Supplementary methods

Clinical characteristics and pathological variables

The patient-related variables included sex, age, American Society of Anesthesiologists (ASA) score, smoking, drinking and body mass index (BMI). The liver-related variables included cirrhosis, portal hypertension, Child-Pugh grade, the positiveness of hepatitis B virus (HBV) or hepatitis C virus (HCV), pre-operative alanine aminotransferase (ALT), and aspartate transaminase (AST), alkaline phosphatase (AKP) and γ-glutamyl transpeptidase (γGT). The tumor-related variables included the presence of cancer-associated symptoms, preoperative alpha-fetoprotein (AFP) level, largest tumor size, satellite nodules, macrovascular invasion, microvascular invasion, tumor differentiation, tumor envelope, AJCC-TNM staging and BCLC staging. Satellite nodules were defined as tumors size less than 1 cm and located less than 1 cm away from the primary focus. The operative variables consisted of intraoperative blood transfusion.

DNA/RNA extraction, WGS and gene expression profiling

Genomic DNA was extracted from tumors and matched adjacent non-tumor samples using QIAamp Fast DNA tissue kit (QIAGEN) according to manufacturer's protocol. DNA purity and concentration were evaluated by the Qubit 4.0 (Invitrogen) and NanoDrop One (Thermo Fisher Scientific™). DNA integrity was assessed by Agilent DNA 1000 Kit (Agilent Technologies) using Agilent 2100 Bioanalyzer System (Agilent Technologies). WGS libraries were prepared using the Hieff NGS® Ultima Pro DNA Library Prep Kit for Illumina (Yeasen Biotechnology). DNBSEQ-T7 platform (MGI Technology) was utilized to sequence the constructed DNA libraries. WGS was shown with a mean coverage depth of 30X-50X for all samples, and median Q20 and Q30 are 96% and 89%.

Total RNA was successfully extracted from fresh frozen tissues using RNeasy®Mini Kit (Qiagen). RNA purity and concentration were quantified using the NanoDrop 2000 (Thermo Fisher Scientific™). RNA integrity was measured with an

Fan Z, *et al. Gut* 2023;0:1–15. doi: 10.1136/gutjnl-2023-330461

Agilent 2100 Bioanalyzer System (Agilent Technologies). Specimens with high RNA integrity number (> 6) and enough amount of RNA (> 1 μg) were included to create DNA library by used the NEB Next® UltraTM RNA Library Prep Kit (NEB). The RNA was processed, labeled and hybridized to Agilent SurePrint G3 Custom Human GE 4x180K chips (Design ID: 085539). The signal was detected Agilent Scanner G2505C (Agilent Technologies) and Feature Extraction (version 10.7.1.1, Agilent Technologies) was utilized to generate raw data of mRNA and lncRNA expression. The quantile normalization was implemented by GeneSpring GX (version 14.9, Agilent Technologies).

Protein preparation and LC-MS/MS analysis

Twelve pairs of tumor and non-tumor tissues were collected for proteomic analysis using previously published methods.[1] briefly, ten formalin-fixed paraffin-embedded (FFPE) sections measuring 2cm x 2cm x 10μm were collected from each sample. Following deparaffinization, a lysis buffer containing 1% protease inhibitor was used to isolate proteins, whose concentrations were determined using the BCA protein assay kit. Equal amounts of proteins were subjected to trypsin digestion at a ratio of 1:50 trypsin to protein, resulting in peptide fragments. The peptide fragments were dissolved in mobile phase A, consisting of a water-based solution with 0.1% formic acid and 2% acetonitrile, and separated using the NanoElute ultra-high-performance liquid chromatography (UHPLC) system by setting a gradient of solvent B (100% acetonitrile-based solution with 0.1% formic acid) as follows: 0-70 min with 6%~24% solvent B; 70-84 min with 24%~35% solvent B; 84-87 min with 35%~80% solvent B; 87-90 min with 80% solvent B at a flow rate of 450nL/min. The peptide fragments were then ionized and injected into the Capillary ion source for analysis on the timsTOF Pro mass spectrometer, with an ion source voltage of 1.75 kV, followed by detection and analysis of both precursor ions and their fragment ions using high-resolution TOF. Secondary mass spectrometry scanning range was set to 100-1700, and data acquisition mode used parallel accumulation-serial fragmentation (PASEF). Maxquant software (v1.6.15.0) was used

to analyze the secondary mass spectrometry data, with the Homo_sapiens_9606 database containing 20,395 entries used as the reference database. A decoy database was included to calculate the false discovery rate (FDR) caused by random matches, and a common contaminant library was added to eliminate any influence of contaminating proteins in the identification result. Peptide identification accuracy was set at FDR of 1% at the spectral, peptide, and protein levels. Protein identification was required to have at least one unique peptide, with search tolerances of 20 ppm for first search and 4.5 ppm for main search, and minimal peptide length of 7. Identification accuracy was set to an FDR of 1% at the spectral, peptide, and protein levels, at least one unique peptide was required for protein identification.

RNA microarray data analysis and proteomics analysis

The gene mRNA expression levels were subjected to hierarchical clustering, and principal component analysis (PCA) was performed using the R packages FactoMineR (version 2.7) and factoextra (version 1.0.7).[2-3] Differentially expressed genes (DEGs) of mRNA and lncRNA between tumor and non-tumor tissues were identified by applying the R package limma (version 3.54.1) with |log2FC|>1 and adjusted *P* value <0.05, whereas a *P* value <0.05 was used when comparing non-tumor tissues in different groups.[3] LncRNA annotation was performed on LncBook 2.0 (https://ngdc.cncb.ac.cn/lncbook/) and LNCipedia (https://lncipedia.org/) using the sequence or chromosomal location of differentially expressed lncRNAs. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses of DEGs and gene set enrichment analysis (GSEA) of all genes were conducted using the R package clusterProfiler (version 4.6.2).[4] To better illustrate the differential enriched pathways among the four types of HCC, we performed gene set variation analysis (GSVA) using the R package GSVA (version 1.46.0) for hallmark, GO and KEGG from MSigDB collections as well as tumor microenvironment score-related gene sets.[5] We used the R package simplifyEnrichment (version 1.8.0) to slim down and visualize the results of GO function enrichment.[6] To explore the tumor microenvironment, we calculated

stromal and immune scores using the R package ESTIMATE (version 1.0.13).[7] Furthermore, we utilized the xCell to estimate the abundances of immune cells of interest present in the tumor milieu. Except special indicating, the visualization was achieved by using the R package ggplot2 (version 3.4.1).[8-9] Construction of a regulatory network involving lncRNAs, microRNAs (miRNAs) and mRNAs was conducted using starBase (https://rnasysu.com/encori/).

Protein relative quantification was performed using the following steps: first, the signal intensity values of peptides were normalized to obtain their relative quantitative values; second, peptide relative quantification values were further corrected by median normalization within each sample; third, protein relative quantification values were calculated as medians of the relative quantification values of protein-specific peptides. FC for each protein between paired tumor and non-tumor samples was calculated by determining the ratio of the mean quantification values of all replicates. Proteins with $|\log2FC| > 0.585$ and P value <0.05 were considered differentially expressed.

WGS data analysis

Low-quality reads from whole genome sequencing data were removed using Trimmomatic (version 0.39).[10] The resulting high-quality reads were aligned to the UCSC hg19 reference sequence with BWA (version 0.7.17) and PCR duplicates were removed and recalibrated using GATK (version 4.2.6.1).[11-12] Somatic variants were identified using Mutect2 on tumor and matched non-tumor pairs, and annotated with Annovar (version 2017 Jul 17), resulting in a total of 7,045 non-silent somatic single nucleotide variant (SNV) calls and 396 indel calls for 49 pairs of tumor and non-tumor liver samples.[13-14] Significantly mutated genes were identified by mutSigCV (version 1.41) using q-values with a threshold of 0.05.[15] Copy number variation (CNV) was determined for each sample using Control-FREEC (version 11.6).[16] Significant focal CNVs across all samples were identified using Genomic Identification of Significant Targets in Cancer (GISTIC, version 2.0.23) with q values < 0.05, indicating regions with significant gains or losses beyond chance.[17]

CNV-mRNA Correlation

Pearson correlation coefficient was utilized to evaluate the correlation between mRNA expression and CNVs. Significance correlation pair was identified as correlation of correlation >0.5 and *P* value <0.05. Visualization was used to display the correlation by shinyCircos (https://yimingyu.shinyapps.io/shinycircos/).

WGCNA analysis

To identify functional modules associated with gross classification, we conducted weighted gene co-expression network analysis (WGCNA) using R package WGCNA (version 1.72-1).[18] We selected the top 5000 genes with the highest median expression levels across all samples and constructed a gene co-expression topology overlap matrix based on gene correlation. K-means clustering was performed to define network modules, with a minimum of 30 genes in each module. We calculated the correlation between clinicopathological features and gene modules using gene significance (GS) and module membership (MM) values. GS and MM were used to estimate the association of individual genes with gross classification by measuring their correlation within each module. Hub-genes for each gross type were identified based on |GS| > 0.2 or 0.3 and |MM|> 0.8.

Prognostic risk Model

To construct prognostic biomarker for HCC, all HCC samples in the The Cancer Genome Atlas (TCGA) database were enrolled as training set, and the R package glmnet (version 4.1.7) was utilized to perform Lasso Cox regression analysis.[19] Univariate and multivariate Cox analyses were conducted on the selected genes which were probably associated with inferior overall survival (OS). Subsequently, a prognostic risk model for predicting OS in patients with was formulated. The performance of the model was assessed by calculating the area under the receiver operating characteristic (ROC) curve using R packages timeROC (version 0.4).[20] The model was further confirmed in the validation cohorts (159 and 231 patients in HBV-related HCC cohort and ICGC cohort, respectively).[21-22]

Follow-up

The postoperative surveillance strategy for tumor recurrence consisted of a serum AFP test, ultrasonography, or contrast-enhanced CT or magnetic resonance imaging (MRI) scan of the abdomen at 3-monthly intervals for the first 2 years, and once every 6 months at 2 years or later after resection. Tumor recurrence was defined as new appearance of intra- or extra-hepatic tumor nodule(s), and these intrahepatic nodules had the typical imaging features consistent with the characteristics of HCC on contrast-enhanced MRI or CT examinations.

H&E, Masson and immunohistochemistry staining

Samples were collected and fixed in formalin for 24 hours, washed in 70% ethanol and embedded in paraffin. Paraffin embedded sections at 5 μm were deparaffinized, rehydrated, and washed in distilled water. The sections were then stained with H&E. For Masson staining, the sections were stained with hematoxylin, ponceau red liquid dye acid complex and aniline blue. For immunohistochemistry staining, the sections were stained using anti-CD8-alpha antibody (Abcam, 1:500), anti-CD68 antibody (Abcam, 1:100), anti-α-SMA antibody (Proteintech, 1:1000), anti-CD45 antibody (Proteintech, 1:2000), anti-CD34 antibody (Proteintech, 1:1000), anti-VEGFA antibody (Proteintech, 1:200), anti-VEGFC antibody (Proteintech, 1:200), anti-PDGFRA antibody (Abcam, 1:500), anti-PDGFRB antibody (Proteintech, 1:400), (Proteintech, 1:1000), anti-HGF antibody (Proteintech, 1:200) or anti-TGFbeta1 antibody (Abcam, 1:400). Horseradish-peroxidase-labeled goat anti-mouse or anti-rabbit secondary antibody (Invitrogen) was then incubated, and the sections were colored by DAB kit (Invitrogen). Images were captured by MoticEasyScan (Motic).

Statistical analyses

The baseline characteristics and operative variables of patients were summarized using frequencies, percentages, mean±standard deviation, or median (range), depending on whether the variables were categorical or continuous. Continuous

variables were compared using either Student's t-test, Mann-Whitney U test, or Kruskal-Wallis H test. Categorical variables were compared using either the $\chi 2$ test with Yates correction or Fisher's exact test using the linear-by-linear association method. Variables that showed a *P* value of less than 0.05 in univariate analysis were selected for multivariate analysis using a Cox proportional hazards regression model with forward stepwise variable selection. The stabilized inverse probability of treatment weighting (IPTW) method was used to create a pseudo-population by weighting the inverse probability of a patient having different gross types or therapeutic strategies based on the propensity score.[23] OS was defined as the time from surgery to death, while recurrence-free survival (RFS) was defined as the time from surgery to death or new tumor occurrence. Hazard ratios (HRs) and 95% confidence intervals (CIs) were reported. The statistical analyses were performed using SPSS software version 25.0 (SPSS, Chicago, IL, USA) and R 4.2.2. Statistical significance was set at $P < 0.05$, two-tailed.

References

1. Xi Y, Zhang D, Liang Y, Shan Z, Teng X, Teng W. Proteomic Analysis of the Intestinal Resistance to Thyroid Hormone Mouse Model With Thyroid Hormone Receptor Alpha Mutations. *Front Endocrinol (Lausanne)* 2022; **13**:773516.
2. S Lê JJ, Husson F. FactoMineR: an R package for multivariate analysis. *Journal of Statistical Software* 2008 .
3. A Kassambara FM. Extract and visualize the results of multivariate data analysis. *Package "factoextra", version* 2017 .
4. T Wu EH, S Xu MC, P Guo ZD, Feng T. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* 2021 .
5. S Hänzelmann RC, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC BIOINFORMATICS* 2013 .
6. Z Gu DH. Simplify enrichment: A bioconductor package for clustering and visualizing functional enrichment results. *GENOMICS PROTEOMICS & BIOINFORMATICS* 2022 .
7. K Yoshihara MS, Martínez E. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature ...* 2013 .
8. H Wickham HW. Data analysis. *ggplot2: elegant graphics for data analysis* 2016 .
9. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome*

Biol 2017; **18**:220.

10. AM Bolger ML, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *BIOINFORMATICS* 2014 .
11. H Li RD. Fast and accurate short read alignment with Burrows–Wheeler transform. *BIOINFORMATICS* 2009 .
12. GA Van der Auwera MOC. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in …* 2013 .
13. D Benjamin TS, K Cibulskis GG, Stewart C. Calling somatic SNVs and indels with Mutect2. *BioRxiv* 2019 .
14. K Wang ML, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *NUCLEIC ACIDS RESEARCH* 2010 .
15. AS Cibulskis SLC. Mutational heterogeneity in cancer and the search for new cancer genes. *NATURE* 2013 .
16. V Boeva TP, K Bleakley PC. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *...* 2012 .
17. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* 2011; **12**:R41.
18. P Langfelder SH. WGCNA: an R package for weighted correlation network analysis. *BMC BIOINFORMATICS* 2008 .
19. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 2010; **33**:1-22.
20. Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat Med* 2013; **32**:5381-97.
21. Gao Q, Zhu H, Dong L*, et al.* Integrated Proteogenomic Characterization of HBV-Related Hepatocellular Carcinoma. *Cell* 2019; **179**:561-77.e22.
22. Zhang J, Bajari R, Andric D*, et al.* The International Cancer Genome Consortium Data Portal. *Nat Biotechnol* 2019; **37**:367-9.
23. Xu S, Ross C, Raebel MA, Shetterly S, Blanchette C, Smith D. Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value Health* 2010; **13**:273-7.