

Integrative multi-cohort transcriptomic analysis identifies conserved molecular programs for detecting hepatocellular carcinoma in chronic liver disease

Yan Liu

Baogang Hospital of Inner Mongolia, The Third Affiliated Hospital of Inner Mongolia Medical University

Yishen Li

China Pharmaceutical University

Jianpeng Wei

benwei12138@gmail.com

Fudan University

Article

Keywords: hepatocellular carcinoma, early detection, single-cell transcriptomics, chronic liver disease, gene-expression programs, machine learning

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10642808/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Integrative multi-cohort transcriptomic analysis identifies conserved molecular programs for detecting hepatocellular carcinoma in chronic liver disease

Yan Liu¹, Yishen Li², Jianpeng Wei^{3*}

¹Department of Pathology, Baogang Hospital of Inner Mongolia, The Third Affiliated Hospital of Inner Mongolia Medical University, 20 Shaoxian Road, Baotou, 014010, Inner Mongolia, People's Republic of China.

²State Key Laboratory of Natural Medicines, Jiangsu Key Laboratory of Carcinogenesis and Intervention, School of Basic Medicine and Clinical Pharmacy, China Pharmaceutical University, 24 Tongjiaxiang, Nanjing, 210009, People's Republic of China.

^{3*}State Key Laboratory of Genetic Engineering, School of Life Sciences, Greater Bay Area Institute of Precision Medicine (Guangzhou), Zhongshan Hospital, Fudan University, Shanghai, 200438, China.

*Corresponding author(s). E-mail(s): benwei12138@gmail.com;

Abstract

Early hepatocellular carcinoma (HCC) must be distinguished from the transcriptional effects of chronic liver injury. We integrated 141,428 cells and nuclei from 36 patients and 41 samples spanning healthy liver, cirrhotic and steatohepatic adjacent tissue, and HCC. Copy-number inference, continuous copy-number-variation burden and marker-supported annotation defined graded malignant hepatocyte states. Recurrent expression programs were identified by sample-wise non-negative matrix factorisation, directionally screened in TCGA-LIHC and ICGC-LIRI-JP, and evaluated using patient-grouped cross-validation of synthetic tumour-background mixtures. This yielded a locked 12-gene panel combining tumour-associated expression with loss of mature hepatocyte identity. An Extra-Trees classifier achieved an out-of-fold AUROC of 0.9983 in TCGA mixtures,

while the direction-aligned score distinguished TCGA tumours from solid-tissue-normal samples (AUROC, 0.947; 95% confidence interval, 0.921–0.968). At 5% nominal bulk-tumour content, median detection probabilities were approximately 0.80 in TCGA and 0.55 in ICGC. A fixed TCGA-derived cutpoint stratified overall survival in both cohorts (log-rank $P < 0.001$). Higher risk scores were associated with T-cell and myeloid/macrophage infiltration and T-cell exhaustion. This CNV-anchored signature remains detectable after substantial dilution and warrants validation in independent chronic-liver-disease surveillance cohorts.

Keywords: hepatocellular carcinoma, early detection, single-cell transcriptomics, chronic liver disease, gene-expression programs, machine learning

1 Introduction

Hepatocellular carcinoma (HCC) is the most common primary liver malignancy and remains a leading cause of cancer-related mortality worldwide[1, 2]. Its aetiological landscape is changing: as viral hepatitis is increasingly controlled through vaccination and antiviral therapy, metabolic liver disease, particularly non-alcoholic fatty liver disease (NAFLD) and its inflammatory form, non-alcoholic steatohepatitis (NASH), is becoming an increasingly important driver of HCC[2–4]. A defining clinical challenge along this disease axis is that the tumour develops within an already diseased organ. The liver of an at-risk patient is not equivalent to normal tissue; it is chronically injured, inflamed and fibrotic or cirrhotic, with regenerating hepatocytes, activated myofibroblasts and infiltrating immune cells already reshaping its transcriptional and genomic landscape[5–8]. Early HCC must therefore be distinguished not from healthy liver, but from an actively remodelling fibrotic and steatohepatitic background.

Current surveillance tools are poorly matched to this challenge. Abdominal ultrasonography, the mainstay of HCC surveillance, has limited sensitivity for early-stage tumours and is strongly affected by operator experience, body habitus and the nodular heterogeneity of the cirrhotic liver[2, 9–11]. Serum alpha-fetoprotein (AFP), the most widely used circulating biomarker, remains within the normal range in a substantial proportion of early HCCs and may also be elevated by active hepatitis or hepatocyte regeneration. Consequently, AFP lacks both sensitivity and specificity in the population for whom surveillance is most critical[9, 11]. Composite scores incorporating AFP-L3 and des-gamma-carboxy prothrombin, including GALAD, improve detection relative to AFP alone but do not fully close the diagnostic gap at the earliest, potentially curable stage[12, 13]. As a result, many patients with fibrosis- or NASH-associated HCC are diagnosed only after the optimal window for curative resection, ablation or transplantation has narrowed or closed.

This unmet need motivates a molecular approach that seeks to detect the transcriptional program of an emerging tumour rather than relying exclusively on visualisation of a discrete mass[4, 14]. Bulk transcriptomic profiling of liver tissue has been used to construct HCC classifiers, but an early lesion embedded within a fibrotic liver may contribute only a small fraction of the total RNA. Its malignant signal is therefore

readily diluted by surrounding non-malignant cells. Moreover, single-gene differential-expression signatures derived from bulk tissue can confound tumour-specific biology with fibrosis, inflammation, regeneration and inter-patient heterogeneity. Single-cell and single-nucleus RNA sequencing (scRNA-seq and snRNA-seq) offer a direct means of resolving this problem by measuring hepatocytes and putative tumour cells individually, thereby enabling the malignant compartment to be identified and characterised separately from the fibrotic and steatotic tissue surrounding it[5, 7, 8, 15–18].

We therefore reasoned that single-cell data provide the appropriate setting in which to *discover* an early-detection signature, because malignant hepatocytes can be resolved according to their genomic and transcriptional identities, whereas bulk RNA sequencing provides a practical setting in which to *deploy* such a signature in clinically accessible tissue specimens. To bridge these settings, we integrated multiple scRNA-seq and snRNA-seq cohorts spanning HCC, NASH-associated HCC and cirrhosis into a unified liver atlas. Malignant cells were identified through concordant copy-number-variation inference and marker-supported annotation, enabling Early-like malignant states to be distinguished from more advanced tumour states. We then applied consensus gene-expression-program analysis to identify coordinated multi-gene modules that separate Early-like malignant cells from the fibrosis/NASH background and track tumour-associated transcriptional changes. Finally, we distilled these programs into a compact gene panel and trained a bulk-RNA-seq tumour-in-background detector designed to retain sensitivity at the low tumour fractions expected in early tissue samples. The choice of HCC versus fibrosis/NASH, rather than HCC versus healthy liver, reflects the intended clinical setting: a molecular signature is useful for early detection only if it can distinguish malignant hepatocyte programs from the chronically injured liver background in which HCC actually arises and is surveyed.

2 Results

2.1 An integrated single-cell and single-nucleus liver atlas spanning HCC, fibrosis and NAFLD/NASH

To establish a discovery framework representative of the clinical setting in which early HCC arises, we assembled an integrated liver atlas from four published datasets spanning healthy liver, cirrhotic adjacent tissue, NAFLD/NASH adjacent tissue and core HCC tumour tissue[5, 15–17]. The atlas comprised 141,428 cells and nuclei from 36 patients across 41 samples, including 128,348 single-nucleus profiles and 13,080 single-cell profiles. The inclusion of multiple tissue contexts, patients and assay modalities exposed subsequent analyses to clinically relevant biological and technical heterogeneity rather than restricting signature discovery to a single cohort or sequencing platform.

Patient-level quality control and scGen integration[19] produced a harmonised representation in which the Leiden solution at resolution 0.8 resolved 28 transcriptionally distinct clusters (Fig. 1A). The integrated embedding reduced dataset-specific separation while retaining substantial structure within the hepatocyte and tumour-associated compartments. Thus, integration did not collapse all between-sample

variation into a homogeneous population, providing a suitable basis for distinguishing disease-associated hepatocyte states from technical batch effects.

Machine-learning-based annotation resolved the major parenchymal, biliary, vascular, stromal and immune compartments (Fig. 1B). The hepatocyte compartment contained a distinct HCC-like state, whereas the non-parenchymal compartment included cholangiocytes, endothelial cells, fibroblasts and pericytes. Immune populations comprised myeloid cells, mast cells, T cells, natural killer cells, B cells and plasma cells. Cycling T/natural-killer cells formed a separate proliferative state, and profiles with weak, mixed or non-coherent lineage evidence remained in an ambiguous or putative-contaminant category rather than being assigned to a definitive lineage.

Tissue-source mapping showed that non-malignant states from healthy, cirrhotic and NAFLD/NASH samples occupied broadly overlapping regions of the integrated representation (Fig. 1C). By contrast, hepatocyte and tumour-associated regions retained clear tissue- and patient-related heterogeneity. Common cellular states were therefore aligned across datasets without obscuring disease-associated variation relevant to tumour detection in chronically injured liver.

Canonical marker expression supported the assigned identities (Fig. 1D). Mature hepatocytes expressed *ALB*, *APOA1* and *TTR*, whereas HCC-like hepatocytes retained hepatic features together with stress-, progenitor- or malignancy-associated programs. Biliary identity was supported by *KRT7*, *KRT19* and *EPCAM*. Endothelial cells expressed *PECAM1*, *VWF*, *EMCN* and *KDR*; fibroblasts expressed collagen and matrix genes, including *COL1A1*, *COL3A1* and *DCN*; and pericytes expressed *RGS5*, *PDGFRB* and *CSPG4*.

Immune identities were likewise supported by lineage-specific programs. Myeloid cells expressed *LST1*, *TYROBP* and *FCER1G*; mast cells expressed *KIT* and *TPSAB1*; and T cells expressed *CD3D*, *CD3E* and T-cell-receptor genes. Natural killer cells expressed *NKG7*, *GNLY* and *KLRD1*; B cells expressed *CD79A*, *MS4A1* and *CD74*; and plasma cells expressed *MZB1* and *JCHAIN*. The cycling lymphocyte state was marked by *MKI67* and *TOP2A*, whereas the ambiguous category lacked a coherent canonical lineage program.

2.2 CNV-based malignancy classification and ordering of hepatocyte and HCC-like states

Distinguishing malignant-state signals from the injury, regeneration and metabolic stress already present in fibrotic or NASH-associated liver required evidence beyond transcriptional identity alone. We therefore combined genome-wide copy-number inference, a continuous CNV-burden measure, malignancy-associated marker expression, functional enrichment and trajectory analysis to construct a graded framework of hepatocyte malignancy.

CopyKatpy analysis identified broad chromosome-scale copy-number alterations within subsets of the hepatocyte and HCC-like compartments (Fig. 2A), while infercnvpy provided complementary CNV evidence for the composite annotation[20, 21]. Putatively malignant cells displayed extensive aneuploid patterns involving multiple chromosomal regions, whereas Background/non-malignant hepatocytes showed comparatively stable inferred copy-number profiles. The genome-wide distribution of these

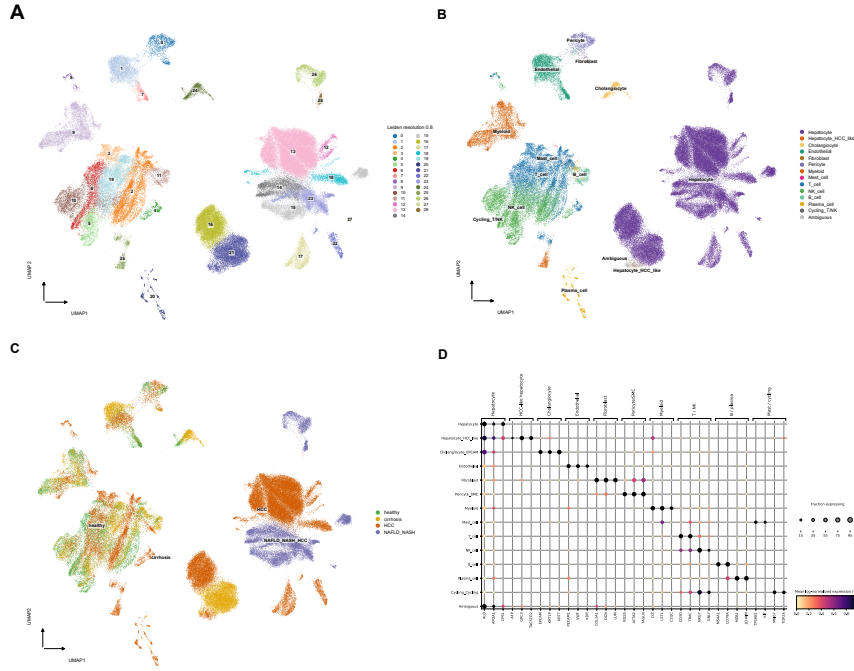


Fig. 1 Construction, integration and annotation of the single-cell and single-nucleus liver atlas. (A) Patient-level quality-controlled data integrated with scGen while retaining sequencing platform and tissue source as covariates. SCCAF-guided evaluation supported Leiden clustering at resolution 0.8, producing 28 clusters. (B) Machine-learning-based annotation of the major parenchymal, biliary, vascular, stromal and immune populations, including a distinct HCC-like hepatocyte state; cycling T/natural-killer and ambiguous or putative-contaminant profiles are shown separately. (C) Distribution of healthy liver, cirrhotic adjacent tissue, NAFLD/NASH adjacent tissue and core HCC tumour tissue across the integrated representation. Shared non-malignant populations co-embed across tissue sources, whereas hepatocyte and tumour-associated compartments retain substantial heterogeneity. (D) Canonical hepatic, biliary, vascular, stromal and immune marker programs supporting the assigned cell identities.

alterations supported inferred aneuploidy as an orthogonal indicator of malignant identity, although transcriptome-derived CNV calls were interpreted as supportive evidence rather than a substitute for direct genomic validation.

Projection of the continuous CNV-burden score onto the hepatocyte embedding revealed a graded distribution rather than two completely separated populations (Fig. 2B). Cells with low, intermediate and high CNV burdens occupied partially

connected regions of the representation, indicating that the inferred malignant signal varied continuously across hepatocyte states. This structure supported retention of intermediate and lower-confidence states instead of imposing a strictly binary malignant/non-malignant classification.

CopyKatpy calls, continuous CNV burden and transcriptional marker evidence were integrated into a composite classification (Fig. 2C). Six operational states were defined. Background/non-malignant cells formed the reference group, while the malignant continuum comprised Early-like, Intermediate-like and Late-like states. Early-like and Late-like cells were further divided into standard- and high-confidence groups. High-confidence states showed concordant genomic and transcriptional evidence, whereas cells with partial or discordant evidence were retained in intermediate or lower-confidence groups. These labels provided the framework for subsequent comparisons among background hepatocytes, Early-like malignant cells and more advanced malignant states.

The damage-associated hepatocyte score^[6] provided an independent biological correlate of this classification (Fig. 2D). Malignant-state groups showed higher damage-associated activity than Background/non-malignant hepatocytes, with a further modest increase across Early-like, Intermediate-like and Late-like states. The strongest overall signal was observed in the Late-like compartment. This trend supported the biological coherence of the CNV-based ordering, but the overlap among states also indicated that damage-associated transcription was not specific to malignancy and could reflect injury programs induced by fibrosis or NASH.

Differential-expression and enrichment analyses further separated the Early-like state from the chronic-liver-disease background (Fig. 2E). Early-like hepatocytes were associated with mitotic and cell-cycle processes, including spindle-checkpoint signalling, mitotic cytoskeletal organisation and cytokinesis. In contrast, Background/non-malignant hepatocytes were enriched for inflammatory and metabolic processes associated with chronic liver injury, including complement and coagulation cascades, NAFLD-related pathways, PPAR signalling, phagocytosis, interleukin-6 regulation and thermogenesis. ATP biosynthesis was also represented among the enriched processes, consistent with altered energetic requirements across the comparison. These enrichments distinguish a proliferation-associated Early-like program from an injury- and metabolism-dominated background but do not establish temporal causality or identify an individual pathway as an initiator of HCC.

Trajectory inference using PAGA, Monocle 2 and Slingshot^[22–24] recovered a broadly concordant state-ordering structure despite the different graph- and curve-based assumptions of the three methods (Fig. 2F). Background/non-malignant hepatocytes were concentrated near the inferred trajectory origin, Early-like cells occupied nearby positions along pseudotime and Late-like cells were distributed further along the continuum. The agreement among complementary trajectory approaches was compatible with a shared background-to-malignant state axis. Because the data were cross-sectional, however, this ordering should be interpreted as a model of transcriptional-state relationships rather than direct evidence of a longitudinal or clonal hepatocyte-to-tumour transition.

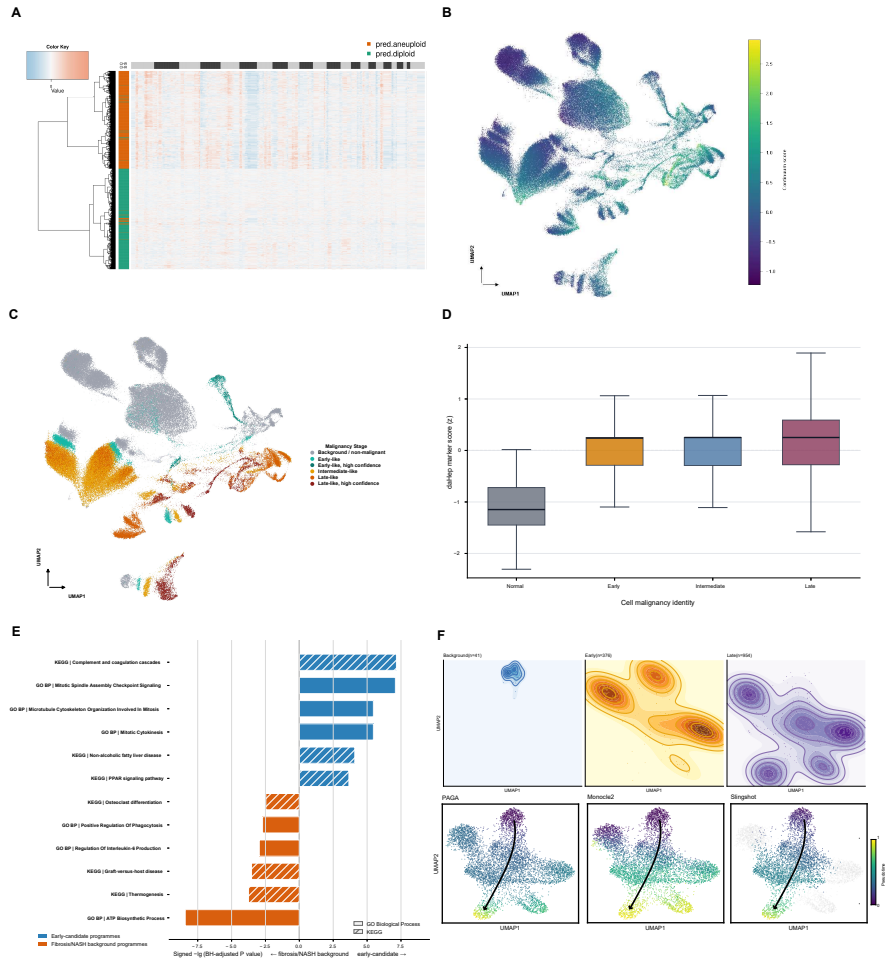


Fig. 2 CNV-based malignancy classification and state-ordering analysis of hepatocyte and HCC-like states. (A) CopyKatpy-inferred genome-wide copy-number profiles showing broad aneuploid patterns in putatively malignant hepatocytes and comparatively stable profiles in predicted diploid cells. (B) Continuous CNV-burden score projected onto the hepatocyte embedding, revealing a graded malignancy-associated axis. (C) Composite classification into Background/non-malignant, Early-like, Early-like high-confidence, Intermediate-like, Late-like and Late-like high-confidence states using CNV and transcriptional evidence. (D) Distribution of the damage-associated hepatocyte score across the malignancy-state groups. (E) Functional enrichment of direction-specific genes from the Early-like versus Background/non-malignant comparison. (F) State-ordering analyses using PAGA, Monocle 2 and Slingshot, with Background/non-malignant cells positioned near the inferred origin and more advanced malignant states distributed further along pseudotime.

2.3 Conserved gene-expression programs yield a directionally locked 12-gene signature

Consensus gene-expression-program analysis resolved transcriptional structure that was not captured by single-gene differential expression alone. The recurrent meta-programs fell into two conceptually distinct classes: fibrosis/NASH reference programs, which represented the chronic-injury context against which malignant cells must be detected, and tumour-internal programs, which described heterogeneity among CNV-defined malignant hepatocytes. This distinction was important because extracellular-matrix remodelling, inflammation, complement/coagulation activity, altered lipid handling and hypoxic stress may occur extensively in non-malignant chronic liver disease and therefore cannot be interpreted as evidence of malignancy on the basis of marker overlap alone.

Repeated sample-wise NMF[25, 26] identified stable programs that recurred across factorisation ranks and independent samples (Fig. 3A). Meta-program consolidation retained 174 background-associated genes and 187 candidate Early-like-associated genes. Collectively, these genes represented coordinated processes involving hypoxia, proliferation, epithelial-mesenchymal transition, coagulation, inflammation, drug metabolism and cholesterol metabolism. The correlation structure separated groups of recurrent programs while also revealing partial sharing across the background and malignant compartments. Thus, the principal value of the analysis was not to designate every program gene as tumour specific, but to define conserved modules and the chronic-liver-disease constraints required for subsequent feature filtering.

Directional survival screening using Cox proportional-hazards models[27, 28] in TCGA-LIHC and ICGC-LIRI-JP reduced the conserved program genes to a locked 12-gene panel (Fig. 3B and Table 1). Nine genes—*CFL1*, *PRDX1*, *CCT3*, *S100A10*, *CLTC*, *PPIA*, *ATP1B3*, *ACTG1* and *HSP90AB1*—were assigned a positive orientation, whereas *ALDOB*, *CYP2C8* and *AFM* were assigned a negative orientation. The pooled hazard ratios were directionally consistent with this assignment and remained significant after false-discovery-rate correction. These orientations were used to lock the feature direction before classifier training; they describe reproducible survival-associated expression signals and do not demonstrate that individual genes cause or prevent tumour initiation. Because the number of observed deaths was lower in ICGC, the external-cohort estimates were interpreted primarily as support for directional reproducibility.

The locked panel was subsequently benchmarked across 101 machine-learning methods or model specifications using patient-grouped 10-fold cross-validation on synthetic tumour-background mixtures[28, 29]. Extra-Trees[30] ranked first and achieved an out-of-fold AUROC of 0.9983 (Fig. 3C). The grouped design prevented mixtures derived from the same tumour donor from being split across training and validation folds. Performance across individual tumour fractions and the scope of cross-cohort transfer are reported in the following subsection.

The importance profile indicated that classification combined increased tumour signals with reduced mature hepatocyte identity (Fig. 3D). *CCT3* contributed most strongly, followed by *HSP90AB1*, *CYP2C8*, *PPIA*, *AFM* and *ALDOB*. The prominence of positively oriented genes was consistent with increased tumour activity. By

Table 1 Directionally locked 12-gene panel derived from pooled survival screening.

Gene	Signature orientation	Pooled hazard ratio	Pooled FDR
<i>CFL1</i>	Positive	1.31	0.0030
<i>CCT3</i>	Positive	1.25	0.0044
<i>PRDX1</i>	Positive	1.31	0.0044
<i>S100A10</i>	Positive	1.24	0.0052
<i>CLTC</i>	Positive	1.27	0.0061
<i>PPIA</i>	Positive	1.25	0.0084
<i>ATP1B3</i>	Positive	1.30	0.0084
<i>ACTG1</i>	Positive	1.27	0.0148
<i>HSP90AB1</i>	Positive	1.23	0.0261
<i>CYP2C8</i>	Negative	0.70	0.0044
<i>ALDOB</i>	Negative	0.71	0.0044
<i>AFM</i>	Negative	0.69	0.0061

contrast, the high importance of *CYP2C8*, *AFM* and *ALDOB* indicated that reduced hepatocyte metabolic differentiation also contributed substantially to classification. The detector therefore integrated two complementary signals: acquisition of malignant transcriptional programs and erosion of mature liver identity.

2.4 Low-fraction tumour detection and prognostic validation of the 12-gene signature

We next evaluated two complementary properties of the locked 12-gene signature: its analytical sensitivity to a diluted tumour-derived signal and its ability to stratify overall survival. The synthetic-mixture analyses assessed tumour detection across decreasing tumour content, whereas the time-dependent receiver operating characteristic and Kaplan–Meier analyses assessed prognostic information in TCGA-LIHC and ICGC-LIRI-JP. These endpoints were analysed separately because successful tumour detection and outcome discrimination are related but non-equivalent properties.

Detection probability increased with nominal tumour fraction in both cohorts (Fig. 4A,B). In TCGA mixtures, the median predicted probability approached 0.80 at 5% tumour content and rose rapidly towards saturation at higher fractions. In ICGC mixtures, the median probability at 5% was approximately 0.55, close to the locked decision threshold of 0.555, and increased progressively as tumour content increased. The reduced probability at low fractions in ICGC was consistent with cross-cohort differences in expression distribution and tissue composition, but the monotonic response showed that the transferred model remained sensitive to increasing tumour-derived signal.

The nominal mixing fraction should not be interpreted as a directly measured malignant-cell fraction. Both the tumour and non-tumour components were bulk RNA-sequencing profiles, and the tumour profiles themselves contained malignant cells

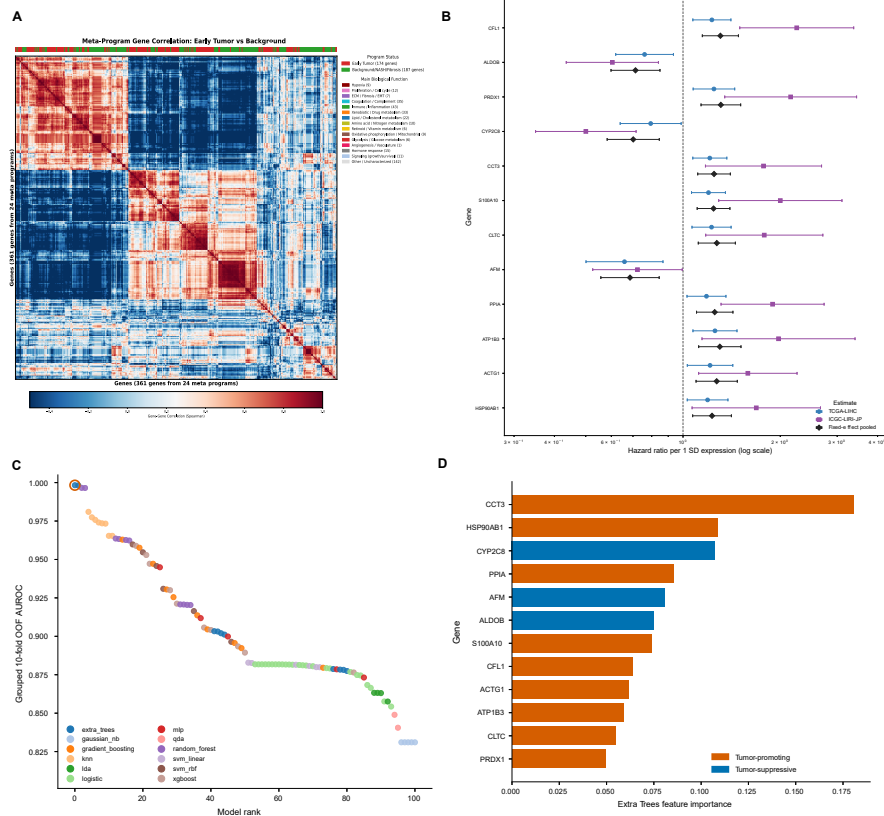


Fig. 3 Conserved gene-expression programs and derivation of the directionally locked 12-gene signature. (A) Correlation structure of recurrent meta-program genes identified by repeated sample-wise NMF in chronic-liver-disease background and malignant hepatocytes. (B) Cohort-specific and pooled hazard-ratio estimates used to assign the direction of the 12 retained genes in TCGA-LIHC and ICGC-LIRI-JP. (C) Out-of-fold AUROC ranking of 101 machine-learning methods or model specifications under patient-grouped 10-fold cross-validation, with Extra-Trees achieving the highest discrimination. (D) Extra-Trees feature importance, showing contributions from positively oriented tumour genes and negatively oriented genes reflecting loss of mature hepatocyte metabolic identity.

together with normal hepatocytes, stromal cells and immune cells. A mixture containing 5% bulk tumour tissue therefore contained less than 5% pure malignant-cell RNA. The retained signal at this dilution supports low-fraction analytical sensitivity within a complex liver-tissue background, but it does not establish a histologically verified or clinical limit of detection. As an orthogonal tissue-level assessment, the direction-aligned score distinguished TCGA primary tumours from solid-tissue-normal samples with an AUROC of 0.947 (95% confidence interval, 0.921–0.968).

In 360 TCGA-LIHC patients with 129 deaths, the risk score showed moderate and relatively stable time-dependent discrimination (Fig. 4C). The 1-year AUROC was 0.697 (95% confidence interval, 0.620–0.771), the 3-year AUROC was 0.663 (0.584–0.741) and the 5-year AUROC was 0.670 (0.571–0.768). The similar point estimates and overlapping confidence intervals indicated that prognostic information was retained across short- and longer-term follow-up rather than being confined to a single time horizon.

Time-dependent discrimination varied more strongly in the 203-patient ICGC cohort (Fig. 4D). The 1-year AUROC was 0.670 (95% confidence interval, 0.493–0.818), increasing to 0.788 (0.685–0.881) at 3 years before declining to 0.359 (0.183–0.591) at 5 years. The broad confidence intervals at 1 and 5 years reflected the smaller number of informative events and patients remaining under observation at these horizons. In particular, the 5-year interval included chance performance and did not support durable late prognostic discrimination. The clearest external-cohort performance was therefore observed at the intermediate 3-year horizon.

Kaplan–Meier analysis further evaluated broad risk-group separation. In TCGA, the median risk-score cutpoint of -0.028 divided the cohort into two groups of 180 patients (Fig. 4E). The high-risk group experienced 84 deaths compared with 45 deaths in the low-risk group, and the survival curves separated early and remained distinct during follow-up (log-rank $P < 0.001$). This separation was consistent with the time-dependent AUROC estimates and indicated persistently poorer overall survival among patients with higher signature scores.

The same TCGA-derived cutpoint was transferred to ICGC without cohort-specific re-optimisation (Fig. 4F). It classified 109 patients as high risk and 94 as low risk. The high-risk group experienced 27 deaths, compared with 8 deaths in the low-risk group, and overall survival differed significantly between the groups (log-rank $P < 0.001$). Separation under a fixed external threshold provided evidence that the risk grouping was transportable across cohorts, even though time-dependent discrimination varied by follow-up horizon.

Together, these analyses support two distinct conclusions. First, the locked Extra-Trees model remained responsive when bulk tumour RNA was strongly diluted by non-tumour liver RNA, including at a nominal 5% tumour fraction. Second, the directionally aligned 12-gene score carried prognostic information that separated broad survival-risk groups in both cohorts. The prognostic findings do not convert the survival score into an early-detection endpoint, and the synthetic mixtures do not reproduce all sources of variation expected in prospective fibrosis/NASH surveillance. Clinical validation will require independently collected specimens with measured tumour-cell content, prespecified processing and longitudinal outcomes.

2.5 The 12-gene risk score is associated with an immune-infiltrated but exhausted tumour microenvironment

To investigate biological processes associated with the prognostic signal, we examined the relationship between the continuous 12-gene risk score and inferred immune and stromal infiltration in TCGA-LIHC and ICGC-LIRI-JP. Across both cohorts,

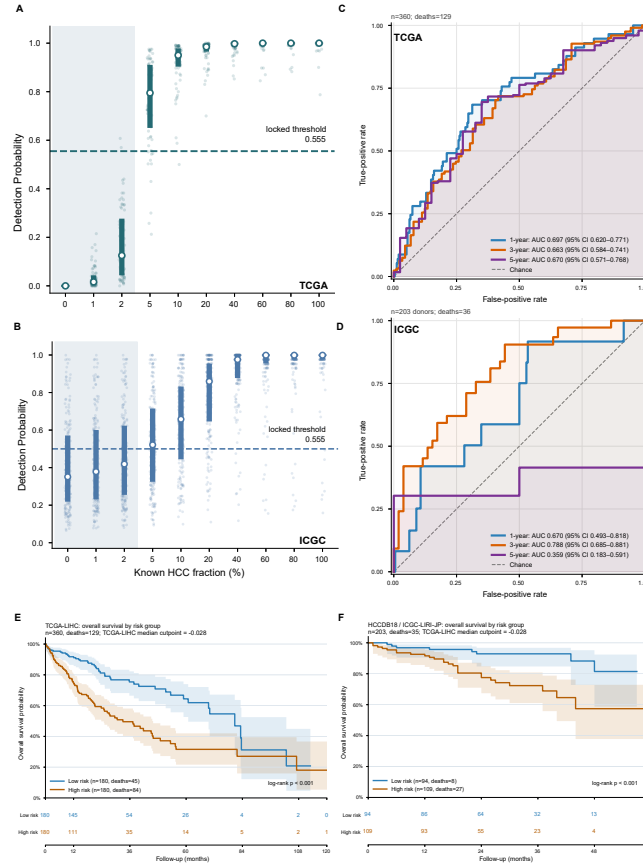


Fig. 4 Low-fraction tumour detection and prognostic validation of the 12-gene signature. (A,B) Detection probability across synthetic mixtures containing 0–100% bulk tumour tissue in TCGA-LIHC and ICGC-LIRI-JP, respectively. The Extra-Trees decision threshold was locked at 0.555. At 5% nominal tumour content, the median probability was approximately 0.80 in TCGA and 0.55 in ICGC. (C) TCGA time-dependent ROC analysis: 1-year AUROC, 0.697 (95% confidence interval, 0.620–0.771); 3-year AUROC, 0.663 (0.584–0.741); and 5-year AUROC, 0.670 (0.571–0.768). (D) ICGC time-dependent ROC analysis: 1-year AUROC, 0.670 (0.493–0.818); 3-year AUROC, 0.788 (0.685–0.881); and 5-year AUROC, 0.359 (0.183–0.591). (E) TCGA Kaplan-Meier analysis using the median risk-score cutpoint of -0.028 : low risk, $n = 180$ with 45 deaths; high risk, $n = 180$ with 84 deaths; log-rank $P < 0.001$. (F) External application of the same cutpoint in ICGC: low risk, $n = 94$ with 8 deaths; high risk, $n = 109$ with 27 deaths; log-rank $P < 0.001$. Nominal tumour fractions refer to bulk tumour tissue and not purified malignant-cell content.

higher risk scores were associated with greater overall immune infiltration rather than an immune-desert phenotype, although the strength and reproducibility of the associations differed among cellular compartments.

In TCGA-LIHC, the risk score correlated positively with pan-immune infiltration (Spearman’s $\rho = 0.21$, FDR = 9.14×10^{-5}) and T-cell infiltration ($\rho = 0.21$, FDR = 9.14×10^{-5} ; Fig. 5A). The strongest association involved the myeloid/macrophage compartment ($\rho = 0.36$, FDR = 1.49×10^{-11}). Smaller positive associations were observed for B-cell infiltration ($\rho = 0.16$, FDR = 0.00406) and stromal/fibroblast infiltration ($\rho = 0.11$, FDR = 0.0473). By contrast, the cytotoxic/natural-killer-cell association was weak and did not remain significant after multiple-testing correction ($\rho = 0.06$, FDR = 0.219).

The ICGC cohort showed a similar overall pattern (Fig. 5A). Higher risk scores were associated with pan-immune infiltration ($\rho = 0.21$, FDR = 0.00684), T-cell infiltration ($\rho = 0.17$, FDR = 0.0263) and myeloid/macrophage infiltration ($\rho = 0.28$, FDR = 0.000342). Associations with cytotoxic/natural-killer cells ($\rho = 0.07$, FDR = 0.328), B cells ($\rho = 0.12$, FDR = 0.0908) and stromal/fibroblast cells ($\rho = 0.14$, FDR = 0.0709) did not pass FDR correction. Increased T-cell and myeloid/macrophage infiltration were therefore the most reproducible cellular correlates of a high-risk score across cohorts.

The positive T-cell association did not imply more effective antitumour immunity. T-cell-exhaustion scores increased significantly with the 12-gene risk score in TCGA ($\rho = 0.32$, FDR = 4.10×10^{-8}) and were independently reproduced in ICGC ($\rho = 0.23$, FDR = 0.00291; Fig. 5B). This concordance helps reconcile the combination of greater T-cell-associated signal and poorer survival: an immune-rich tumour microenvironment may still be functionally compromised when infiltrating T cells exhibit an exhaustion program.

Together, these results indicate that high-risk tumours are immune infiltrated rather than immune excluded. Increased T-cell exhaustion and myeloid/macrophage infiltration suggest that abundant immune cells may have impaired effector activity or may support tumour-promoting inflammation, consistent with cell-resolved descriptions of dysfunctional T-cell states in liver cancer[31]. Because the scores were derived from bulk transcriptomes, they represent relative molecular associations and do not directly enumerate immune cells or establish their functional state at single-cell resolution.

3 Discussion

The main contribution of this study is a disease-aware route from cell-resolved discovery to a compact tissue-RNA panel. Early HCC must be detected within a liver already altered by injury, regeneration, inflammation and fibrosis, not against an idealised healthy reference[3, 5–8]. Treating these background programs as a specificity constraint changes the biomarker question from whether a panel distinguishes tumour from normal tissue to whether it retains tumour-associated information after chronic-liver-disease signals and tissue dilution are considered. The 12-gene panel is therefore

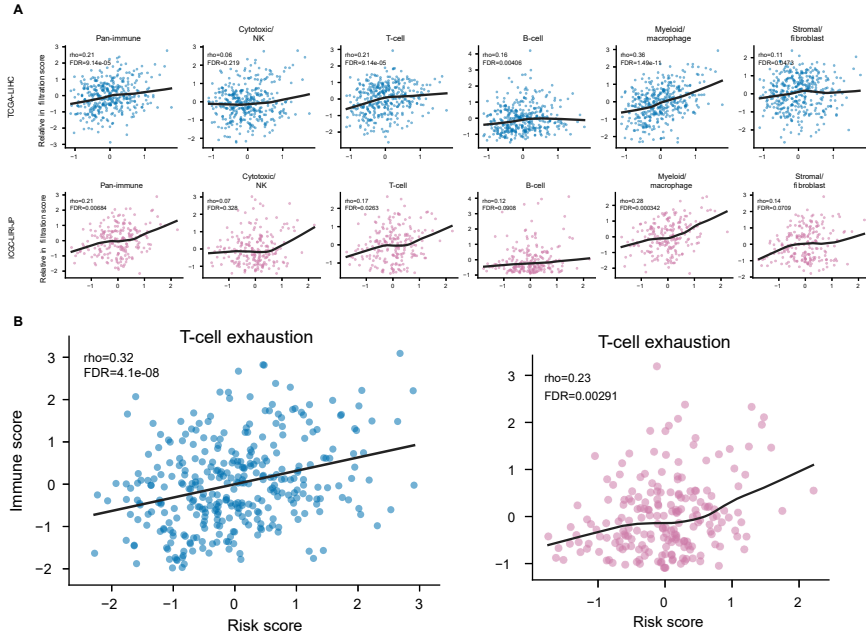


Fig. 5 Association of the 12-gene risk score with immune infiltration and T-cell exhaustion. (A) Spearman correlations between the risk score and relative infiltration scores in TCGA-LIHC (top) and ICGC-LIRI-JP (bottom). In TCGA, significant positive associations were observed for pan-immune, T-cell, B-cell, myeloid/macrophage and stromal/fibroblast infiltration, whereas cytotoxic/natural-killer-cell infiltration was not significant after FDR correction. In ICGC, significant associations were reproduced for pan-immune, T-cell and myeloid/macrophage infiltration; the cytotoxic/natural-killer-cell, B-cell and stromal/fibroblast associations did not pass FDR correction. (B) T-cell-exhaustion scores correlated positively with the risk score in TCGA ($\rho = 0.32$, $\text{FDR} = 4.10 \times 10^{-8}$) and ICGC ($\rho = 0.23$, $\text{FDR} = 0.00291$). Black lines show fitted trends; correlation statistics are Spearman coefficients with Benjamini–Hochberg-adjusted probabilities.

most important as a testable bridge between single-cell disease biology and an assay format that could ultimately be evaluated in surveillance specimens.

Two design choices strengthen that bridge. First, combining inferred copy-number alterations with transcriptional evidence reduces reliance on markers that may also be induced in reactive hepatocytes. Second, selecting recurrent programs rather than isolated differentially expressed genes favours coordinated signals that persist across patients and assay modalities[20, 25, 26]. The bidirectional panel also captures both acquisition of tumour-associated expression and erosion of mature hepatocyte identity, an informative feature when malignant RNA is scarce. These choices improve biological plausibility, but they do not establish tumour initiation mechanisms. “Early-like” denotes an operational state with weaker CNV and transcriptional evidence near

the background state; cross-sectional pseudotime cannot convert that ordering into lineage, chronology or clinical stage[22–24].

The bulk analyses should consequently be viewed as stress tests of the discovery framework rather than as clinical validation. Synthetic dilution asks whether tumour-associated signal survives increasing background, but the nominal mixture fraction is not a measured malignant-cell fraction and does not define a clinical limit of detection. Moreover, the ICGC analysis used quantile-mapped tumour profiles and the TCGA normal background, while both cohorts contributed to survival-based feature orientation. It is therefore a cross-cohort transfer analysis, not an untouched external validation of specificity in fibrosis or NASH. Prognostic separation and the association with an immune-infiltrated, exhaustion-associated state add biological context, but neither establishes surveillance performance or predicts response to immunotherapy. Bulk immune scores cannot distinguish changes in cell abundance from altered cellular state, and the survival endpoint is distinct from early detection[7, 8, 31].

Several limitations set the evidential boundary. Public single-cell datasets differed in platform, tissue processing and sampling, so residual cohort effects may remain after integration. Malignant identity was inferred from RNA rather than confirmed by matched DNA sequencing, and the cross-sectional, tumour-enriched atlas cannot estimate lead time or reconstruct within-patient evolution. Disease controls were too limited to represent the aetiological and histological diversity of cirrhosis and metabolic liver disease. At the assay level, computational mixtures omit RNA degradation, sampling error, field effects and shifts in stromal and immune composition; reuse of the TCGA background further limits claims of cross-cohort generalisation. Model selection among many specifications may also yield optimistic performance unless the entire selection and thresholding process is evaluated in a fully nested or external design. Finally, the panel was derived from tissue RNA. Its behaviour cannot be extrapolated to blood RNA, circulating DNA or serum proteins without direct measurement.

Future work should proceed in stages. Analytical validation should first lock the assay, preprocessing, classifier and operating threshold, then test independent HCC and disease-control tissues with measured tumour purity under blinded conditions. The decisive study is a prospective fibrosis/NASH surveillance cohort enriched for early-stage and AFP-negative HCC, with longitudinal time-to-diagnosis, matched imaging and serum biomarkers, and prespecified handling of indeterminate cases. Evaluation should emphasise calibration and specificity at clinically acceptable thresholds, incremental value over AFP or composite scores, and net clinical benefit rather than AUROC alone; prospective biomarker studies illustrate why these quantities must be fixed before outcome assessment[11, 13]. Sampling should also reflect MASLD-associated HCC outside conventional cirrhosis-based pathways[32].

Mechanistic validation is a separate priority. Matched single-cell RNA and DNA profiling, longitudinal sampling and phylogenetic analysis could determine whether the Early-like state precedes more advanced malignant states. Spatial or cell-resolved immune profiling could test whether high-score tumours contain exhausted T cells and immunosuppressive myeloid niches, ideally in treatment-linked cohorts. Comparisons with circulating cell-free DNA and other emerging omics assays would clarify whether

the tissue panel offers complementary information or mainly recapitulates existing signals[4, 14].

In summary, this study establishes a compact, disease-aware discovery framework and a candidate tissue-RNA panel. Its value lies in defining a focused hypothesis for rigorous prospective testing, not in supporting immediate use as a screening test.

4 Materials and methods

4.1 Study design and data sources

This study used a staged discovery and validation design. Single-cell and single-nucleus RNA-sequencing datasets deposited in the Gene Expression Omnibus under accession numbers GSE125449, GSE151530, GSE189175 and GSE212046 were used to resolve malignant and non-malignant hepatocyte states and identify recurrent transcriptional programs[5, 15–17]. Bulk RNA-sequencing data from The Cancer Genome Atlas Liver Hepatocellular Carcinoma cohort (TCGA-LIHC) and the International Cancer Genome Consortium Liver Cancer—RIKEN, Japan cohort (ICGC-LIRI-JP) were subsequently used for directional gene screening, synthetic dilution experiments, classifier development, survival analysis and immune-context analysis[33, 34]. The integrated atlas comprised 141,428 cells and nuclei from 36 patients and 41 samples, including 128,348 profiles generated by single-nucleus RNA sequencing and 13,080 generated by single-cell RNA sequencing. Tissue categories comprised healthy liver, cirrhotic adjacent tissue, NAFLD/NASH adjacent tissue and core HCC tissue.

4.2 Patient-level quality control and atlas integration

Quality control was performed separately for each patient to accommodate differences among studies, tissue preparations and single-cell versus single-nucleus assays. Processing and data handling were implemented with Scanpy and Seurat[35, 36], together with MedHorizon utilities. Low-complexity profiles, profiles with disproportionate mitochondrial signal and likely multiplets or mixed profiles were excluded according to patient-specific distributions of gene detection, transcript counts and library complexity. After within-sample preprocessing, datasets were integrated using scGen[19]. Dataset-specific effects were attenuated while sequencing platform and tissue source were retained as covariates, thereby reducing technical variation without removing biologically relevant differences associated with assay modality or disease context.

Clustering solutions were compared using the Single Cell Clustering Assessment Framework (SCCAF)[37], after which Leiden clustering was performed at a resolution of 0.8. Cell identities were assigned using a machine-learning annotation workflow and checked against canonical lineage programs. The annotation reference encompassed hepatocyte, cholangiocyte, endothelial, fibroblast, pericyte, myeloid, mast-cell, T-cell, natural-killer-cell, B-cell and plasma-cell programs, together with HCC-associated hepatocyte markers. Cycling lymphocytes and profiles without coherent lineage evidence were retained as separate cycling and ambiguous or putative-contaminant categories rather than being forced into a mature lineage.

4.3 Copy-number inference and malignant-state annotation

Hepatocytes and HCC-like hepatocytes were evaluated for large-scale copy-number alterations using two complementary transcriptome-based approaches implemented in infercnvpy and CopyKatpy[20, 21]. Genome-wide aneuploid patterns provided orthogonal evidence of malignant identity. Copy-number burden was additionally represented as a continuous cell-level score to preserve intermediate states rather than imposing a strictly binary malignant/non-malignant boundary.

Copy-number calls, continuous copy-number burden and canonical malignancy-associated transcriptional features were combined to define six operational states: Background/non-malignant, Early-like, Early-like high-confidence, Intermediate-like, Late-like and Late-like high-confidence. High-confidence labels required concordant genomic and transcriptional evidence, whereas cells with partial or discordant evidence were retained in intermediate or lower-confidence categories. A damage-associated hepatocyte score[6] was calculated as an additional biological correlate but was not used as a stand-alone malignancy criterion because injury-associated programs may also be induced by fibrosis and NASH.

4.4 Differential expression, functional enrichment and trajectory inference

Differential-expression analysis was performed with Scanpy and Seurat[35, 36] to compare Background/non-malignant and Early-like hepatocytes. Genes were ranked by effect size and statistical support, and multiple comparisons were controlled using the Benjamini–Hochberg procedure[38]. Functional enrichment was then applied separately to direction-specific gene sets to characterise the biological processes associated with each cell state. Enrichment terms were treated as population-level associations and not as evidence that an individual pathway initiates malignant transformation.

Cell-state relationships were examined using three complementary trajectory approaches: partition-based graph abstraction (PAGA) in Scanpy, Monocle 2 and Slingshot[22–24]. Background-labelled hepatocytes were specified as the reference state for pseudotime orientation, and malignant-state labels were projected onto each inferred trajectory. Concordance among graph-based and curve-based approaches was used to assess the robustness of the resulting state ordering. Because all datasets were cross-sectional, trajectories were interpreted as hypotheses of state ordering rather than observed lineage histories.

4.5 Sample-wise gene-expression-program discovery

Gene-expression-program discovery was performed using the CSI-GEP workflow[26]. Non-negative matrix factorisation[25] was applied separately within each sample to chronic-liver-disease background hepatocytes and malignant hepatocytes. Factorisation was repeated across multiple ranks, and programs were compared across runs and samples. Programs were matched on the basis of gene-loading similarity, and those recurring across factorisation settings and independent samples were grouped into meta-programs. Genes contributing consistently to background-associated or

malignant-state meta-programs were carried forward as candidates for downstream screening.

Program interpretation was constrained by direct state comparisons. For retained cross-assay programs, stage direction was evaluated from the difference in mean patient-run usage between Early-like and Late-like malignant cells. Positive Early-minus-Late values indicated Early-like association. Programs without direct early enrichment were not described as early drivers, and background specificity was not inferred where an assay-balanced background-usage matrix was unavailable.

4.6 Directional survival screening and definition of the 12-gene panel

Candidate program genes were evaluated by univariable Cox proportional-hazards analysis[27] in TCGA-LIHC and ICGC-LIRI-JP using the survival package and the Mime modelling framework[28]. Cohort-specific hazard ratios were estimated independently and compared for directional consistency. Pooled effect estimates and multiple-testing-adjusted probabilities were used as a replication and feature-locking procedure. Genes meeting the prespecified consistency criteria were assigned a fixed positive or negative orientation and assembled into the locked 12-gene panel before classifier evaluation. The survival screen was used only to define a reproducible direction of association and was not interpreted as evidence that individual genes causally determine tumour initiation or clinical outcome.

4.7 Synthetic tumour-background mixtures and machine-learning evaluation

Bulk tumour profiles were computationally mixed with non-tumour liver profiles at nominal tumour fractions of 0%, 1%, 2%, 5%, 10%, 20%, 40%, 60%, 80% and 100%. TCGA solid-tissue-normal samples provided the background for both the TCGA mixtures and the ICGC transfer experiment. Before ICGC mixtures were constructed, ICGC tumour profiles were feature-wise quantile-mapped to the TCGA tumour-expression scale. The ICGC experiment therefore assessed transfer of the tumour-associated signal but did not constitute fully independent background validation.

The locked 12-gene panel was evaluated across 101 machine-learning methods or model specifications using Mime and scikit-learn[28, 29]. Mixtures with a nominal tumour fraction of at least 5% were assigned to the detection-positive class. Model comparison used patient-grouped 10-fold cross-validation so that mixtures derived from the same tumour donor could not be divided between training and validation folds. Out-of-fold AUROC was the primary model-selection metric, and all preprocessing and model fitting were performed within the corresponding training folds. The Extra-Trees specification[30] used 400 trees with unrestricted depth. An operating threshold was estimated using TCGA only and locked before application to the ICGC mixtures. Feature importance was extracted from the fitted classifier and interpreted as model-specific predictive contribution rather than causal importance.

As an orthogonal tissue-level assessment, the direction-aligned 12-gene score was evaluated in TCGA primary tumours and solid-tissue-normal samples. AUROC, 95% confidence intervals, sensitivity and specificity were reported. Synthetic-mixture performance was interpreted as analytical sensitivity to a diluted bulk-tumour signal and not as a histologically measured malignant-cell limit of detection.

4.8 Prognostic analysis

A continuous risk score was calculated from the locked 12-gene signature using the pre-specified gene directions. Time-dependent receiver operating characteristic curves were estimated at 1, 3 and 5 years in TCGA-LIHC and ICGC-LIRI-JP using timeROC[39]. For Kaplan–Meier analysis[40], the median TCGA risk score defined the high- and low-risk groups. This cutpoint was transferred unchanged to ICGC, and no external-cohort threshold optimisation was performed. Overall-survival distributions were estimated and compared using survival and survminer, with group differences assessed by the log-rank test. Time-dependent discrimination and risk-group separation were treated as prognostic analyses distinct from the tumour-detection endpoint.

4.9 Immune-infiltration and T-cell-exhaustion analyses

Relative transcriptomic infiltration scores were evaluated for overall immune, cytotoxic and natural-killer, T-cell, B-cell, myeloid and macrophage, and stromal and fibroblast compartments in TCGA and ICGC. T-cell exhaustion was evaluated using a transcriptome-derived exhaustion score, with biological interpretation informed by published single-cell analyses of liver-cancer T cells[31]. Associations between the continuous 12-gene risk score and each immune or stromal score were quantified by Spearman rank correlation. Multiple comparisons were controlled using the Benjamini–Hochberg false-discovery rate[38]. These analyses were interpreted as bulk-transcriptomic associations; they do not directly enumerate immune cells or establish their functional states at single-cell resolution.

4.10 Statistical analysis and interpretation

All statistical tests were two-sided unless otherwise specified. Multiple-testing-adjusted values are reported as false-discovery rates according to the Benjamini–Hochberg procedure[38]. Classification performance was summarised using AUROC and, at the locked operating threshold, sensitivity and specificity. Survival discrimination was assessed using time-dependent AUROC with 95% confidence intervals and Kaplan–Meier analysis with log-rank testing[39, 40]. Correlation analyses used Spearman’s rho. The patient or donor, rather than an individual synthetic mixture, was retained as the independent grouping unit during cross-validation. Analyses of public, de-identified datasets were conducted as secondary computational research.

4.11 Software and computational resources

The software packages and source-code repositories used in the computational analyses are listed in Table 2.

Table 2 Software packages, versions and source-code repositories used in the analysis.

Package	Version	GitHub repository
CSI-GEP	Unreleased (main)	https://github.com/geeleherlab/CSI-GEP
scGen	2.1.0	https://github.com/theislab/scgen
SCCAF	0.0.10	https://github.com/SCCAF/sccaf
MedHorizon	Unreleased (main)	https://github.com/medhorizon/medhorizon
CopyKatpy	0.1.0 (source)	https://github.com/medhorizon/CopyKatpy
infercnvpy	0.6.1	https://github.com/icbi-lab/infercnvpy
Scanpy	1.12.3	https://github.com/scverse/scanpy
Seurat	5.5.1	https://github.com/satijalab/seurat
Monocle 2	2.40.0	https://github.com/cole-trapnell-lab/monocle-release
Slingshot	2.20.0	https://github.com/kstreet13/slinsshot
Mime	0.0.0.9000 (development)	https://github.com/l-magnificence/Mime
scikit-learn	1.9.0	https://github.com/scikit-learn/scikit-learn
survival	3.8-9	https://github.com/therneau/survival
survminer	0.5.2	https://github.com/kassambara/survminer
timeROC	0.4.1	https://github.com/cran/timeROC

References

- [1] Sung, H., Ferlay, J., Siegel, R. L. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
- [2] Singal, A. G., Llovet, J. M., Yarrow, M. *et al.* AASLD practice guidance on prevention, diagnosis, and treatment of hepatocellular carcinoma. *Hepatology* **78**, 1922–1965 (2023).
- [3] European Association for the Study of the Liver, European Association for the Study of Diabetes & European Association for the Study of Obesity. EASL–EASD–EASO clinical practice guidelines on the management of metabolic dysfunction-associated steatotic liver disease (MASLD). *J. Hepatol.* **81**, 492–542 (2024).
- [4] Lan, T. & Tacke, F. Diagnostics and omics technologies for the detection and prediction of metabolic dysfunction-associated steatotic liver disease-related malignancies. *Metabolism* **161**, 156015 (2024).
- [5] Filliol, A., Saito, Y., Nair, A. *et al.* Opposing roles of hepatic stellate cell subpopulations in hepatocarcinogenesis. *Nature* **610**, 356–365 (2022).
- [6] Carlessi, R., Denisenko, E., Boslem, E. *et al.* Single-nucleus RNA sequencing of pre-malignant liver reveals disease-associated hepatocyte state with HCC prognostic potential. *Cell Genom.* **3**, 100301 (2023).
- [7] Martin, O. P., Wallace, M. S., Oetheimer, C. *et al.* Single-cell atlas of human liver and blood immune cells across fatty liver disease stages reveals distinct signatures linked to liver dysfunction and fibrogenesis. *Nat. Immunol.* **26**, 1596–1611 (2025).

- [8] Huang, Y., Xie, Y., Zhang, Y. *et al.* Single-cell transcriptome reveals the reprogramming of immune microenvironment during the transition from MASH to HCC. *Mol. Cancer* **24**, 177 (2025).
- [9] Tzartzeva, K., Obi, J., Rich, N. E. *et al.* Surveillance imaging and alpha fetoprotein for early detection of hepatocellular carcinoma in patients with cirrhosis: a meta-analysis. *Gastroenterology* **154**, 1706–1718.e1 (2018).
- [10] Singal, A. G., Ng, M. & Kulkarni, A. Advancing surveillance strategies for hepatocellular carcinoma: a new era of efficacy and precision. *J. Clin. Exp. Hepatol.* **14**, 101448 (2024).
- [11] Hirode, G., Zangneh, H. F., Cerocchi, O. *et al.* Utility of serum biomarkers in addition to ultrasound for early detection of hepatocellular carcinoma in high-risk patients: a randomized controlled trial. *Gastroenterology* **170**, 595–605 (2026).
- [12] Best, J., Bechmann, L. P., Sowa, J.-P. *et al.* GALAD score detects early hepatocellular carcinoma in an international cohort of patients with nonalcoholic steatohepatitis. *Clin. Gastroenterol. Hepatol.* **18**, 728–735.e4 (2020).
- [13] Marsh, T. L., Parikh, N. D., Roberts, L. R. *et al.* A phase 3 biomarker validation of GALAD for the detection of hepatocellular carcinoma in cirrhosis. *Gastroenterology* **168**, 316–326.e6 (2025).
- [14] Peng, J. & Pan, X. Genome-wide 5-hydroxymethylcytosine in circulating cell-free DNA reflects DNA methylation dynamics for non-invasive detection of early hepatocellular carcinoma. *Chin. Clin. Oncol.* **15**, 32 (2026).
- [15] Ma, L., Hernandez, M. O., Zhao, Y. *et al.* Tumor cell biodiversity drives microenvironmental reprogramming in liver cancer. *Cancer Cell* **36**, 418–430.e6 (2019).
- [16] Ma, L., Wang, L., Khatib, S. A. *et al.* Single-cell atlas of tumor cell evolution in response to therapy in hepatocellular carcinoma and intrahepatic cholangiocarcinoma. *J. Hepatol.* **75**, 1397–1408 (2021).
- [17] Alvarez, M., Benhammou, J. N., Darci-Maher, N. *et al.* Human liver single nucleus and single cell RNA sequencing identify a hepatocellular carcinoma-associated cell-type affecting survival. *Genome Med.* **14**, 50 (2022).
- [18] Long, R., Suoangbaji, Ng, I. O.-L. & Ho, D. W.-H. LiverSCA: a comprehensive and user-friendly cell atlas in human hepatocellular carcinoma. *Comput. Struct. Biotechnol. J.* **23**, 2740–2745 (2024).
- [19] Lotfollahi, M., Wolf, F. A. & Theis, F. J. scgen predicts single-cell perturbation responses. *Nat. Methods* **16**, 715–721 (2019).

- [20] Gao, R., Bai, S., Henderson, Y. C. *et al.* Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat. Biotechnol.* **39**, 599–608 (2021).
- [21] ICBI Lab. infercnvpy: scalable inference of copy-number variation from single-cell transcriptomics data. Software repository (2026). URL <https://github.com/icbi-lab/infercnvpy>.
- [22] Wolf, F. A., Hamey, F. K., Plass, M. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).
- [23] Qiu, X., Mao, Q., Tang, Y. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
- [24] Street, K., Risso, D., Fletcher, R. B. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
- [25] Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
- [26] Liu, X. *et al.* CSI-GEP: a GPU-based unsupervised machine learning approach for recovering gene expression programs in atlas-scale single-cell RNA-seq data. *Cell Genom.* **5**, 100739 (2025).
- [27] Cox, D. R. Regression models and life-tables. *J. R. Stat. Soc. B* **34**, 187–220 (1972).
- [28] Liu, H., Zhang, W., Zhang, Y. *et al.* Mime: a flexible machine-learning framework to construct and visualize models for clinical characteristics prediction and feature selection. *Comput. Struct. Biotechnol. J.* **23**, 2798–2810 (2024).
- [29] Pedregosa, F., Varoquaux, G., Gramfort, A. *et al.* Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
- [30] Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).
- [31] Zheng, C., Zheng, L., Yoo, J.-K. *et al.* Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. *Cell* **169**, 1342–1356.e16 (2017).
- [32] Indre, M.-G., Stefanini, B., Boe, M. *et al.* HCC is the predominant liver-related event in MASLD: two-step non-invasive algorithms to stratify risk in non-cirrhotic patients. *J. Hepatocell. Carcinoma* **13**, 561956 (2026).
- [33] Cancer Genome Atlas Research Network. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell* **169**, 1327–1341.e23 (2017).

- [34] Fujimoto, A., Furuta, M., Totoki, Y. *et al.* Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat. Genet.* **48**, 500–509 (2016).
- [35] Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
- [36] Hao, Y., Hao, S., Andersen-Nissen, E. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
- [37] Miao, Z., Moreno, P., Huang, N. *et al.* Putative cell type discovery from single-cell gene expression data. *Nat. Methods* **17**, 621–628 (2020).
- [38] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
- [39] Blanche, P., Dartigues, J.-F. & Jacqmin-Gadda, H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat. Med.* **32**, 5381–5397 (2013).
- [40] Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53**, 457–481 (1958).

Author contributions

Yan Liu and Yishen Li designed the study, performed the analyses and interpreted the data. Jianpeng Wei designed the study, supervised the project and wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (No. 82073877).

Data availability

The single-cell and single-nucleus RNA-sequencing datasets analysed in this study are publicly available through the NCBI Gene Expression Omnibus under accession numbers GSE125449, GSE151530, GSE189175 and GSE212046. The bulk RNA-sequencing and associated clinical data are available from the Genomic Data Commons under the TCGA-LIHC project and from the International Cancer Genome Consortium under the LIRI-JP project, subject to the access conditions of the respective repositories. No new primary sequencing data were generated. All data generated by the analyses and required to support the findings of this study are included in this article.

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval

Not applicable. This study utilised publicly available, anonymised datasets from the GEO, TCGA and ICGC databases and did not require additional ethics approval.

Consent to participate

Not applicable. No new participants were recruited and no identifiable participant data were used in this secondary analysis.

Consent for publication

All authors approved the final manuscript and its submission to this journal.