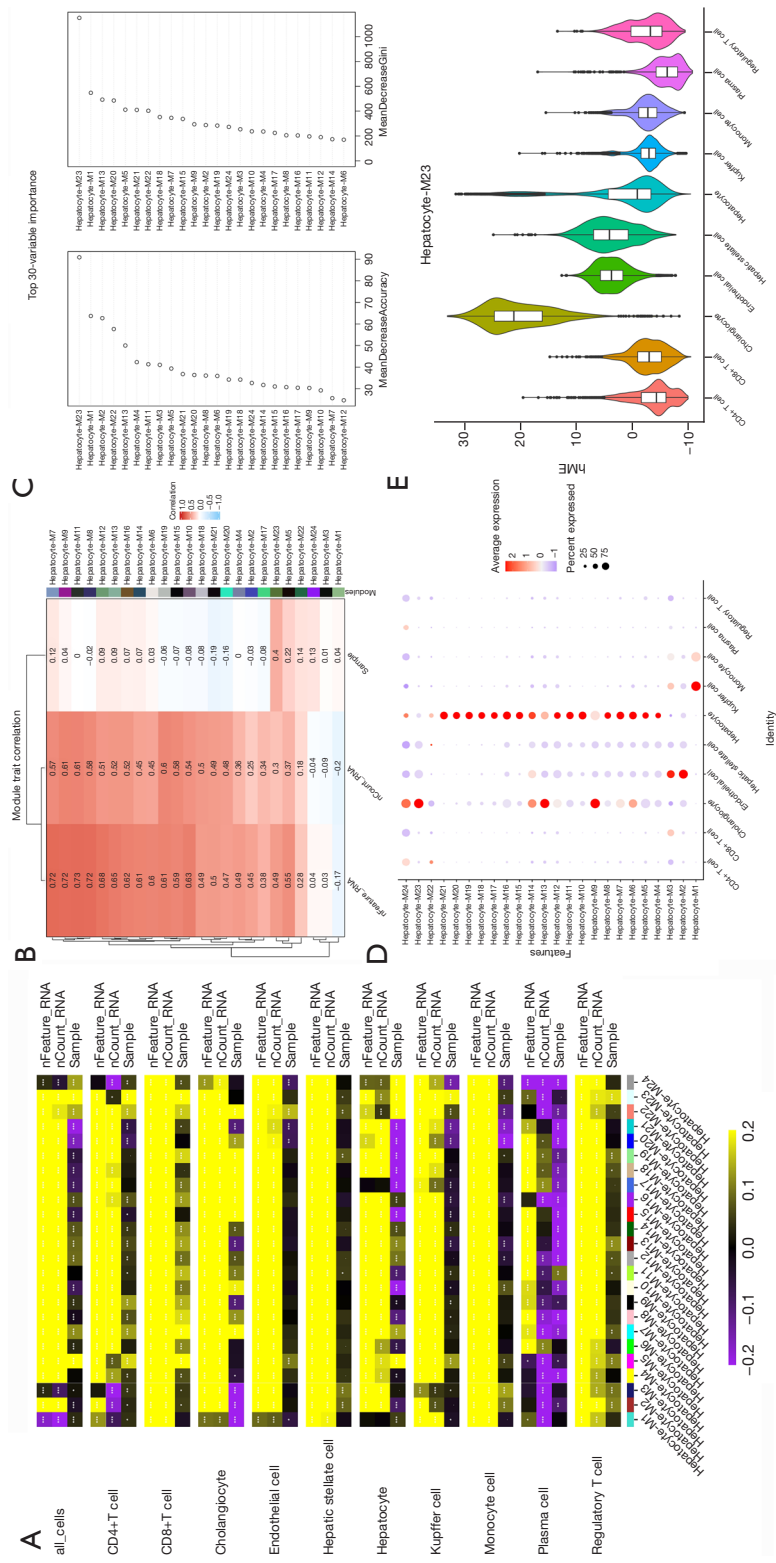
Figure 3 hdWGCNA identifies M23 as a key module in PVTT formation in HCC patients. (A) The correlation heatmap drawn by the PlotModuleTraitCorrelation in the hdWGCNA package. *, P<0.05; **, P<0.01; ***, P<0.001. (B) Correlation heat map based on traditional WGCNA method. (C) Importance screening of modules based on random forests. (D) The expression level of each module gene in various cells. (E) Expression level of module 23 in each cell type. hdWGCNA, high dimensional WGCNA; PVTT, portal vein tumor thrombus; HCC, hepatocellular carcinoma; WGCNA, weighted gene co-expression network analysis.
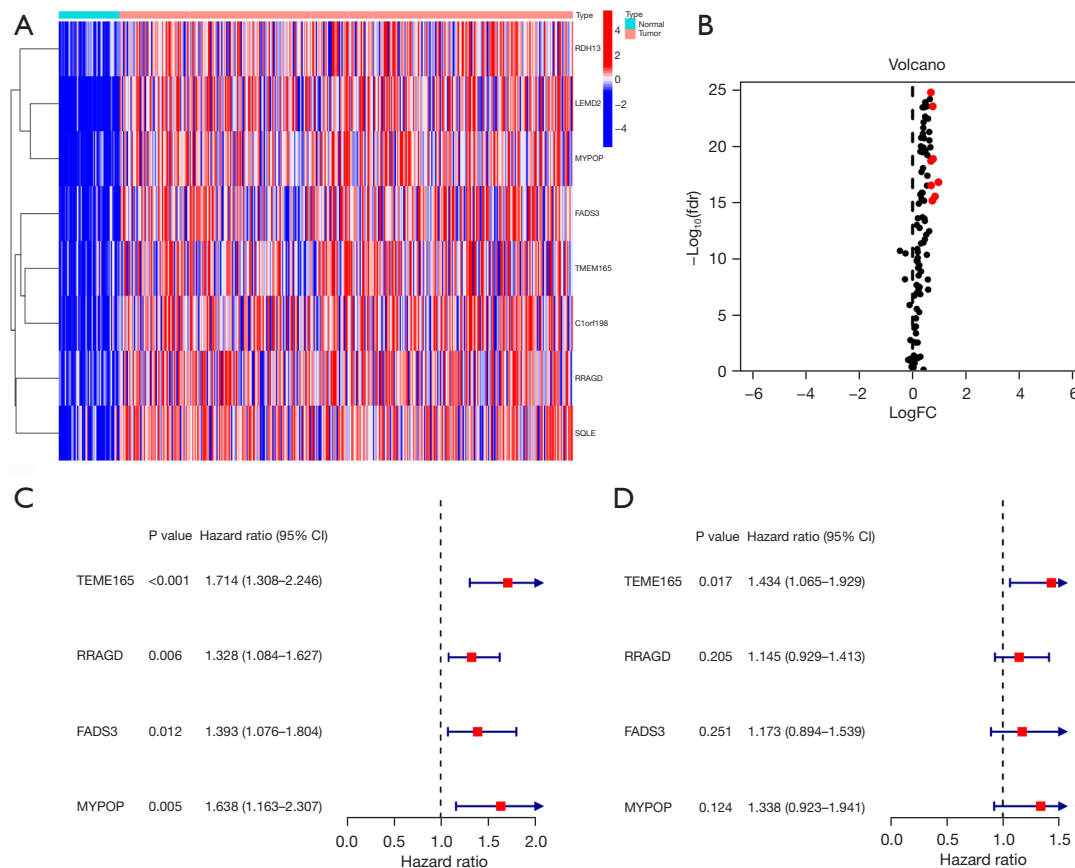
Figure 4 Differential analysis and univariate/multivariate COX regression for further screening of key module genes. (A) Gene heatmaps showing differential expression of 110 module genes in normal and tumor tissues (TCGA cohort). (B) The gene volcano plot of 110 module genes showing differential expression between normal and tumor tissues (TCGA cohort). Black dots represent genes with no significant changes. Red dots represent upregulated genes. (C) Univariate COX regression results of eight differentially expressed genes. (D) Multivariate COX regression results of four genes proved to have prognostic significance in univariate COX regression. fdr, false discovery rate; FC, fold change; CI, confidence interval; TCGA, The Cancer Genome Atlas.

4C). Subsequently, these four genes underwent multivariate COX regression analysis to eliminate the influence of other variables and to identify independent prognostic genes. Only *TMEM165* (P=0.02) was able to predict the prognosis of HCC patients and was deemed an independent prognostic factor for HCC patients according to the COX regression analysis (*Figure 4D*). As a result, *TMEM165* was identified as a PRG in HCC.

Construction of ceRNA network involving TMEM165 and screening of important mRNA-miRNA-lncRNA regulatory axis

Based on the Targetscan and StarBase databases, we initially identified potential upstream regulatory molecules of *TMEM165*. The Targetscan database revealed 92 upstream regulatory miRNAs for *TMEM165*, while the StarBase database provided 427 lncRNAs associated with these 92 miRNAs. The mRNA-miRNA-lncRNA regulatory network was visualized using the Cytoscape program (available online: https://cdn.amegroups.cn/static/public/tcr-23-1589-1.xlsx).

To identify key regulatory axes within the regulatory network, we divided the TCGA cohort into high-risk and low-risk groups based on the median expression of *TMEM165*. Subsequently, we conducted WGCNA on the miRNA and lncRNA expression matrices of the two patient groups. In the miRNA expression data, we constructed a co-expression network with a soft threshold of 26 (*Figure 5A,5B*) and identified modules associated with *TMEM165*. We
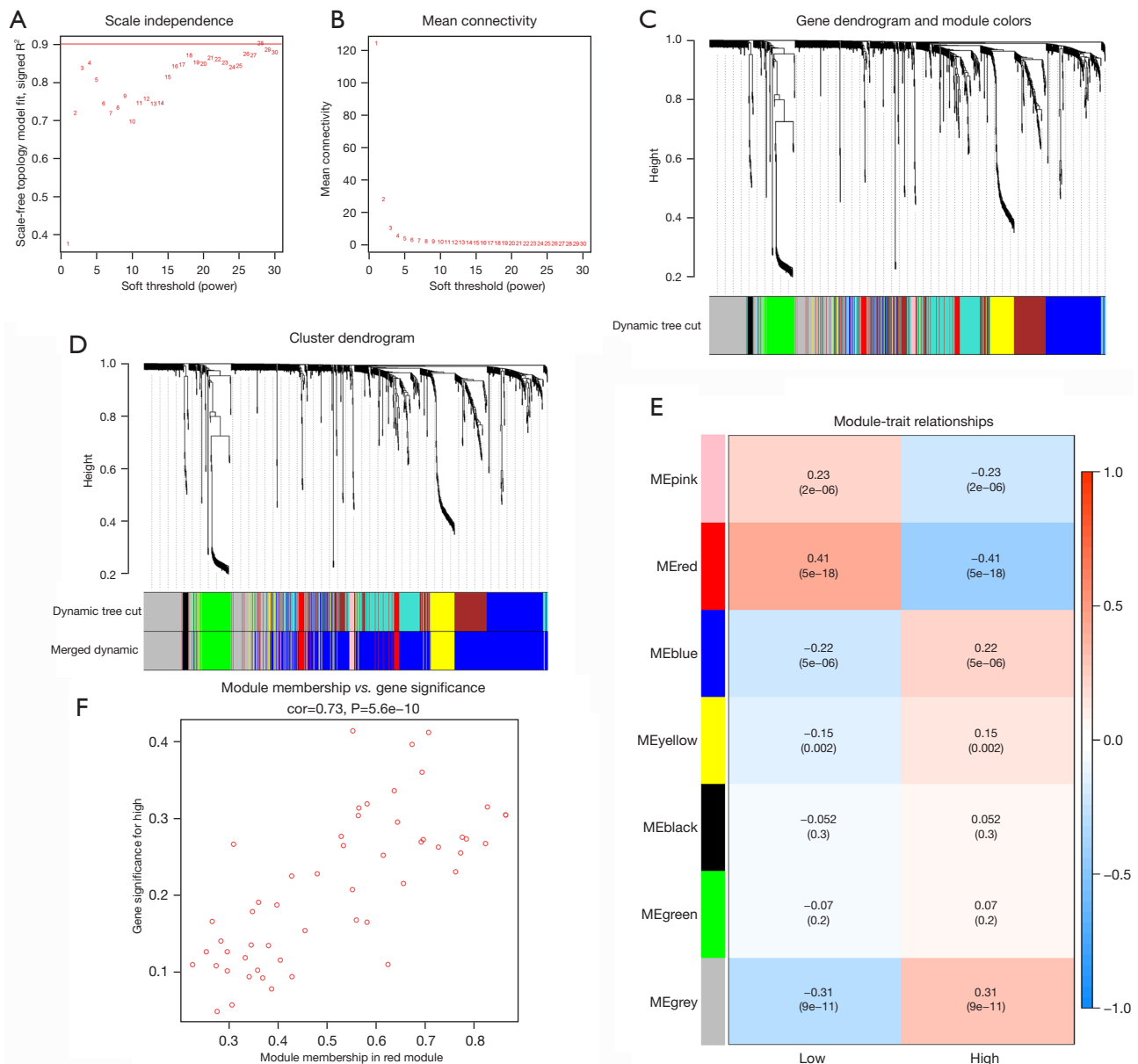
Figure 5 WGCNA recognized six miRNAs associated with *TMEM165*. (A,B) The optimal soft thresholding or power was determined to make the constructed network more consistent with the scale-free topology. (A) Scale-free fit index (y-axis) under different soft threshold (x-axis). The red line represents the subjectively selected scale-free fitting index value, which is 0.9 in this study. (B) Mean connectivity. (C) The co-expression network was constructed based on the optimal soft threshold, and the gene clustering tree was drawn after genes were divided into different modules. The upper part was the hierarchical clustering tree of genes, and the lower part was gene module, namely network module. (D) Gene clustering tree after merging similar modules. (E) Calculate the correlation and significance between the module and the expression level of *TMEM165*, and draw a correlation heat map. The first-row number in each module was the correlation coefficient, and the second-row number was the P value. Red represented positive correlation; blue represented negative correlation. (F) Scatter plot showed that there was a highly significant correlation between the GS of MM and the expression level of *TMEM165* in red module. WGCNA, weighted gene co-expression network analysis; GS, gene significance; MM, module member.
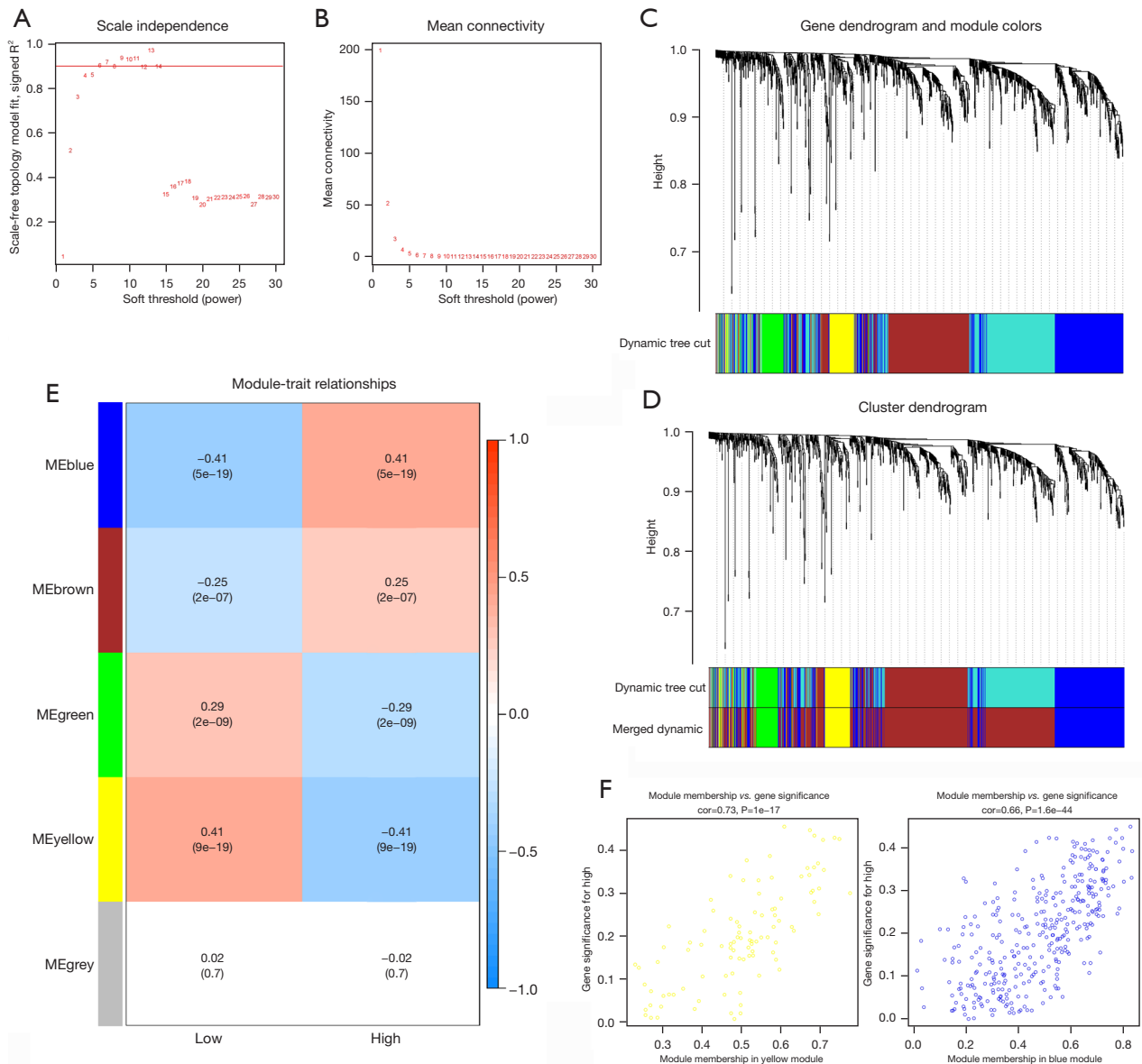
Figure 6 WGCNA recognized 36 lncRNA related with *TMEM165*. (A,B) The optimal soft thresholding or power was determined to make the constructed network more consistent with the scale-free topology. (A) Scale-free fit index (y-axis) under different soft threshold (x-axis). The red line represents the subjectively selected scale-free fitting index value, which is 0.9 in this study. (B) Mean connectivity. (C) The co-expression network was constructed based on the optimal soft threshold, and the gene clustering tree was drawn after genes were divided into different modules. The upper part was the hierarchical clustering tree of genes, and the lower part was gene module, namely network module. (D) Gene clustering tree after merging similar modules. (E) Calculate the correlation and significance between the module and the expression level of *TMEM165*, and draw a correlation heat map. The first-row number in each module was the correlation coefficient, and the second-row number was the P value. Red represented positive correlation; blue represented negative correlation. (F) The scatter plot showed that there was a highly significant correlation between the gene significance of module members and the expression level of *TMEM165* in yellow and blue modules. WGCNA, weighted gene co-expression network analysis.
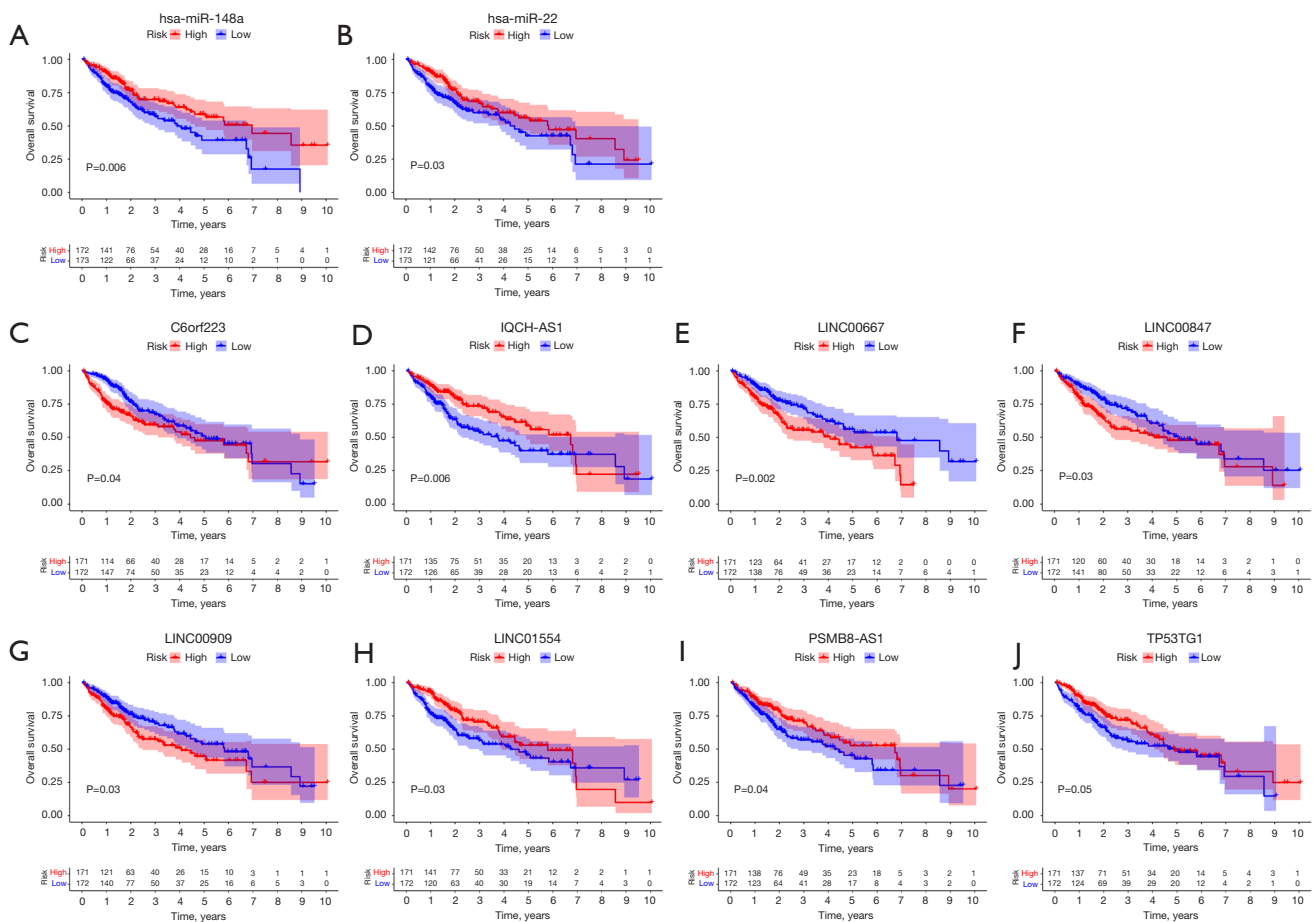
Figure 7 The survival analysis of candidate miRNA and lncRNA. (A,B) KM survival analysis was performed on two miRNAs. (A) *hsa-miR-148a*; (B) *hsa-miR-22*. (C-J) KM survival analysis was performed on eight lncRNAs. (C) *C6orf223*; (D) *IQCH-AS1*; (E) *LINC00667*; (F) *LINC00847*; (G) *LINC00909*; (H) *LINC01554*; (I) *PSMB8-AS1*; (J) *TP53TG1*. KM, Kaplan-Meier.

then merged related modules (*Figure 5C,5D*) and ultimately selected the red module (R=−0.41, P=5e−18) for further analysis (*Figure 5E,5F*). In the lncRNA expression data, we constructed a co-expression network with a soft threshold of 6 (*Figure 6A,6B*) and identified modules associated with *TMEM165*. Similar to the miRNA analysis, we merged related modules (*Figure 6C,6D*) and ultimately selected the blue module (R=0.41, P=5e−19) and the yellow module (R=−0.41, P=9e−19) for further analysis (*Figure 6E,6F*). By merging the WGCNA results of miRNA and lncRNA with the projected findings, we identified a total of 6 miRNAs and 36 lncRNAs.

To narrow down the key miRNAs and lncRNAs, a survival analysis using the KM approach was conducted. Among the six miRNAs, *hsa-miR-22* and *hsa-miR-148a* were found to be associated with the prognosis of HCC

patients (*Figure 7A,7B*). Similarly, eight out of the 36 lncRNAs (*C6orf223*, *IQCH-AS1*, *LINC00667*, *LINC00847*, *LINC00909*, *LINC01554*, *PSMB8-AS1*, and *TP53TG1*) were found to be related to the prognosis of HCC patients (*Figure 7C-7J*). To further analyze the remaining large quantity of lncRNAs, three machine-learning methods were employed. LASSO, SVM-RFE, and RF analyses classified five out of the eight lncRNAs as significant (*Figure 8A-8F*). By merging the screening results from the three different machine-learning approaches, three important lncRNAs were identified (*Figure 8G*). Finally, we determined the *TMEM165/hsa-miR-148a/LINC00909* axis as the most crucial regulatory axis in the regulatory network, playing a significant role in the formation of PVTT in HCC patients based on the miRNA-lncRNA correspondence in the ceRNA network.
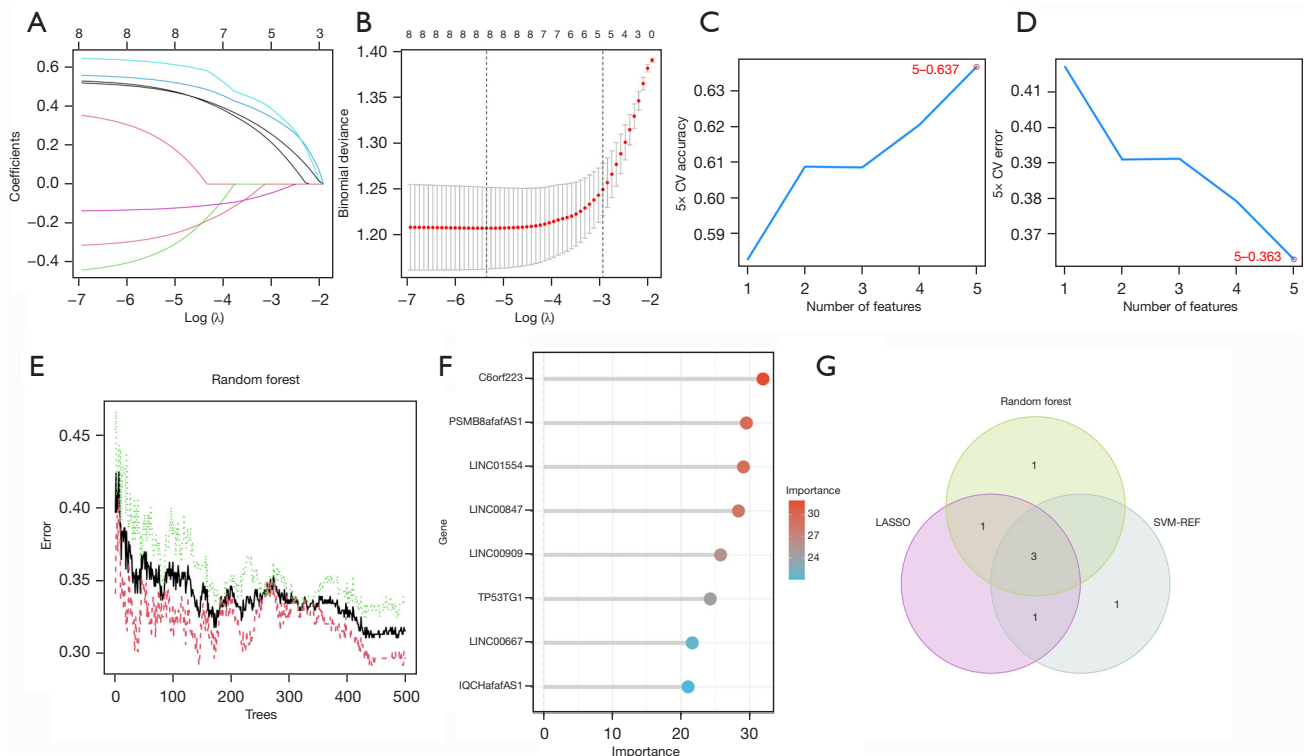
Figure 8 Identification of *LINC00909* as a *TMEM165* upstream targeting lncRNA based on multiple machine learning methods. (A,B) LASSO regression selected the number of variables filtered based on the minimum lambda value. Each colored line represents one gene out of eight genes. (C,D) SVM-RFE selected five variables. (E,F) The importance of eight lncRNAs was ranked by random forest. The green line represents the variation of the Out-Of-Bag error rate with the number of decision trees. The black line represents the mean Out-Of-Bag error. The red line represents the validation set error, i.e., how the model's error on the validation set changes with the number of decision trees. (G) The Venn diagram showed the common results of the three variable screening methods. CV, coefficient of variation; LASSO, least absolute shrinkage and selection operator; SVM-RFE, support vector machine-recursive feature elimination.

Various methods to determine that TMEM165 promotes PVTT progression through the Notch pathway

Using Mfuzz cluster analysis, we stratified the TCGA cohort into 50 subgroups based on the expression levels of *TMEM165* (*Figure 9A*). Through correlation and difference analyses, we determined that subgroup 22 was the most significant (*Figure 9B-9D*). To conduct a KEGG enrichment analysis, we retrieved all the genes (n=624) from subgroup 22. The enrichment analysis revealed that these 624 genes were primarily associated with the FoxO signaling pathway, PPAR signaling system, Hippo signaling pathway, Notch signaling pathway, and pentose phosphate pathway (*Figure 9E*). Additionally, we performed GSVA analysis directly on the TCGA cohort using *TMEM165* as the input. The GSVA results indicated that the neurotrophin signaling pathway, NOTCH signaling

pathway, MTOR signaling pathway, NOD-like receptor signaling pathway, adipocytokine signaling pathway, T cell receptor signaling pathway, and PPAR signaling pathway were significantly influenced by *TMEM165* expression (*Figure 9F*). Subsequently, we conducted cell communication analysis to identify differentially expressed signaling pathways between primary tumor tissues and PVTT tissues using single-cell sequencing data. In the tumor microenvironment, intercellular communication was found to be notably stronger in the PVTT group compared to the PT group (*Figure 10A,10B*). However, the changes in signal communication of HCC cells within tumors were not as significant as those observed in other cell types, such as hepatic stellate cells (*Figure 10C*). We analyzed the differences in important signaling pathways between the two groups, and the results are depicted in *Figure 10D,10E*. Combining the findings from the three pathway analyses,
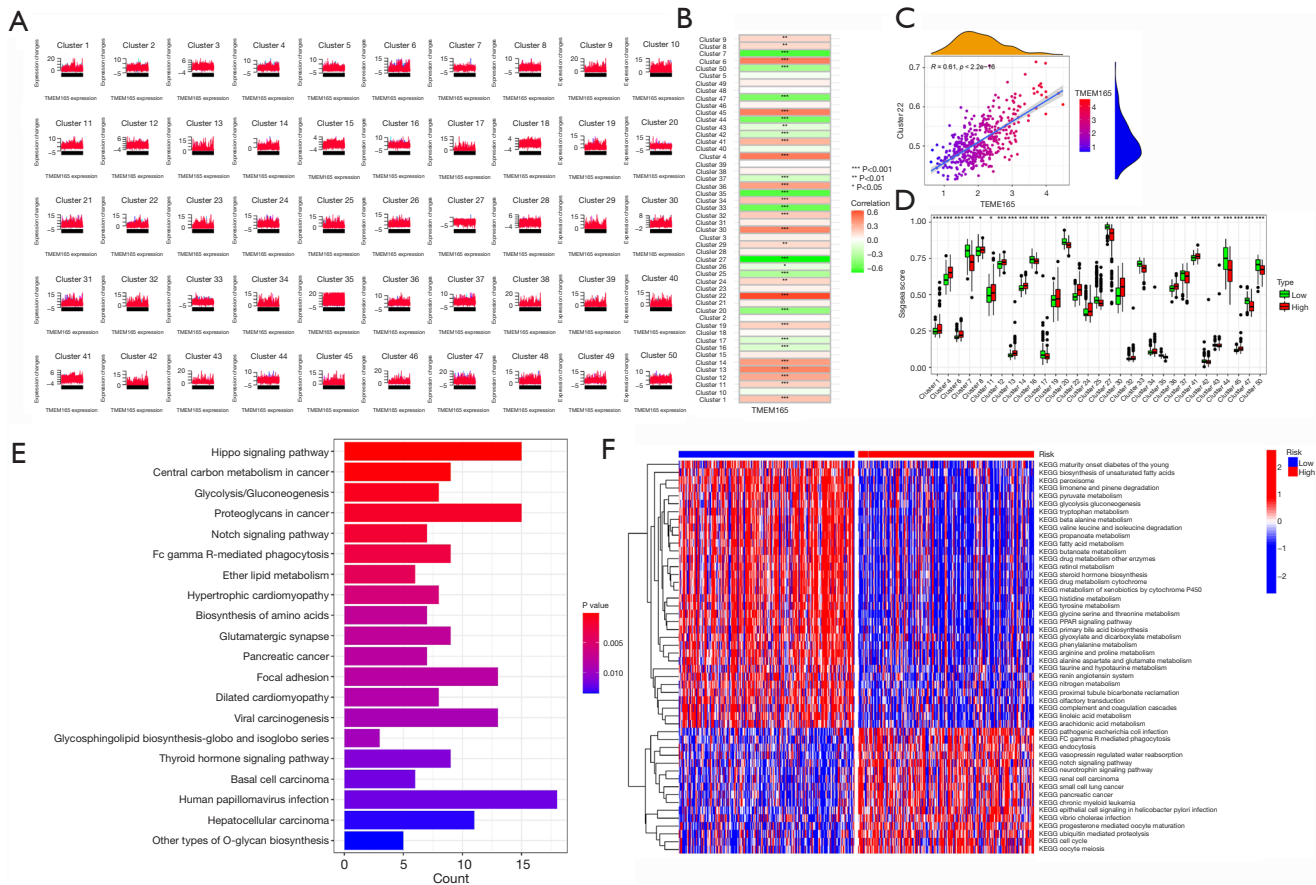
Figure 9 GSVA analysis based on Mfuzz cluster analysis identified *TMEM165* involvement in Notch signaling pathway activation. (A) Mfuzz cluster analysis extracted the characteristics of all genes and divided them into 50 subgroups. (B) The correlation analysis between 50 subgroups and *TMEM165* expression. (C) The scatter plot between subgroup 22 and *TMEM165* expression. (D) The subgroups with differences between the two groups were displayed. *, P<0.05; **, P<0.01; ***, P<0.001. (E) The results of KEGG enrichment analysis of the genes in subgroup 22. (F) The expression matrix was subjected to GSVA results directly according to the expression of *TMEM165*. GSVA, gene set variation analysis; Mfuzz, fuzzy clustering of time series gene expression data; KEGG, Kyoto Encyclopedia of Genes and Genomes.

we identified the Notch signaling pathway as the sole pathway of significance. Therefore, we concluded that *TMEM165*'s involvement in the Notch signaling pathway played a critical role in triggering PVTT development.

Susceptibility response revealed differences in efficacy of different drugs for patients in different risk groups

From the intersection of pharmaceuticals available in the pRRophetic package and commonly used drugs in clinical practice, we selected eight medications for a drug sensitivity study. These medications included brivanib, camptothecin, cisplatin, doxorubicin, erlotinib, gemcitabine, sorafenib, and sunitinib. Through sensitivity analysis, we found

variations in drug efficacy between high-risk and low-risk individuals. Erlotinib demonstrated higher effectiveness in high-risk patients (*Figure 11A*), while gemcitabine, cisplatin, camptothecin, doxorubicin, and brivanib were more effective in low-risk patients (*Figure 11B-11F*). On the other hand, there were no significant differences in treatment response between the two risk groups for brivanib, sorafenib, and sunitinib (*Figure 11G,11H*).

TMEM165 revealed two different immune infiltration patterns in HCC patients

Using the CIBERSORT deconvolution method, we investigated immune infiltration in the TCGA cohort and
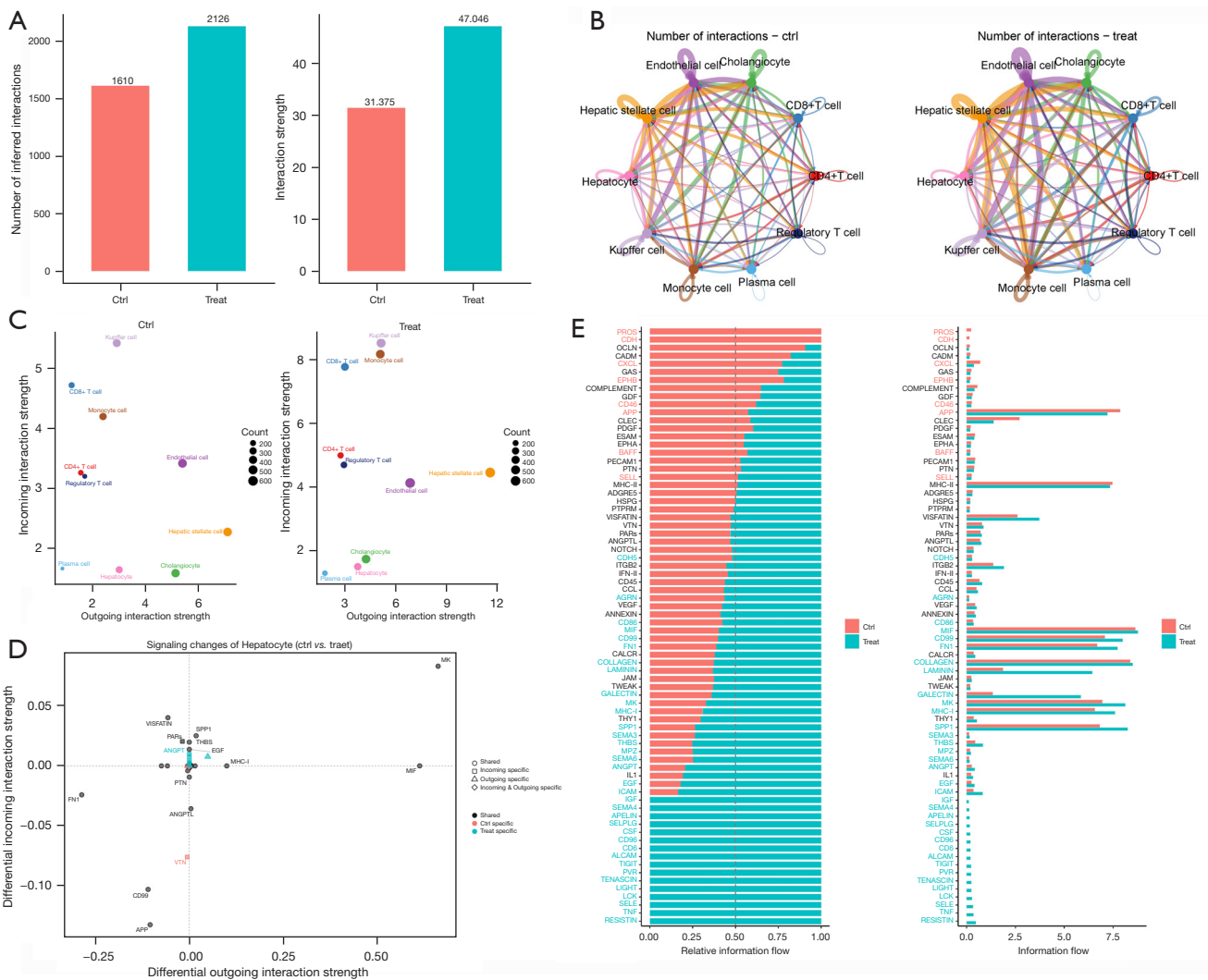
Figure 10 Notch signaling pathway involved in *TMEM165* demonstrated by cell communication analysis. (A) The tumor between PT group and PVTT group was the comparison of the number and intensity of cell communication in patients (bar plot). (B) The tumor between PT group and PVTT group was a network of cell communication in patients. (C) The tumor between PT group and PVTT group was the comparison of cell communication intensity in patients in two-dimensional coordinate system. (D) Difference analysis of significantly activated signal pathways between PT group and PVTT group. (E) Display of differential signaling pathways. ctrl, PT; treat, PVTT; PT, primary tumor; PVTT, portal vein tumor thrombus.

identified 22 immune cell types that had infiltrated the tumor microenvironment (*Figure 12A*). Upon performing differential analysis and excluding cells with low expression, we observed distinct patterns of immune cell infiltration between the low-risk and high-risk groups. In the low-risk group, CD8+ T cells, gamma delta T cells, and M2 macrophages were found to be highly infiltrated, while the high-risk group exhibited significant infiltration of activated CD4+ memory T cells, M0 macrophages, resting dendritic

cells, and neutrophils (*Figure 12B*). These immune cell populations may play important roles in the development of PVTT. Additionally, the two groups displayed differences in various immunological functions, with higher immune checkpoint activity observed in the high-risk group compared to the low-risk group (*Figure 12C*). Subsequently, we examined the expression of immunological checkpoints in the two patient groups and found that patients in the high-risk group exhibited overall higher expression of
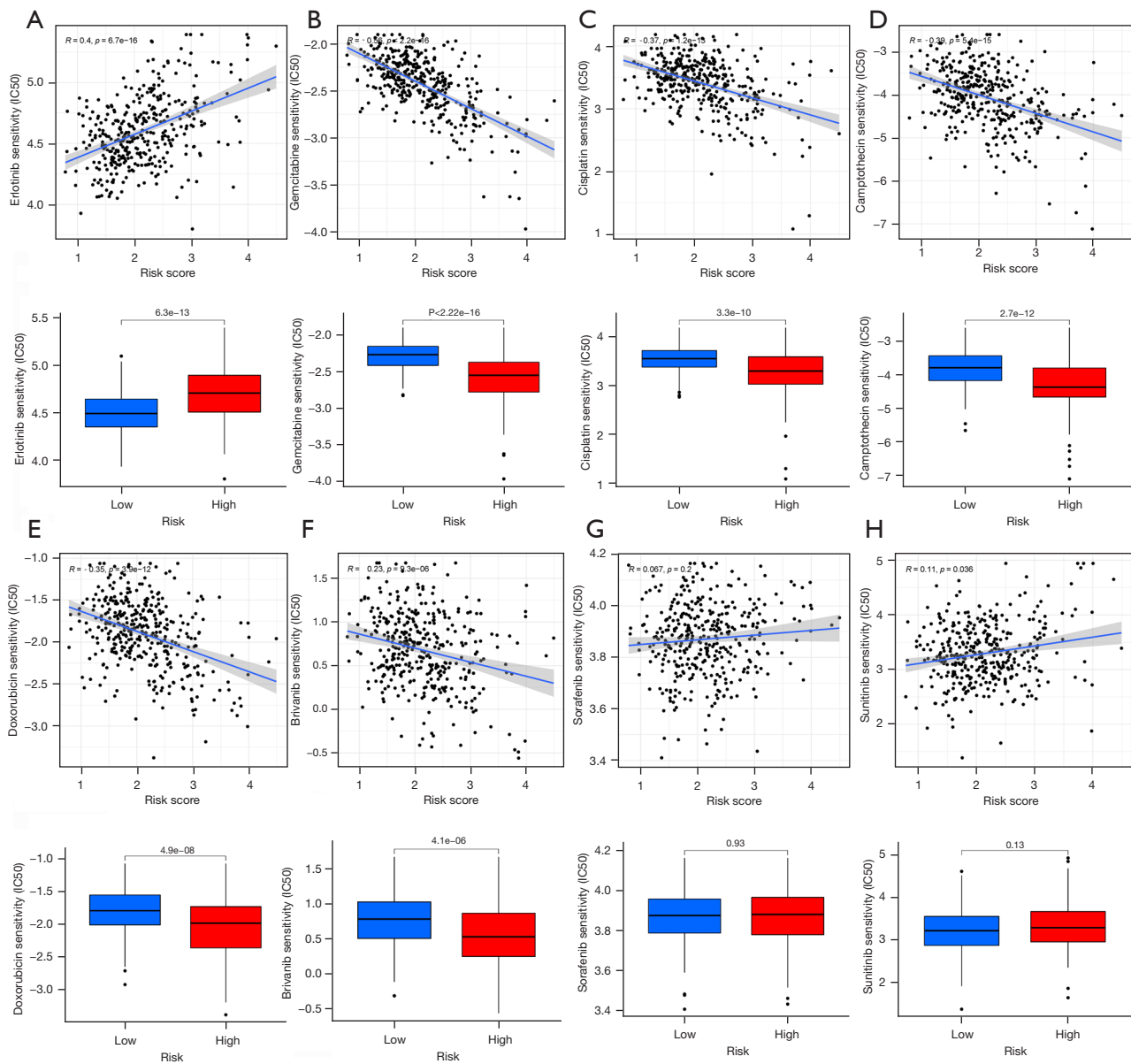
Figure 11 Drug sensitivity analysis of eight common drugs. (A-H) Drug correlation and sensitivity. (A) Erlotinib; (B) gemcitabine; (C) cisplatin; (D) camptothecin; (E) doxorubicin; (F) brivanib; (G) sorafenib; (H) sunitinib. IC50, half-inhibitory concentration.

immune checkpoints, including well-known ones such as CTLA4, CD28, and PDCD1 (*Figure 12D*). This suggests that patients at high-risk experience more pronounced immunosuppression. Finally, we predicted the responsiveness of the two patient groups to immunotherapy and found that the high-risk group may exhibit less favorable response to immunotherapy compared to the low-risk group (*Figure 12E*).

TMEM165/hsa-miR-148a/LINC00909 axis for HCC survival pattern classification

Despite the association between the three RNA molecules of the *TMEM165/hsa-miR-148a/LINC00909* axis and patient prognosis, their individual impact differed. *Hsa-miR-148a* exhibited a positive correlation with patient prognosis, whereas *TMEM165* and *LINC00909* showed
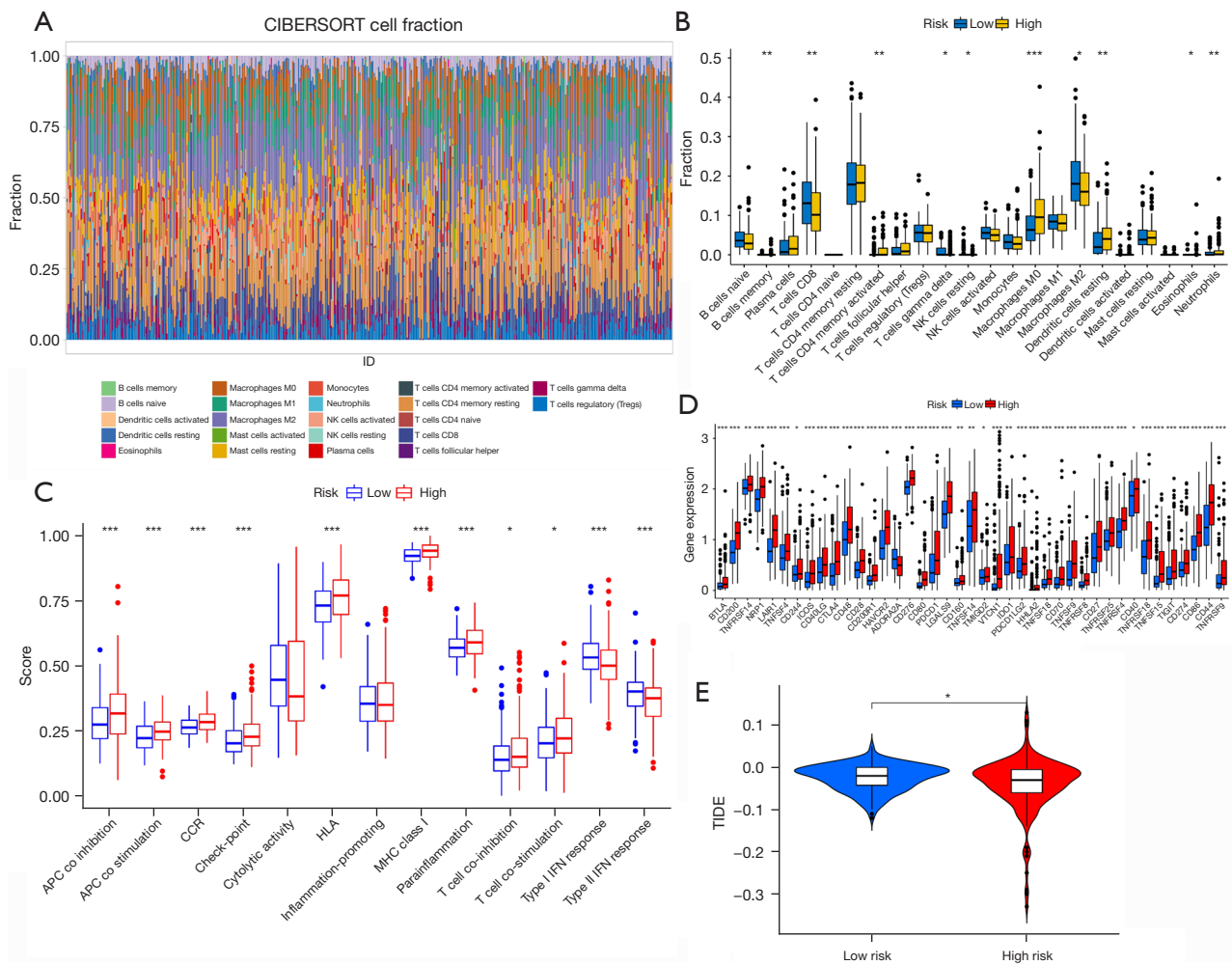
Figure 12 Changes of immune microenvironment in patients with high risk of PVTT formation. (A) The infiltration of immune cells in tumor microenvironment of HCC patients was demonstrated based on CIBERSORT deconvolution method. (B) The box plot qualitatively analyzed the difference in infiltration of immune cells in 22 high-risk groups and low-risk groups. *, P<0.05; **, P<0.01; ***, P<0.001. (C) Identify differences in immune function between high-risk and low-risk groups. *, P<0.05; ***, P<0.001. (D) Identify different immune checkpoints in the high-risk group and the low-risk group. *, P<0.05; **, P<0.01; ***, P<0.001. (E) Predict whether there was a difference between high-risk and low-risk groups for immunotherapy. *, P<0.05. NK, natural killer; APC, antigen-presenting cell; CCR, chemokine receptor; HLA, human leukocyte antigen; MHC, major histocompatibility complex; IFN, interferon; TIDE, Tumor Immune Dysfunction and Exclusion; HCC, hepatocellular carcinoma; PVTT, portal vein tumor thrombus.

negative correlations. In light of these findings, we devised an innovative analytical model to explore the relationship between different combinations of the three RNA expression patterns and overall survival (OS) in patients with HCC. The results revealed significant variations among the eight models generated. HCC patients in models 2 and 4 experienced poorer prognosis, as did those in models 7 and 3 (*Figure 13A*). The survival analysis heatmap exhibited notable P values for several models (*Figure 13B*). Overall,

our classification results aligned with the conclusions drawn from our investigation, affirming the validity and accuracy of the identified RNA molecules.

Discussion

HCC is a highly aggressive gastrointestinal tumor known for its propensity to invade the liver vascular system. Once tumor cells infiltrate the main portal vein or its
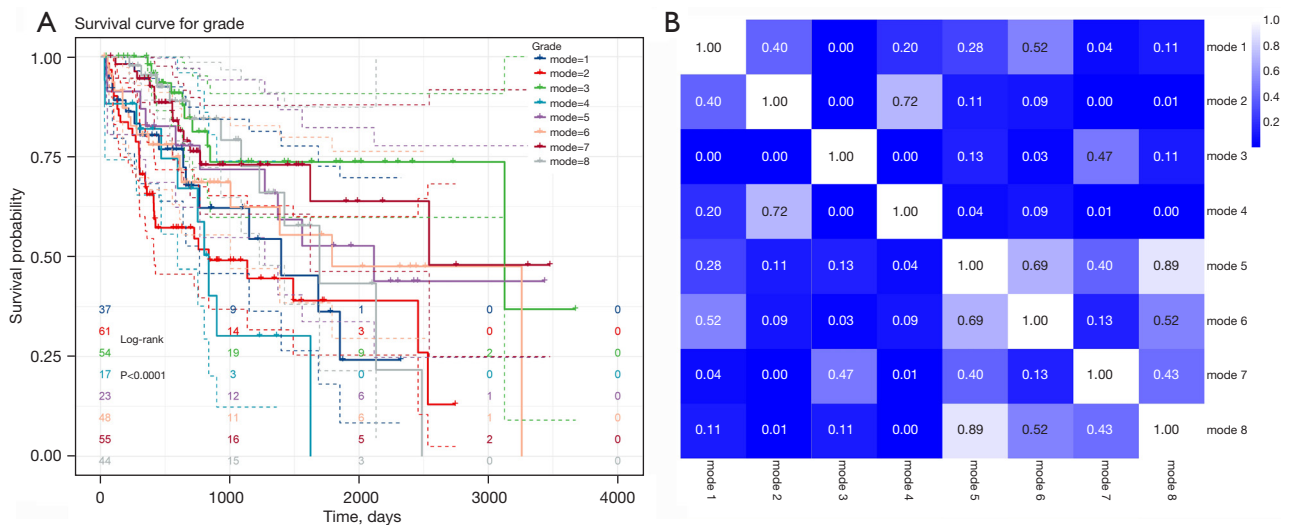
Figure 13 A new survival classification based on the *TMEM165/hsa-miR-148a/LINC00909* axis. (A) The KM analysis and log-rank test results of eight survival subtypes. (B) The heatmap of P value of pairwise comparison of eight survival subtypes. Mod1: mRNAhigh/miRNAhigh/lncRNAhigh (n=37); mod2: mRNAhigh/miRNAlow/lncRNAhigh (n=61); mod3: mRNAlow/miRNAhigh/lncRNAhigh (n=54); mod4: mRNAlow/miRNAlow/lncRNAhigh (n=17); mod5: mRNAhigh/miRNAhigh/lncRNAlow (n=23); mod6: mRNAhigh/miRNAlow/lncRNAlow (n=48); mod7: mRNAlow/miRNAhigh/lncRNAlow (n=55); mod8: mRNAlow/miRNAlow/lncRNAlow (n=44). KM, Kaplan-Meier.

branches, the hepatic vein or its branches, or the inferior vena cava within the liver, HCC patients enter the phase of macrovascular invasion (MVI) (23). MVI serves as a strong indicator of poor prognosis and advanced-stage HCC (24,25). Among the various forms of MVI, PVTT is the most prevalent and severe type (26). The formation of PVTT in HCC primary tumor cells typically involves multiple stages, including cellular invasion acquisition, anti-apoptosis within the extracellular matrix (ECM), remodeling and migration, infiltration, and growth along the portal vein (27). Transcriptomic alterations in HCC cells play a critical role in the entire process of PVTT formation, particularly during the initial stage. It is evident that changes in cancer-related genes associated with PVTT, such as RMP and CXCR4 (9,28), as well as non-coding RNAs like MiR-135a and ICAM-1-related non-coding RNA (29,30), directly facilitate PVTT development. Moreover, interactions between external factors like the immunosuppressive microenvironment, hypoxia, hepatitis virus infection (31-33), and the tumor's interactions with normal cells/components (e.g., endothelial cells, platelets, neutrophils, etc.) (34-36), can also contribute to genetic alterations in HCC cells. Therefore, further exploration of the transcriptomic differences between PVTT cells and primary tumor cells holds great significance in elucidating

the mechanism of PVTT formation, improving clinical treatments, and enhancing the prognosis for HCC patients.

hdWGCNA is a versatile R package developed by Morabito *et al.* (21) from the Swarup laboratory, specifically designed for conducting WGCNA on high-dimensional transcriptomic data, including single-cell RNA-seq or spatial transcriptomics. In the module screening process, in addition to the correlation-based method proposed by Morabito *et al.*, we also employed the RF method to validate the module screening results. The modules that demonstrate consistency between the two analysis methods are considered the most reliable modules in the process of PVTT formation. Since PVTT formation significantly impacts the prognosis of HCC patients, it is essential for each RNA in the mRNA/miRNA/lncRNA regulatory axis to be associated with HCC survival. Therefore, whether through hdWGCNA, subsequent WGCNA, or machine learning-based variable selection, the results should align with COX regression or KM analysis. To ensure rigor in the screening process, the identification of miRNA and lncRNA cannot solely rely on their correlation with *TMEM165* or the presence of binding sites with *TMEM165* in online databases. Both conditions must be met to establish them as upstream regulatory RNAs of *TMEM165*.

TMEM165 is a transmembrane protein localized in

the Golgi apparatus, where it contributes to maintaining Golgi ion homeostasis and vesicle trafficking. Initially, *TMEM165* mutations were identified in individuals with a congenital glycosylation disorder. Subsequent investigations have suggested a potential association between *TMEM165* expression and aggressive behavior in certain tumors (37,38). In a study by Lee *et al.* conducted in 2018 (39), reverse transcription polymerase chain reaction (RT-PCR) analysis was performed on tumor and normal tissues obtained from 88 patients with HCC. The results confirmed the high expression of *TMEM165* in HCC tissues. Furthermore, statistical analysis revealed a correlation between *TMEM165* overexpression and macroscopic vascular invasion as well as serosal invasion in HCC patients. Through experiments conducted on HCC cell lines, the researchers demonstrated that the depletion of *TMEM165* reduced the invasive activity of cancer cells by inhibiting the expression of matrix metalloproteinase 2 (MMP-2). Our study not only validated their findings but also identified *TMEM165* as a key gene involved in the formation of PVTT. Additionally, we identified upstream regulatory non-coding RNAs for *TMEM165*. Previous research has indicated that *hsa-miR-148a* is associated with the development and progression of various tumors (40,41), as well as non-neoplastic diseases (42,43). In the context of liver diseases, *hsa-miR-148a* has been specifically linked to alcohol-induced liver injury (44). On the other hand, *LINC00909* is a recently discovered lncRNA with an unknown mechanism of action. Nonetheless, evidence suggests that *LINC00909* is implicated in the progression of several neoplastic diseases, including colon cancer (45), ovarian cancer (46), and glioma (47).

In order to elucidate the signaling pathways associated with *TMEM165*, we performed a comprehensive analysis utilizing three methods and two types of datasets. To ensure result comparability, we utilized reference datasets based on previous literature and the KEGG database for enrichment analysis following Mfuzz clustering, GSVA, and cell communication analysis. By employing Mfuzz cluster analysis based on the enrichment analysis of DEGs, we were able to identify the most relevant subgroup closely associated with *TMEM165* from multiple subgroups. The Notch signaling pathway is an evolutionarily conserved cascade that plays a crucial role in normal biological processes such as cell differentiation, development, and homeostasis. Dysregulation of the Notch pathway has been associated with tumor progression and represents a promising target for cancer therapy (48,49). Notably,

the Notch pathway is significant in liver tissue cell fate determination from embryonic to adult stages and can regulate liver repair and regeneration (50). However, aberrant activation of the Notch pathway can contribute to HCC development and promote HCC cell migration (51). Our findings once again demonstrate that Notch pathway activation is not only involved in the initiation and progression of HCC but also serves as a critical mediator for *TMEM165* in promoting PVTT formation. This not only confirms previous research findings but also provides important evidence for understanding the mechanisms underlying PVTT formation in HCC patients in the future.

Both systemic drug therapy and TACE play significant roles in the treatment of HCC. In this study, we selected eight candidate drugs based on previous RCTs. The patients in the TCGA cohort were stratified into two groups according to their risk of developing PVTT in HCC. Through a sensitivity analysis of these eight drugs in treating both groups of patients, we aimed to predict their efficacy in managing PVTT. In comparison to the low-risk group, sorafenib, which is commonly recommended by the Barcelona Clinic Liver Cancer (BCLC) guidelines for PVTT patients, did not demonstrate a significant advantage in treating the high-risk patients. Similar results were observed for brivanib (52) and sunitinib (53). Other drugs such as gemcitabine (54), cisplatin (55) and camptothecin (56) were found to be less effective in the high-risk group compared to the low-risk group. Interestingly, erlotinib (57) exhibited greater sensitivity in the high-risk group than in the low-risk group. Based on these findings, we believe that erlotinib shows promise as a potential treatment option for PVTT and warrants further investigation.

The tumor microenvironment encompasses non-cancerous cells and components present within tumors, along with the molecules they produce and release. The dynamic interaction between tumor cells and the tumor microenvironment plays a pivotal role in tumorigenesis, tumor development, metastasis, and response to treatment (58). Immune cells represent a crucial component of the tumor microenvironment, given their diverse functions (59). Single-cell sequencing has emerged as a vital tool for investigating the immune microenvironment. However, due to the limited number of PVTT samples included in this study, there were insufficient tumor cells from other tumor microenvironments, aside from HCC cells, for comprehensive analysis. Nonetheless, our CIBERSORT method revealed the infiltration of immune cells in different risk groups. Notably, activated CD4$^+$

memory T cells, M0 macrophages, resting dendritic cells, and neutrophils exhibited high infiltration levels in the high-risk group, suggesting their potential unique roles in PVTT formation.

Finally, our study still has certain limitations. Due to the loss of surgical indications once portal vein tumor thrombosis occurs in HCC patients, collecting PVTT tissues in clinical settings is extremely challenging. Furthermore, progress in animal modeling of PVTT is slow, and currently, it is difficult to establish animal models with significant effects. These factors contribute to the lack of validation of our study results at the tissue level. However, our research provides a methodology for future investigations into the association between *TMEM165* and portal vein tumor thrombosis in HCC. We anticipate that more in-depth studies will be conducted to further explore the relationship between *TMEM165* and the formation of portal vein tumor thrombosis in liver cancer.

Conclusions

The regulatory axis consisting of *TMEM165*, *hsa-miR-148a*, and *LINC00909* exhibits a strong correlation with immune infiltration within the HCC tumor microenvironment, contributing to immune dysfunction and potential failures in immunotherapy. Hence, targeting *TMEM165* holds promise in ameliorating immune dysregulation in HCC patients and potentially mitigating the occurrence of portal vein thrombosis.

Acknowledgments

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://tcr.amegroups.com/article/view/10.21037/tcr-23-1589/rc

Peer Review File: Available at https://tcr.amegroups.com/article/view/10.21037/tcr-23-1589/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://tcr.amegroups.com/article/view/10.21037/tcr-23-1589/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

References

1. Sayiner M, Golabi P, Younossi ZM. Disease Burden of Hepatocellular Carcinoma: A Global Perspective. Dig Dis Sci 2019;64:910-7.
2. Khemlina G, Ikeda S, Kurzrock R. The biology of Hepatocellular carcinoma: implications for genomic and immune therapies. Mol Cancer 2017;16:149.
3. Vyas M, Zhang X. Hepatocellular Carcinoma: Role of Pathology in the Era of Precision Medicine. Clin Liver Dis 2020;24:591-610.
4. Minagawa M, Makuuchi M. Treatment of hepatocellular carcinoma accompanied by portal vein tumor thrombus. World J Gastroenterol 2006;12:7561-7.
5. Liu PH, Huo TI, Miksad RA. Hepatocellular Carcinoma with Portal Vein Tumor Involvement: Best Management Strategies. Semin Liver Dis 2018;38:242-51.
6. Intagliata NM, Caldwell SH, Tripodi A. Diagnosis, Development, and Treatment of Portal Vein Thrombosis in Patients With and Without Cirrhosis. Gastroenterology 2019;156:1582-1599.e1.
7. Zhang ZM, Lai EC, Zhang C, et al. The strategies for treating primary hepatocellular carcinoma with portal vein

tumor thrombus. Int J Surg 2015;20:8-16.

8. Xu X, Wei X, Ling Q, et al. Identification of two portal vein tumor thrombosis associated proteins in hepatocellular carcinoma: protein disulfide-isomerase A6 and apolipoprotein A-I. J Gastroenterol Hepatol 2011;26:1787-94.

9. Li N, Guo W, Shi J, et al. Expression of the chemokine receptor CXCR4 in human hepatocellular carcinoma and its role in portal vein tumor thrombus. J Exp Clin Cancer Res 2010;29:156.

10. Hou YK, Wang Y, Cong WM, et al. Expression of tumor metastasis-suppressor gene KiSS-1 and matrix metalloproteinase-9 in portal vein tumor thrombus of hepatocellular carcinoma. Ai Zheng 2007;26:591-5.

11. Chen L, Heikkinen L, Wang C, et al. Trends in the development of miRNA bioinformatics tools. Brief Bioinform 2019;20:1836-52.

12. Hill M, Tran N. miRNA interplay: mechanisms and consequences in cancer. Dis Model Mech 2021;14:dmm047662.

13. Mishra S, Yadav T, Rani V. Exploring miRNA based approaches in cancer diagnostics and therapeutics. Crit Rev Oncol Hematol 2016;98:12-23.

14. Lu TX, Rothenberg ME. MicroRNA. J Allergy Clin Immunol 2018;141:1202-7.

15. Qi X, Zhang DH, Wu N, et al. ceRNA in cancer: possible functions and clinical implications. J Med Genet 2015;52:710-8.

16. Karreth FA, Pandolfi PP. ceRNA cross-talk in cancer: when ce-bling rivalries go awry. Cancer Discov 2013;3:1113-21.

17. Tay Y, Rinn J, Pandolfi PP. The multilayered complexity of ceRNA crosstalk and competition. Nature 2014;505:344-52.

18. Thomson DW, Dinger ME. Endogenous microRNA sponges: evidence and controversy. Nat Rev Genet 2016;17:272-83.

19. Zhang X, Lan Y, Xu J, et al. CellMarker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res 2019;47:D721-8.

20. Morabito S, Reese F, Rahimzadeh N, et al. High dimensional co-expression networks enable discovery of transcriptomic drivers in complex biological systems. bioRxiv 2022. doi: 10.1101/2022.09.22.509094.

21. Morabito S, Miyoshi E, Michael N, et al. Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. Nat Genet 2021;53:1143-55.

22. Geeleher P, Cox N, Huang RS. pRRophetic: an R package for prediction of clinical chemotherapeutic response from tumor gene expression levels. PLoS One 2014;9:e107468.

23. Roayaie S, Blume IN, Thung SN, et al. A system of classifying microvascular invasion to predict outcome after resection in patients with hepatocellular carcinoma. Gastroenterology 2009;137:850-5.

24. Llovet JM, Bustamante J, Castells A, et al. Natural history of untreated nonsurgical hepatocellular carcinoma: rationale for the design and evaluation of therapeutic trials. Hepatology 1999;29:62-7.

25. Schöniger-Hekele M, Müller C, Kutilek M, et al. Hepatocellular carcinoma in Central Europe: prognostic features and survival. Gut 2001;48:103-9.

26. Costentin CE, Ferrone CR, Arellano RS, et al. Hepatocellular Carcinoma with Macrovascular Invasion: Defining the Optimal Treatment Strategy. Liver Cancer 2017;6:360-74.

27. Zhou XH, Li JR, Zheng TH, et al. Portal vein tumor thrombosis in hepatocellular carcinoma: molecular mechanism and therapy. Clin Exp Metastasis 2023;40:5-32.

28. Zhang J, Pan YF, Ding ZW, et al. RMP promotes venous metastases of hepatocellular carcinoma through promoting IL-6 transcription. Oncogene 2015;34:1575-83.

29. Xia L, Huang W, Tian D, et al. Upregulated FoxM1 expression induced by hepatitis B virus X protein promotes tumor metastasis and indicates poor prognosis in hepatitis B virus-related hepatocellular carcinoma. J Hepatol 2012;57:600-12.

30. Guo W, Liu S, Cheng Y, et al. ICAM-1-Related Noncoding RNA in Cancer Stem Cells Maintains ICAM-1 Expression in Hepatocellular Carcinoma. Clin Cancer Res 2016;22:2041-50.

31. Hinshaw DC, Shevde LA. The Tumor Microenvironment Innately Modulates Cancer Progression. Cancer Res 2019;79:4557-66.

32. Tang Y, Liu S, Li N, et al. 14-3-3ζ promotes hepatocellular carcinoma venous metastasis by modulating hypoxia-inducible factor-1α. Oncotarget 2016;7:15854-67.

33. Zhang C, Gao Y, Du C, et al. Hepatitis B-Induced IL8 Promotes Hepatocellular Carcinoma Venous Metastasis and Intrahepatic Treg Accumulation. Cancer Res 2021;81:2386-98.

34. Reymond N, d'Água BB, Ridley AJ. Crossing the endothelial barrier during metastasis. Nat Rev Cancer 2013;13:858-70.

35. Xue TC, Ge NL, Xu X, et al. High platelet counts increase metastatic risk in huge hepatocellular carcinoma undergoing transarterial chemoembolization. Hepatol Res 2016;46:1028-36.

36. Chen H, Zhou XH, Li JR, et al. Neutrophils: Driving inflammation during the development of hepatocellular carcinoma. Cancer Lett 2021;522:22-31.

37. Foulquier F, Amyere M, Jaeken J, et al. TMEM165 deficiency causes a congenital disorder of glycosylation. Am J Hum Genet 2012;91:15-26.

38. Murali P, Johnson BP, Lu Z, et al. Novel role for the Golgi membrane protein TMEM165 in control of migration and invasion for breast carcinoma. Oncotarget 2020;11:2747-62.

39. Lee JS, Kim MY, Park ER, et al. TMEM165, a Golgi transmembrane protein, is a novel marker for hepatocellular carcinoma and its depletion impairs invasion activity. Oncol Rep 2018;40:1297-306.

40. Bayrak OF, Gulluoglu S, Aydemir E, et al. MicroRNA expression profiling reveals the potential function of microRNA-31 in chordomas. J Neurooncol 2013;115:143-51.

41. Song B, Du J, Song DF, et al. Dysregulation of NCAPG, KNL1, miR-148a-3p, miR-193b-3p, and miR-1179 may contribute to the progression of gastric cancer. Biol Res 2018;51:44.

42. Luo J, Xie M, Hou Y, et al. A novel epigenetic mechanism unravels hsa-miR-148a-3p-mediated CYP2B6 downregulation in alcoholic hepatitis disease. Biochem Pharmacol 2021;188:114582.

43. Vonk LA, Kragten AH, Dhert WJ, et al. Overexpression of hsa-miR-148a promotes cartilage production and inhibits cartilage degradation by osteoarthritic chondrocytes. Osteoarthritis Cartilage 2014;22:145-53.

44. Luo J, Hou Y, Ma W, et al. A novel mechanism underlying alcohol dehydrogenase expression: hsa-miR-148a-3p promotes ADH4 expression via an AGO1-dependent manner in control and ethanol-exposed hepatic cells. Biochem Pharmacol 2021;189:114458.

45. Zhang Y, Guan B, Wu Y, et al. LncRNAs Associated with Chemoradiotherapy Response and Prognosis in Locally Advanced Rectal Cancer. J Inflamm Res 2021;14:6275-92.

46. Yang X, Wu G, Yang F, et al. Elevated LINC00909 Promotes Tumor Progression of Ovarian Cancer via Regulating the miR-23b-3p/MRC2 Axis. Oxid Med Cell Longev 2021;2021:5574130.

47. Liu Z, Lu C, Hu H, et al. LINC00909 promotes tumor progression in human glioma through regulation of miR-194/MUC1-C axis. Biomed Pharmacother 2019;116:108965.

48. Aggarwal V, Tuli HS, Varol M, et al. NOTCH signaling: Journey of an evolutionarily conserved pathway in driving tumor progression and its modulation as a therapeutic target. Crit Rev Oncol Hematol 2021;164:103403.

49. Zhou B, Lin W, Long Y, et al. Notch signaling pathway: architecture, disease, and therapeutics. Signal Transduct Target Ther 2022;7:95.

50. Valizadeh A, Sayadmanesh A, Asemi Z, et al. Regulatory Roles of the Notch Signaling Pathway in Liver Repair and Regeneration: A Novel Therapeutic Target. Curr Med Chem 2021;28:8608-26.

51. Jin M, Wang J, Ji X, et al. MCUR1 facilitates epithelial-mesenchymal transition and metastasis via the mitochondrial calcium dependent ROS/Nrf2/Notch pathway in hepatocellular carcinoma. J Exp Clin Cancer Res 2019;38:136.

52. Llovet JM, Decaens T, Raoul JL, et al. Brivanib in patients with advanced hepatocellular carcinoma who were intolerant to sorafenib or for whom sorafenib failed: results from the randomized phase III BRISK-PS study. J Clin Oncol 2013;31:3509-16.

53. Cheng AL, Kang YK, Lin DY, et al. Sunitinib versus sorafenib in advanced hepatocellular cancer: results of a randomized phase III trial. J Clin Oncol 2013;31:4067-75.

54. Guan Z, Wang Y, Maoleekoonpairoj S, et al. Prospective randomised phase II study of gemcitabine at standard or fixed dose rate schedule in unresectable hepatocellular carcinoma. Br J Cancer 2003;89:1865-9.

55. Aramaki O, Takayama T, Moriguchi M, et al. Arterial chemoembolisation with cisplatin versus epirubicin for hepatocellular carcinoma (ACE 500 study): A multicentre, randomised controlled phase 2/3 trial. Eur J Cancer 2021;157:373-82.

56. Chen F, Wang H, Zhu J, et al. Camptothecin suppresses NRF2-ARE activity and sensitises hepatocellular carcinoma cells to anticancer drugs. Br J Cancer 2017;117:1495-506.

57. Zhu AX, Rosmorduc O, Evans TR, et al. SEARCH: a phase III, randomized, double-blind, placebo-controlled trial of sorafenib plus erlotinib in patients with advanced hepatocellular carcinoma. J Clin Oncol 2015;33:559-66.

58. Xiao Y, Yu D. Tumor microenvironment as a therapeutic target in cancer. Pharmacol Ther 2021;221:107753.

59. Garcia-Gomez A, Rodríguez-Ubreva J, Ballestar E. Epigenetic interplay between immune, stromal and cancer cells in the tumor microenvironment. Clin Immunol 2018;196:64-71.