Appendix 1 Methods

Tissue dissociation and preparation

The specimens were washed for three times with Hanks balanced salt solution, minced into small pieces, and then digested with 3 mL sCelLive^{TM} tissue dissociation solution (Singleron, Nanjing, China) at 37 ℃ for 15 minutes by PythoN™ tissue dissociation system (Singleron, Nanjing, China). The cell suspension was collected and filtered through a 40-micron sterile strainer. Afterwards, the GEXSCOPE® red blood cell lysis buffer (RCLB, Singleron, Nanjing, China) was added, and the mixture (cell: RCLB =1:2 volume ratio) was incubated at room temperature for 5–8 minutes to remove red blood cells. The mixture was then centrifuged at 300 ×g 4 ℃ for 5 minutes to remove supernatant and suspended softly with PBS.

Primary analysis of raw read data

Raw reads were processed to generate gene expression profiles using CeleScope (v1.14.0, Singleron, Nanjing, China) with default parameters. Briefly, Barcodes and UMIs were extracted from R1 reads and corrected. Adapter sequences and poly A tails were trimmed from R2 reads and the trimmed R2 reads were aligned against the GRCh38 (hg38) transcriptome using STAR (v2.6.1b). Uniquely mapped reads were then assigned to exons with FeatureCounts (v2.0.1). Successfully Assigned Reads with the same cell barcode, UMI and gene were grouped together to generate the gene expression matrix for further analysis.

Quality control, dimension-reduction and clustering

Scanpy (v1.8.1) was used for quality control, dimensionality reduction and clustering under Python 3.7 (1). For each sample dataset, we filtered expression matrix by the following criteria: (I) cells with gene count less than 200 or with top 2% gene count were excluded; (II) cells with top 2% UMI count were excluded; (III) cells with mitochondrial content over 30% were excluded; (IV) genes expressed in less than 5 cells were excluded. After filtering, 26,800 cells were retained for the downstream analyses, with on average 1,784.391 genes and 6,437.798 UMIs per cell. The raw count matrix was normalized by total counts per cell and logarithmically transformed into normalized data matrix. Top 2000 variable genes were selected by setting flavor = ’seurat’. Principle component analysis (PCA) was performed on the scaled variable gene matrix, and top 20 principle components were used for clustering and dimensional reduction. Cells were separated into 22 clusters by using Louvain algorithm and setting resolution parameter at 1.2. Cell clusters were visualized by using Uniform Manifold Approximation and Projection (UMAP).

Differentially expressed genes (DEGs) analysis

To identify DEGs by Scanpy (v1.8.1), we used the “scanpy.tl.rank_genes_groups” function based on Wilcoxon rank sum test with default parameters, and selected the genes expressed in more than 10% of the cells in either of the compared groups of cells and with an average log (fold change) value greater than 0.25 as DEGs. Adjusted P value was calculated by Benjamini-Hochberg correction and the value 0.05 was used as the criterion to evaluate the statistical significance.

Pathway enrichment analysis

To investigate the potential functional pathways, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis were used with the “clusterProfiler” R package (v 4.0.0) (2). Pathways with p_adj value less than 0.05 were considered as significantly enriched. Selected significant pathways were plotted as bar plots. Gene set enrichment analysis (GSEA) was performed on ATP related genes in the CancerCell 6 cluster (3). For gene set variation analysis (GSVA), the average gene expression of each cell type was used as input data (4). GO gene sets were including molecular function (MF), biological process (BP), and cellular component (CC) categories.

Cell-type recognition with cell-ID

Cell-ID is multivariate approach that extracts gene signatures for each individual cell and perform cell identity recognition using hypergeometric tests (HGT) (5). Dimensionality reduction was performed on normalized gene expression matrix through multiple correspondence analysis, where both cells and genes were projected in the same low dimensional space. Then a gene ranking was calculated for each cell to obtain most featured gene sets of that cell. HGT were performed on these gene sets against brain reference from SynEcoSys database, which contains all cell-type's featured genes. Identity of each cell was determined as the cell-type has the minimal HGT p value. For cluster annotation, Frequency of each cell-type was calculated in each cluster, and cell-type with highest frequency was chosen as cluster's identity.

Cell-cell interaction analysis

Cell-cell interaction was predicted based on known ligand-receptor pairs by Cellphone DB (v4.0.0) (6). Permutation number for calculating the null distribution of average ligand-receptor pair expression in randomized cell identities was set to 1,000. Individual ligand or receptor expression was thresholded by a cutoff based on the average log gene expression distribution for all genes across each cell type. Predicted interaction pairs with P value <0.05 and of average log expression >0.1 were considered as significant and visualized by heat map and dot plot in CellphoneDB.

Pseudo-time trajectory analysis

Cell differentiation trajectory of hepatocyte and MP subtypes was reconstructed with the Monocle2 (v2.10.0) (7). For constructing the trajectory by using Seurat (v3.1.2), top 2,000 highly variable genes were selected by "FindVairableFeatures" function, and dimension-reduction was performed by "DDRTree" function. The trajectory was visualized by "plot_cell_trajectory" function in Monocle2.

Functional gene module analysis

Hotspot was used to identify functional gene modules which illustrate heterogeneity within hepatocyte (cancer cell) subpopulations (8). Briefly, we used the "danb" model and selected the top 500 genes with highest autocorrelation z-score for module identification. Modules were then identified using the "create_modules" function, with "min_gene_threshold"=15 and "fdr_threshold"= 0.05. Module scores were calculated by using "calculate_module_scores" function.

UCell gene set scoring

Gene set scoring was performed by using UCell (v2.2.0) (9). UCell scores are based on the Mann-Whitney U statistic by ranking query genes in order of their expression levels in individual cells. Because UCell is a rank-based scoring method, it is suitable to be used in large datasets containing multiple samples and batches.

Transcription factor regulatory network analysis

Transcription factor network was constructed by pySCENIC (v0.11.0) using scRNA expression matrix and transcription factors in AnimalTFDB (10). First, GRNBoost2 predicted a regulatory network based on the co-expression of regulators and targets. CisTarget was then applied to exclude indirect targets and to search transcription factor binding motifs. After that, UCell was used for regulon activity quantification for every cell. Cluster-specific regulons were identified according to the regulon specificity scores and the activity of these regulons were visualized in heat maps.

Copy number alterations (CNAs) detection

The InferCNV package was applied to detect the CNAs in malignant cells (hepatocytes) (11). T and NK cells were used as baselines to estimate the CNAs of malignant cells. Genes expressed in more than 20 cells were sorted based on their loci on each chromosome. The relative expression values were centered to 1, using 1.5 standard deviation from the residualnormalized expression values as the floor and ceiling. A slide window size of 101 genes was used to smoothen the relative expression on each chromosome, to remove the effect of gene-specific expression. The inferred CNAs on each short or long arm or full length of the chromosomes were visualized with heat maps generated by "pheatmap" function. The clonal relationships of malignant cell clusters in each sample were determined by the accumulation of CNAs.

References

1. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 2018;19:15.
2. Yu G, Wang LG, Han Y, et al. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16:284-7.
3. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545-50.
4. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics 2013;14:7.
5. Cortal A, Martignetti L, Six E, et al. Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID. Nat Biotechnol 2021;39:1095-102.
6. Efremova M, Vento-Tormo M, Teichmann SA, et al. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. Nat Protoc 2020;15:1484-506.
7. Qiu X, Hill A, Packer J, et al. Single-cell mRNA quantification and differential analysis with Census. Nat Methods 2017;14:309-15.
8. DeTomaso D, Yosef N. Hotspot identifies informative gene modules across modalities of single-cell genomics. Cell Syst 2021;12:446-456.e9.
9. Andreatta M, Carmona SJ. UCell: Robust and scalable single-cell gene signature scoring. Comput Struct Biotechnol J 2021;19:3796-8.
10. Van de Sande B, Flerin C, Davie K, et al. A scalable SCENIC workflow for single-cell gene regulatory network analysis. Nat Protoc 2020;15:2247-76.
11. Tirosh I, Venteicher AS, Hebert C, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature 2016;539:309-13.

Table S1 Clinical information of the HCC samples

Patient characteristics	Sample	
	01	02
Gender	Male	Female
Age (years)	67	74
Treatment strategies	TACE + T + A	TACE + T + A
ECOG PS score	0	0
BCLC stage	C	C
Child-Pugh classification	A	A
HBV infection	Yes	Yes
Cirrhosis	Yes	Yes
Tumor number	1	1
Tumor largest length (cm)	6.7	6.1
PVTT	Yes	Yes
Metastasis	No	No
AFP (ng/mL)	125.7	704.8
ALT (U/L)	149.1	37.7
AST (U/L)	179.9	84.2
ALB (g/L)	39.3	35.9
TBIL (μmol/L)	18	12.8
Cr (μmol/L)	73.4	55.9
Na (mmol/L)	141.7	135.6
WBC ($\times 10^9$/L)	5.36	4.87
Hb (g/L)	123	114
PLT ($\times 10^9$/L)	210	153
N ($\times 10^9$/L)	3.33	3.23
L ($\times 10^9$/L)	1.57	0.87
M ($\times 10^9$/L)	0.32	0.52
PT (seconds)	11.7	14.3
INR	1.02	1.26
Time to progression (days)	34	37

TACE+T+A, transarterial chemoembolization combined with atezolizumab (Tecentriq) and bevacizumab (Avastin). ECOG PS, Eastern Cooperative Oncology Group performance status; BCLC; Barcelona Clinic Liver Cancer; HBV, hepatitis B virus; PVTT, portal vein tumor thrombus; AFP, alpha fetoprotein; ALT, alanine aminotransferase; AST, aspartate aminotransferase; ALB, albumin; TBIL, total bilirubin; Cr, creatinine; Na, serum sodium; WBC, white blood cell; Hb, hemoglobin; PLT, platelet; N, neutrophil; L, lymphocyte; M, monocyte; PT, prothrombin time; INR, international normalized ratio.

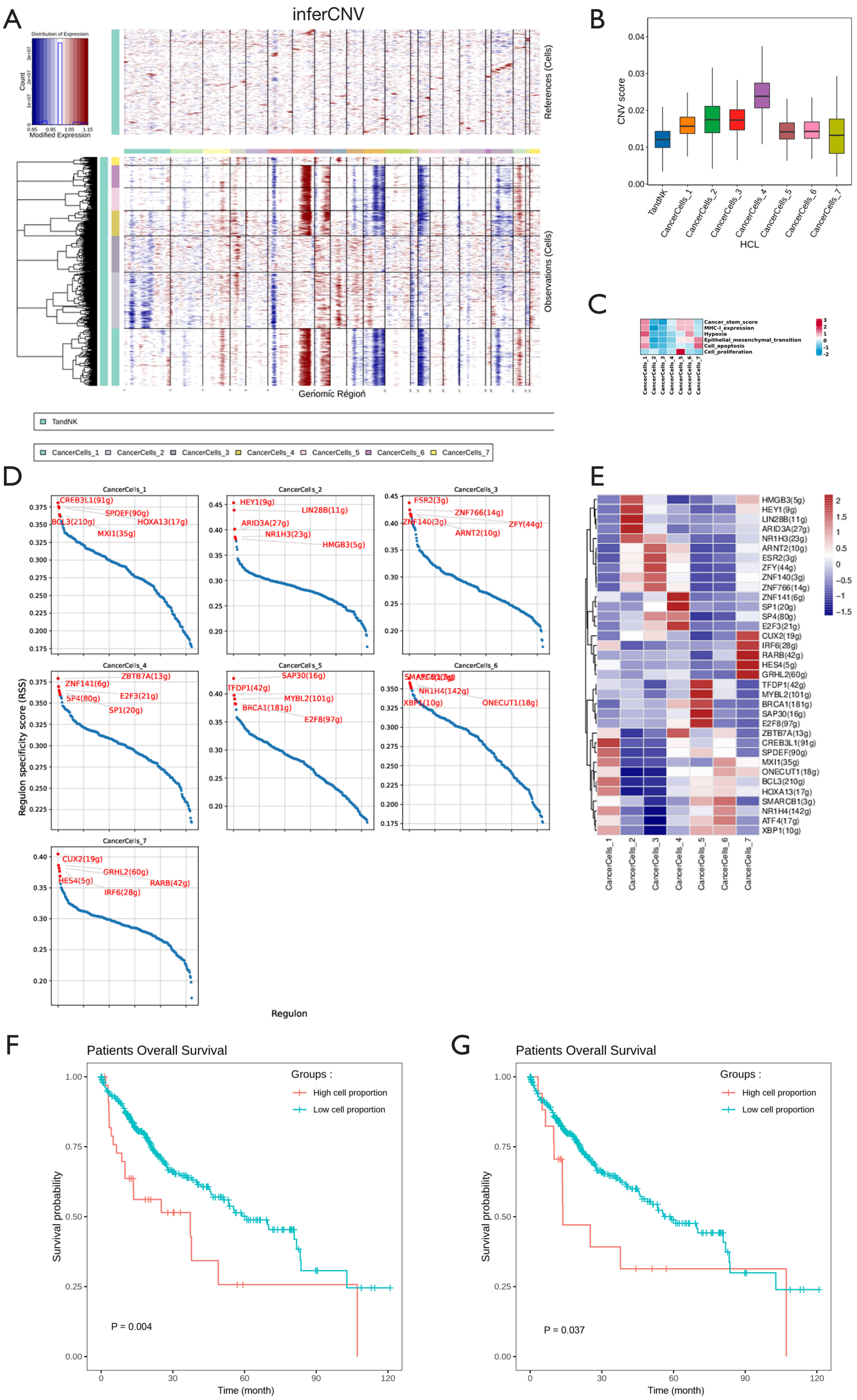

Figure S1 The characteristics of sub-clusters of hepatocytes. The CancerCell 4 cluster was the highest degree of malignancy and the CancerCell 7 cluster was the lowest degree of malignancy according to the copy number variation (CNV) analysis (A,B). The CancerCell 6 cluster had strong signaling pathways of tumor stem score, hypoxia, MHC-I expression, and epithelial mesenchymal transition by gene set variation analysis (GSVA) (C). The CancerCell 6 cluster had several high expressions of transcription factors, such as *XBP1*, *ATF4*, and *MXI1* (D,E). Except the CancerCell 6 cluster, the CancerCell 4 and 5 clusters also showed significant differences between high cell proportion and low cell proportion for overall survival (F,G).

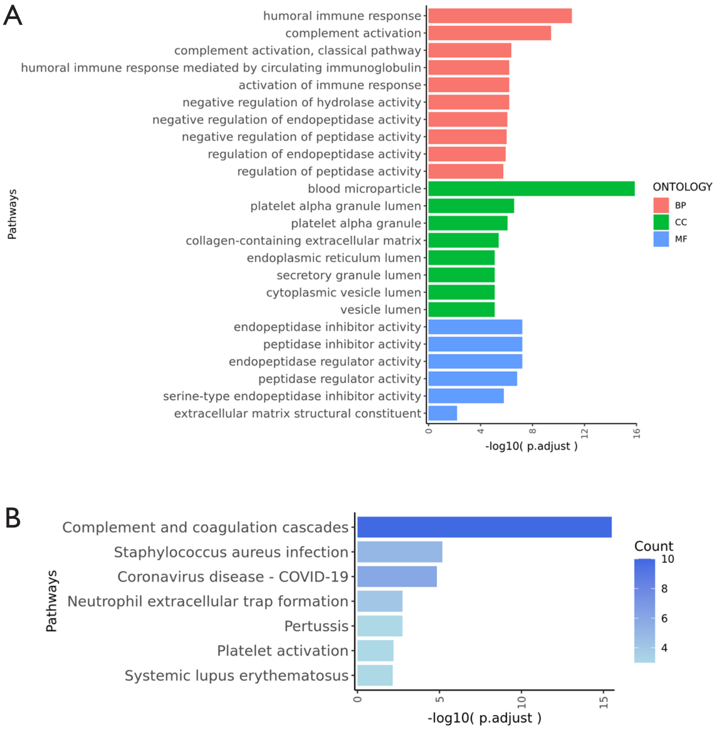

Figure S2 The GO and KEGG gene enrichment analyses of the module 9. The GO (A) and KEGG (B) gene enrichment analyses showed enrichment of endopeptidase-related signaling pathways.

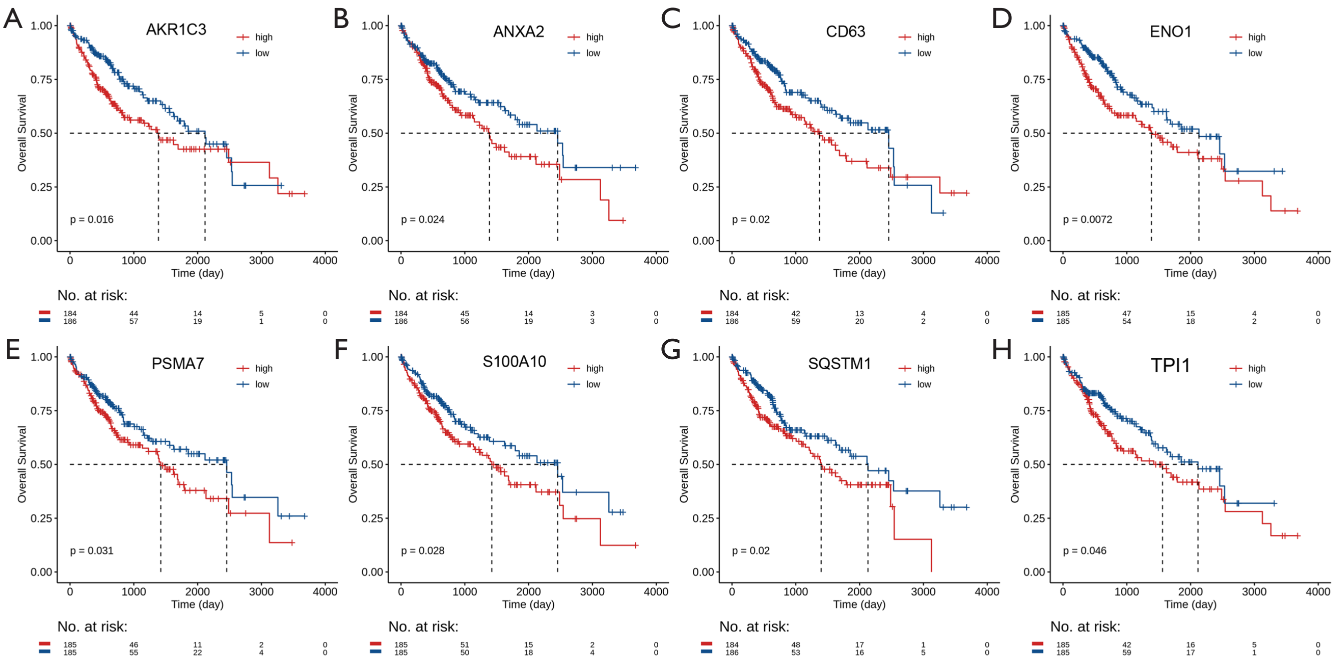

Figure S3 The prognosis analyses of different gene expression groups. The Kaplan-Meier curves showed that the high gene expression groups had poor OS than the low gene expression groups, including *AKR1C3*, *ANXA2*, *CD63*, *ENO1*, *TPI1*, *SQSTM1*, *S100A10* and *PSMA7* (A-H).

https://dx.doi.org/10.21037/cco-24-58

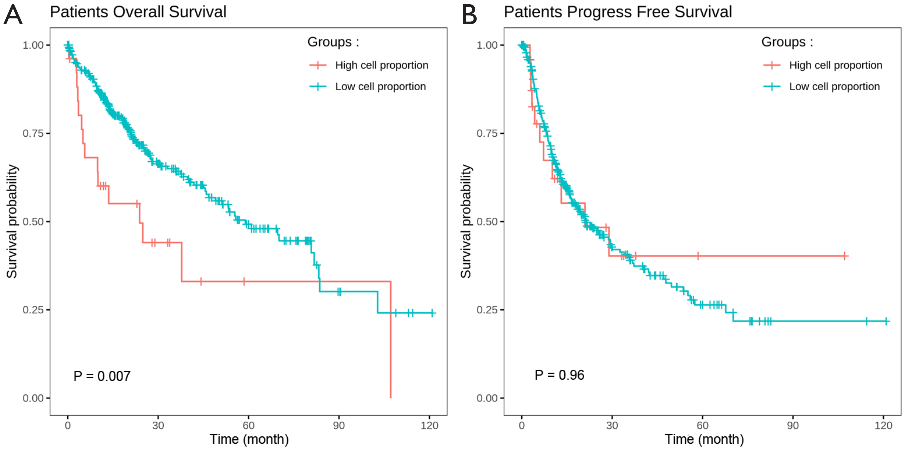

Figure S4 The prognosis analysis of MPs_TREM2. The Kaplan-Meier curve showed high cell proportion of MPs_TREM2 had poor OS than the low cell proportion of MPs_TREM2 (P=0.007, A), but it didn't show significant difference of PFS (P=0.96, B).

https://dx.doi.org/10.21037/cco-24-58