RESEARCH

Open Access

Integration of transcriptomic analysis and multiple machine learning approaches identifies NAFLD progression-specific hub genes to reveal distinct genomic patterns and actionable targets

Jing Sun[1,2†], Run Shi[3†], Yang Wu[4†], Yan Lou[1†], Lijuan Nie[1], Chun Zhang[2,5], Yutian Cao[1,2], Qianhua Yan[1], Lifang Ye[1], Shu Zhang[1], Xuanbin Wang[6], Qibiao Wu[7], Xuehua Jiao[8], Jiangyi Yu[1*], Zhuyuan Fang[2,9*] and Xiqiao Zhou[1*]

†Jing Sun, Run Shi, Yang Wu, Yan Lou contributed equally to this work.

*Correspondence:
Jiangyi Yu
yujiangyi2007@163.com
Zhuyuan Fang
jsszyyfzy@163.com
Xiqiao Zhou
zhouxiqiao@njucm.edu.cn

Full list of author information is available at the end of the article

Abstract

Background Nonalcoholic fatty liver disease (NAFLD) is a leading public health problem worldwide. Approximately one fourth of patients with nonalcoholic fatty liver (NAFL) progress to nonalcoholic steatohepatitis (NASH), an advanced stage of NAFLD. Hence, there is an urgent need to make a better understanding of NAFLD heterogeneity and facilitate personalized management of high-risk NAFLD patients who may benefit from more intensive surveillance and preventive intervene.

Methods In this study, a series of bioinformatic methods were performed to identify NAFLD progression-specific pathways and genes, and three machine learning approaches were combined to construct a risk-stratification gene signature to quantify risk assessment. In addition, bulk RNA-seq, single-cell RNA-seq (scRNA-seq) transcriptome profiling data and whole-exome sequencing (WES) data were comprehensively analyzed to reveal the genomic alterations and altered pathways between distinct molecular subtypes.

Results Two distinct subtypes of NAFL were identified with the NAFLD progression-specific genes, and one subtype has a high similarity of the inflammatory pattern and fibrotic potential with NASH. The established risk-stratification gene signature could discriminate advanced samples from overall NAFLD. COL1A2, one key gene closely related to NAFLD progression, is specifically expressed in fibroblasts involved in hepatocellular carcinoma (HCC), and significantly correlated with EMT and angiogenesis in pan-cancer. Moreover, the β-catenin/COL1A2 axis might play a critical role in fibrosis severity and inflammatory response during NAFLD-HCC progression.

Conclusion In summary, our study provided evidence for the necessity of molecular classification and established a risk-stratification gene signature to quantify risk assessment of NAFLD, aiming to identify different risk subsets and to guide personalized treatment.

Keywords Nonalcoholic fatty liver disease, Drug repositioning, Bioinformatic analysis, Machine learning, Hepatocellular carcinoma, Mutational signatures

Introduction

Nonalcoholic fatty liver disease (NAFLD), which ranges from simple steatosis to nonalcoholic steatohepatitis (NASH), is the most common cause of chronic liver disease worldwide and affects approximately one fourth of the global population [1, 2]. About 25% of patients with NAFL progress to NASH, an advanced stage of NAFLD [3]. NAFLD is characterized by hepatic steatosis, necroinflammation, and liver fibrosis, and these pathological features coordinate in the progression of NAFLD and jointly contribute to the development to end-stage liver disease such as cirrhosis and hepatocellular carcinoma (HCC) [4, 5].

A "multi-hit" theory has been proposed to explain NAFLD pathogenesis [6]. Hepatic steatosis is generally attributed to the dysfunction of intrahepatic lipid metabolism, which acts as the first hit and sensitizes the liver to the subsequent hits including oxidative stress, inflammation, and injury [7, 8]. A number of key molecules serve a significant role in NAFLD progression. For example, chemokine CCL2 and its receptor CCR2 are abnormally elevated during NAFLD progression, and the therapeutic strategy of targeting CCL2 and CCR2 has been shown as a promising approach for the treatment of NASH [9, 10]. Important as these molecules are, however, they are not enough to comprehensively explain the initiation and mechanism of NAFLD. NAFLD is a complex heterogeneous disease resulted from both intrinsic susceptibility and environmental background, during which numerous molecules such as CCL2 and CCR2 within the gene regulation network mediate NAFLD progression at different disease stages. Hence, there is an urgent need to identify more molecular drivers and coordinators involved in liver steatosis and inflammation, thus to make a better understanding of NAFLD heterogeneity and facilitate personalized management of high-risk NAFLD patients who may benefit from more intensive surveillance and preventive intervene.

In this study, we aimed to investigate the dysregulated signaling pathways and identify key genes involved in the progression of NAFLD. With a series of bioinformatic approaches, a robust NAFLD progression-specific gene signature was established to classify the NAFL patients into different risk subgroups. Intriguingly, more severe inflammatory response, alteration of extracellular matrix organization and cell-cell adhesion were detected in the high-risk NAFL samples compared to the low-risk subset, and similar infiltrating patterns of fibroblasts and Th1/2 cell populations were observed between high-risk NAFL and NASH samples. Furthermore, an integrative strategy of machine learning was used to screen for the most robust biomarkers, and their capacity of discrimination in progressive stages of NAFLD was validated in different independent cohorts. In addition, distinct immune and stromal patterns were observed among different CTNNB1/COL1A2 groups in HCC. From these comprehensive analyses, we provide evidence for the necessity of molecular classification and its potential clinical utility in the risk stratification of NAFLD, aiming to guide preventive intervene in the early stage of the disease. We hope this work could facilitate the personalized management and treatment of NAFLD patients.

Methods

Collection and preprocessing of transcriptome profiling data

Available transcriptome profiling data of healthy liver, NAFLD, and HCC samples were systematically searched in public databases. In summary, four microarray datasets named GSE167523 (51 NAFL and 47 NASH) [11], GSE163211 (88 simple steatosis samples, 72 NASH with F0, and 82 NASH with F1-4) [12], GSE135251 (10 normal, 51 NAFLs, and 155 NASHs with F0-4) [13], and GSE164760 (6 healthy liver tissues, 74 NASH, and 53 NASH-HCC) [14], and two bulk RNA-seq datasets named TCGA-LIHC [15] and the Genotype-Tissue Expression (GTEx) project (110 donated liver tissues, 50 adjacent normal tissues, and 369 HCC samples) [16] were included in our study. The detailed clinicopathological characteristics are summarized in supplementary Table 1. In addition, three single-cell RNA-seq (scRNA-seq) datasets of HCC samples were obtained from GSE125449, GSE146409, and GSE166635 [17–19]. All the raw CEL files and clinical information of microarray datasets were downloaded from the Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/). Probe IDs were mapped to gene symbols according to the corresponding annotation file, and the maximal measurement was selected as the final gene expression value if one gene symbol has multiple probes. All the microarray and RNA-seq data included in this study were normalized and log2 transformed as previously reported [20–22].

Identification of NAFLD progression-specific genes and pathways

GSE167523 was used as a training cohort. Using Gene Ontology Biological Process (GOBP) enrichment analysis, the transcriptome profiling data of GSE167523 which contains 51 NAFL and 47 NASH samples were used to explore the enriched pathways which represent the alterations of biological processes involved in NAFLD progression.

Two strategies were combined to screen for NAFLD progression-specific genes. Firstly, the weighted gene co-expression network analysis (WGCNA) [23] was used to construct a scale-free co-expression network using R package "WGCNA" and to identify a gene module which is mostly correlated with sample category (NAFL or NASH), and this gene module was considered as "gene module of NAFLD progression". Secondly, using R package "limma", differentially expressed genes (DEGs) were identified between NAFL and NASH samples with a filtering threshold of FDR (false discovery rate) q value less than 0.01 and fold change $(FC)>2$ or <0.5. The overlapping genes in "gene module of NAFLD progression" and "DEGs between NAFL and NASH" were finally considered as NAFLD progression-specific genes. A protein-protein interaction (PPI) network was generated to reveal the functional and physical linkages among these DEGs using the STRING database.

Connectivity map (CMap) analysis

CMap is a resource that uses transcriptional expression data to probe the relationships between disease, cell physiology, and therapeutics [24]. The above-mentioned NAFLD progression-specific genes were submitted to CMap for analysis to explore potential targets and applicable drugs for high-risk NAFLD patients. Top 10 compounds with the highest predictive scores and corresponding descriptions of mode-of-actions (MoAs) were displayed in a dot diagram.

Estimation of immune infiltration and tumor purity

Several computational algorithms including TIMER [25], Cibersort [26], quanTIseq [27], MCP-counter [28], xCell [29], and EPIC [30] were used to quantify the infiltrating abundance of different cell types, and the ESTIMATE algorithm [31] was applied to infer the tumor purity and immune score. In detail, the infiltration abundance of cancer-associated fibroblast (CAF) was compared among different clusters using EPIC, MCP-counter and xCell algorithms, respectively.

Establishment of a risk-stratification gene signature

To establish a risk-stratification gene signature to discriminate high-risk NAFLD, we combined multiple machine learning approaches including random forest (RF), support vector machine (SVM), and logistic regression (LR) with Lasso regularization to screen for optimal candidates. Recursive feature elimination (RFE) was applied in the RF and SVM algorithms to remove the weakest features and retain the optimal features. To ensure the high accuracy and stability of the predictive model, we introduced the leave-one-out cross-validation (LOOCV) framework in the training cohort. Subsequently, Lasso regularization adds a penalty parameter (λ) to the LR model, and this action can lead to zero coefficients, i.e. some of the candidate genes are completely neglected for evaluation. In our analysis, 8 candidates remained after the integrative selection of RF-RFE and SVM-RFE, and 4 genes retained their logistic coefficients after Lasso regularization. Finally, the discriminative score for risk stratification was calculated with the relative expression and Lasso logistic coefficient of the four genes as follows:

$$\mathrm{Discriminativescore} = \sum_{i} \mathrm{Coefficient}\left(\mathrm{mRNA_i}\right) \times \mathrm{Expression}\left(\mathrm{mRNA_i}\right)$$

scRNA-seq analysis

Three scRNA-seq datasets (GSE125449, GSE146409, and GSE166635) were used to reveal the components of HCC tumor microenvironment (TME) and depict the gene expression characteristics in different cell types. GSE125449 contains single-cell transcriptome profiling data of nine HCC specimens, GSE146409 six specimens, and GSE166635 two specimens. The scRNA-seq expression matrix was processed with R package "Seurat". At first, the "NormalizeData" function was applied to normalize the gene expression data, then the "FindVariableFeatures" function was applied to identify the top 2,000 highly variable genes (HVGs). After performing "RunPCA" for dimension reduction, R package "Harmony" was used to eliminate batch effect. "FindNeighbors" was used to determine the k-nearest neighbors of each cell, and "FindClusters" was used to determine optimal clusters. UMAP reduction was used for cluster visualization, and "SingleR" package was used for cluster annotation. In addition, "FeaturePlot" and "VlnPlot" functions were used to visualize the gene expression of COL1A2 in each cell type.

Mutation analysis

Somatic variant data of TCGA-LIHC, which was called using MuTect2, were sorted in a mutation annotation format (MAF) file, and were visualized using R package "maftools" [32]. The mutation variants of "Frame_Shift_Del", "Frame_Shift_Ins", "In_Frame_Del", "In_Frame_Ins", "Missense_Mutation", "Nonsense_Mutation", "Nonstop_Mutation", "Splice_Region", "Splice_Site" and "Translation_Start_Site" were defined as non-synonymous, and the other variants were defined as synonymous. Furthermore, R package "sigminer" was used to extract mutational signatures from the whole-exome sequencing (WES) data [33]. Bayesian variant of nonnegative matrix factorization (NMF) algorithm was used to decipher mutational signatures, and the optimal factorization of k value is selected when the magnitude of the cophenetic correlation coefficient begins to drop significantly. Mutational signatures were annotated by computing cosine similarity against validated single base substitution (SBS) mutational catalogues retrieved from the COSMIC database as previously reported [34].

In addition, the mutation frequency and mutual exclusivity of TP53 and CTNNB1 in HCC were further shown in the TCGA-HCC cohort and another two cohorts (MSK-HCC [35] and INSERM-HCC [36]) with oncoplots.

Pan-cancer analysis

The transcriptome data and annotation files of 32 malignant solid cancers (ACC, BLCA, BRCA, CESC, CHOL, COAD, DLBC, ESCA, GBM, HNSC, KICH, KIRC, KIRP, LGG, LIHC, LUAD, LUSC, MESO, OV, PAAD, PCPG, PRAD, READ, SARC, SKCM, STAD, TGCT, THCA, THYM, UCEC, UCS, UVM) were obtained from the Genomic Data Commons portal (https://gdc.cancer.gov/). The activities of ten cancer hallmarks including EMT, angiogenesis, apoptosis, inflammation, hypoxia, glycolysis, cell cycle progression (CCP), senescence, DNA repair and oxidative phosphorylation were quantified using the single-sample gene set enrichment analysis (ssGSEA) with the corresponding gene sets retrieved from the Molecular Signatures Database (MSigDB).

Additional bioinformatic and statistical analyses

The heatmaps of gene expression and cell abundance were generated using R package "pheatmap". R packages "GOSemSim" and "ggtree" were used to measure and visualize the semantic similarity among GO terms [37, 38]. Different risk subgroups of NAFL were identified with the expression matrix of NAFLD progression-specific genes using NMF, and the optimal factorization of k value was selected when the magnitude of the cophenetic correlation coefficient begins to fall significantly. Principal component analysis (PCA) was used to visualize the dissimilarity between NAFL and NASH samples. The activities of some biological features such as "inflammatory response", "ECM organization" and "cell-cell adhesion" in NAFLD samples were quantified using the ssGSEA method based on the transcriptome profiling data and corresponding gene sets retrieved from the MSigDB [39]. Pearson correlation analysis was performed to evaluate the correlation between two variables subject to normal distribution. Student's t-test or one-way analysis of variance (ANOVA) was used to analyze differences among groups with variables subject to normal distribution, otherwise Mann-Whitney U test or Kruskal-Wallis test. The receiver operating characteristic (ROC) analysis was used to evaluate the predictive accuracy of the discriminative score established in our study. Categorical variables between the two groups were compared using the chi-square test. Two-sided p value or FDR q value less than 0.05 was considered statistically significant. All analyses were performed in the GraphPad Prism 8 and R 4.1.0 software.

Results

Identification of NAFLD progression-specific pathways and genes

Using R package “limma”, significantly upregulated genes were identified in NASH compared to NAFL, and then submitted to GO analysis for pathway exploration. R package “GOSemSim” was used to measure the similarity among all the biological processes (BP), and R package “ggtree” was used to visualize the clusters, and similar BPs were clustered into a mutual branch (Fig. 1a). According to the significance, we extracted the top five altered GOBPs and displayed them in a Circos diagram (Fig. 1b). We observed that alterations in two biological features are mainly responsible for the NAFLD progression: extracellular matrix (ECM) organization and cell cycle process.

Two filtering strategies were combined to identify NAFLD progression-specific genes. Firstly, WGCNA was performed with the transcriptome profiling data of 98 NAFLD samples and their sample category (NAFL or NASH) to construct a scale-free co-expression network. A total of 36 gene modules were generated with a power of 6 as the optimal soft threshold (Supplementary Fig. 1). Among these modules, the brown module exhibited a highest correlation with sample category ($|r| = 0.61$, p=6e-11) and was considered as “gene module of NAFLD progression” (Fig. 1c). Secondly, the “limma” algorithm was used to identify a total of 378 DEGs between NAFL and NASH samples with a filtering threshold of q value less than 0.01 and fold change (FC) > 2 or < 0.5 (Fig. 1d). Finally, 182 overlapping genes in the intersection of “gene module of NAFLD progression” and “DEGs between NAFL and NASH” were considered as “NAFLD progression-specific genes” (Fig. 1e) and the detailed information is shown in supplementary Table 2. A PPI network was generated to depict the functional and physical linkages among these DEGs, and we observed that the hub (the red circle) in the network is mainly composed of collagen family members (Fig. 1f).

To explore potential targets and applicable drugs for high-risk NAFLD patients, the above-mentioned 182 NAFLD progression-specific genes were submitted to CMap for further investigation. Top 10 compounds with the highest predictive scores and corresponding 7 mode-of-actions (MoAs) were displayed in a dot diagram (Fig. 2g). The 7 MoAs were annotated with “Angiotensin receptor antagonist”, “Aromatic hydrocarbon derivative”, “Aurora kinase inhibitor”, “EGFR inhibitor”, “HDAC inhibitor”, “NFκB pathway inhibitor”, and “Sphingosine kinase inhibitor”. In particular, four compounds named HDAC3-selective, entinostat, mocetinostat, and Merck60 share a common MoA of HDAC inhibitor (HDACi), which indicates HDACi might be a potentially applicable drug for advanced NAFLD patients.

Different risk subgroups with distinct inflammatory and fibrotic patterns were identified in NAFL

The identified 182 genes concerning NAFLD progression were further analyzed using GO method, and a Circos illustrated that they were mainly enriched in five GOBPs annotated with “ECM organization”, “Blood vessel development”, “Wnt signaling pathway”, “Cell adhesion”, and “Cell morphogenesis” (Fig. 2a). We observed that the five GOBPs include the main biological alterations occurred in NASH, and this result further indicated that the 182 genes could commendably represent the pathological development of NAFLD. Based on the expression profile of 182 NAFLD progression genes and using the NMF algorithm, the 51 NAFLs in the training cohort were divided into

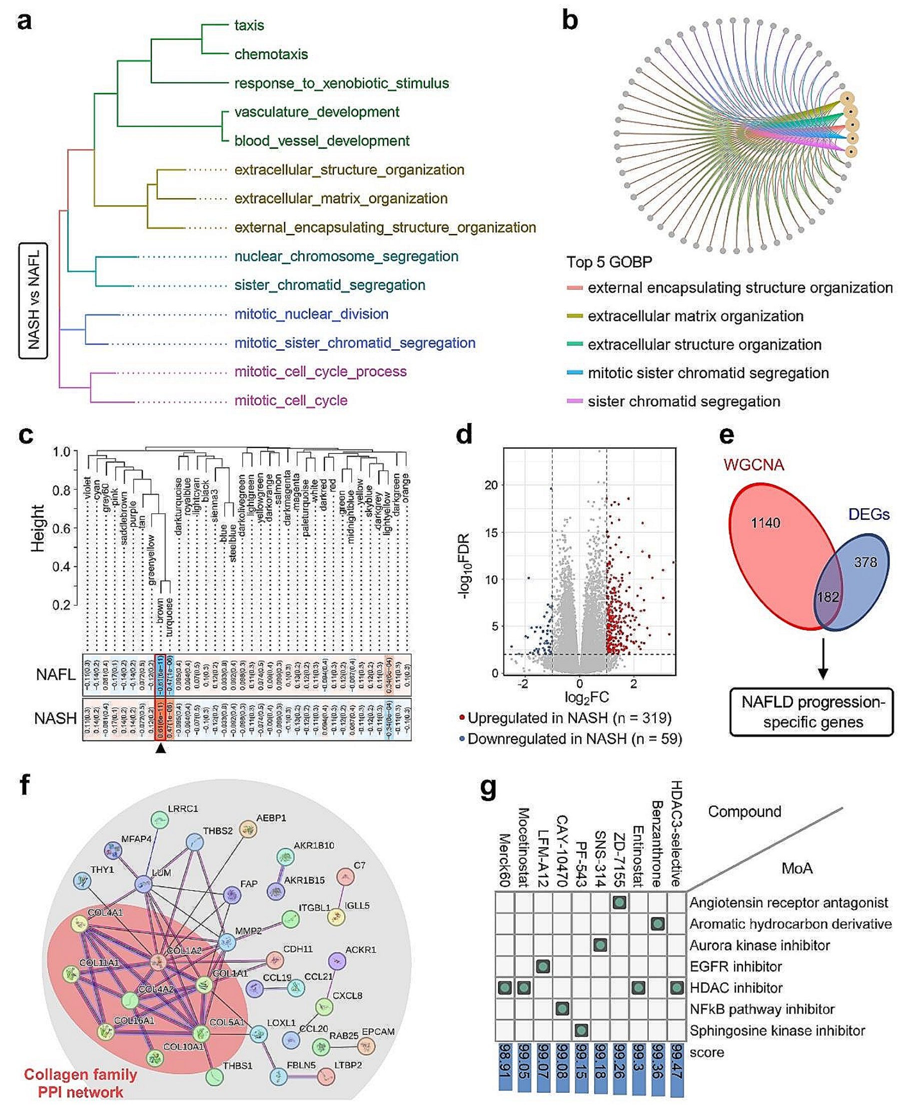

Fig. 1 Identification of NAFLD progression-specific pathways and genes. **(a)** The similarities among all the biological processes (BPs) were measured and visualized, and similar terms were clustered into a common branch. **(b)** The top five altered GOBPs were displayed in a Circos diagram. **(c)** WGCNA was performed with the transcriptome profiling data of 98 NAFLD samples and a total of 36 gene modules were identified. The brown module which exhibited the highest correlation with sample category ($|r| = 0.61$, $p = 6e-11$) and was considered as "WGCNA-identified gene module of NAFLD progression". **(d)** A volcano plot showed a total of 378 DEGs between NAFL and NASH samples with a filtering threshold of q value less than 0.01 and fold change (FC) > 2 or < 0.5. **(e)** 182 overlapping genes in the intersection of "WGCNA-identified gene module of NAFLD progression" and "DEGs between NAFL and NASH" are considered as "NAFLD progression-specific genes". **(f)** A PPI network was generated to depict the functional and physical linkages among these DEGs, and the hub (the red circle) in the network is mainly composed of collagen family members. **(g)** CMap algorithm showed top 10 compounds with the highest predictive scores and corresponding 7 mode-of-actions (MoAs) in a dot diagram. The 7 MoAs were annotated with "Angiotensin receptor antagonist", "Aromatic hydrocarbon derivative", "Aurora kinase inhibitor", "EGFR inhibitor", "HDAC inhibitor", "NFκB pathway inhibitor", and "Sphingosine kinase inhibitor"

two subclusters (C1=21, C2=30; Fig. 2b). Using ssGSEA algorithm, we found that the quantified scores of "Inflammatory response", "ECM organization", and "Cell-cell adhesion" were significantly and progressively elevated from C2 to C1 to NASH (Fig. 2c - e; *** $p<0.001$). In particular, positive correlations between "Inflammatory response" and "ECM organization" or "Cell-cell adhesion" were observed in NAFL-C1, C2 and NASH

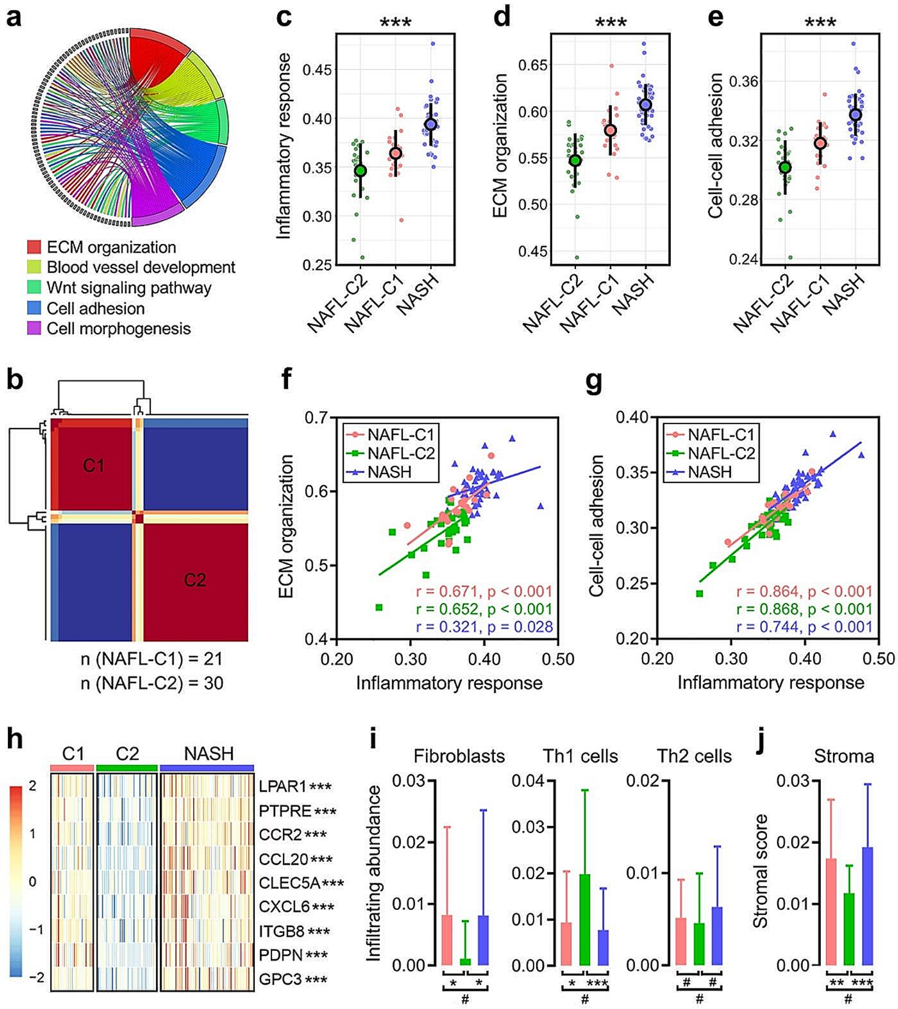

Fig. 2 Different risk subgroups with distinct inflammatory and fibrotic patterns were identified in NAFL. **(a)** The 182 NAFLD progression-specific genes were analyzed using GO method, and the Circos illustrated that they were mainly enriched in five GOBPs annotated with "ECM organization", "Blood vessel development", "Wnt signaling pathway", "Cell adhesion", and "Cell morphogenesis". **(b)** Based on the expression profile of the 182 NAFLD progression genes and using the NMF algorithm, the 51 NAFLs in the training cohort were divided into two subclusters (C1 = 21, C2 = 30). **(c-e)** Using ssGSEA quantification, we observed that the ssGSEA scores of "Inflammatory response", "ECM organization", and "Cell-cell adhesion" were progressively elevated from C2 to C1 to NASH. **(f & g)** Positive correlations between "Inflammatory response" and "ECM organization" or "Cell-cell adhesion" were observed in NAFL-C1, C2, and NASH samples. **(h)** A group of widely acknowledged inflammatory factors including LPAR1, PTPRE, CCR2, CCL20, CLEC5A, CXCL6, ITGB8, PDPN, and GPC3 were significantly decreased in the NAFL-C2 group compared to either NAFL-C1 or NASH. **(i)** The fibroblasts abundance was significantly downregulated in NAFL-C2. Th1 cell infiltration was significantly upregulated in NAFL-C2, while no significant difference of Th2 cell infiltration was observed among the three groups. **(j)** The NAFL-C2 group exhibited the lowest stroma score, while no significant difference of stroma score was observed between NAFL-C1 and NASH. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, # not significant

samples (Fig. 2f & g). Furthermore, among all correlations, NASH exhibited the weakest correlation between "Inflammatory response" and "ECM organization" (Fig. 2f).

Considering inflammation as a driver in NAFLD progression, we selected a list of widely acknowledged inflammatory factors (LPAR1, PTPRE, CCR2, CCL20, CLEC5A, CXCL6, ITGB8, PDPN, and GPC3) and showed their expression profiles in NAFL-C1, C2 and NASH (Fig. 2h; *** $p < 0.001$). Using one-way ANOVA test, we found that all these inflammatory factors were significantly decreased in the NAFL-C2 group compared to either NAFL-C1 or NASH. In addition, we evaluated the abundance of fibrosis-related cell populations including fibroblast and Th1/2 cells in the three groups using xCell algorithm. The fibroblasts abundance was significantly downregulated in NAFL-C2. Th1 cell infiltration was significantly upregulated in NAFL-C2, while no significant difference of Th2 cell infiltration was observed among the three groups (Fig. 2i; * $p<0.05$, *** $p<0.001$, # not significant). Furthermore, the NAFL-C2 group exhibited the lowest stroma score, while no significant difference of stroma score was observed between NAFL-C1 and NASH (Fig. 2j; ** $p<0.01$, *** $p<0.001$, # not significant).

Establishment and validation of the discriminative score for risk stratification

To assess the risk stratification of NAFLD in a quantitative method, we combined different machine learning (ML) approaches to screen for robust biomarkers. The training method of the LOOCV framework was introduced in Fig. 3a, and the iteration number in our study is $N=98$ (the sample number in the training cohort). The identified 182 genes regarding NAFLD progression were trained in the RF and SVM algorithms with feature selection of recursive feature elimination (RFE) respectively, and 8 overlapping genes (COL1A1, COL1A2, COL4A1, COL4A2, COL5A1, DTNA, THBS1, and UBD) remained in the outputs of the two ML algorithms. Finally, Lasso logistic regression (LR) analysis was applied on the 8 genes, and only 4 genes (DTNA, COL4A2, UBD, and COL1A2) retained their coefficients. Among them, COL1A2 exhibited the highest coefficient (Fig. 3b). PCA analysis showed a clear separation of NAFL and NASH samples with the expression matrix of the four genes (Fig. 3c). Furthermore, dotplots showed that all four genes were significantly and stepwisely elevated from NAFL-C2 to C1 to NASH samples (Fig. 3d). In the training cohort, the discriminative score was calculated for each sample according to the formula stated in the methods section, and ROC analysis indicated that the score could discriminate NASH from NAFL accurately (Fig. 3e).

A dataset named GSE163211, which contains 88 steatosis samples, 72 NASH with F0, and 82 NASH with F1-4, was used as an external testing cohort. The GSE163211 dataset was produced from the GPL29503 platform of NanoString Technologies and contains RNA-seq data of 800 custom genes. Among the four genes (DTNA, COL4A2, UBD, and COL1A2), only COL1A2 and COL4A2 were detected in the platform. COL1A2 and COL4A2 were significantly upregulated in NASH with fibrosis, and the 2-gene discriminative score was also significantly higher than steatosis and NASH samples without fibrosis (Fig. 3f). In the total of 154 NASH samples, the ROC analysis demonstrated that the 2-gene discriminative score could identify advanced fibrotic samples with a favorable performance (AUC=0.724, 95% CI=0.644–0.804; Fig. 3g).

In addition, another dataset named GSE135251 containing 10 normal tissues, 51 NAFLs, and 155 NASHs with F0-4 was used as the second external testing cohort. COL1A2, COL4A2 and DTNA were detected in the platform, and we calculated the "3-gene score" for each sample, and compared the score among normal samples, NAFL, and NASH with different fibrosis levels (F0-4). We observed that the "3-gene score" was significantly and stepwisely elevated from normal to NAFL to NASH with advanced fibrosis levels ($p<0.001$; Fig. 3h). Furthermore, we evaluated its discriminative capacity for NAFLD and fibrosis levels using ROC analysis, and the "3-gene score" exhibited favorable performances in discriminating advanced stages (AUC=0.737, 95% CI=0.665–0.810; Fig. 3i) and advanced fibrosis levels (AUC=0.729, 95% CI=0.650–0.808; Fig. 3j).

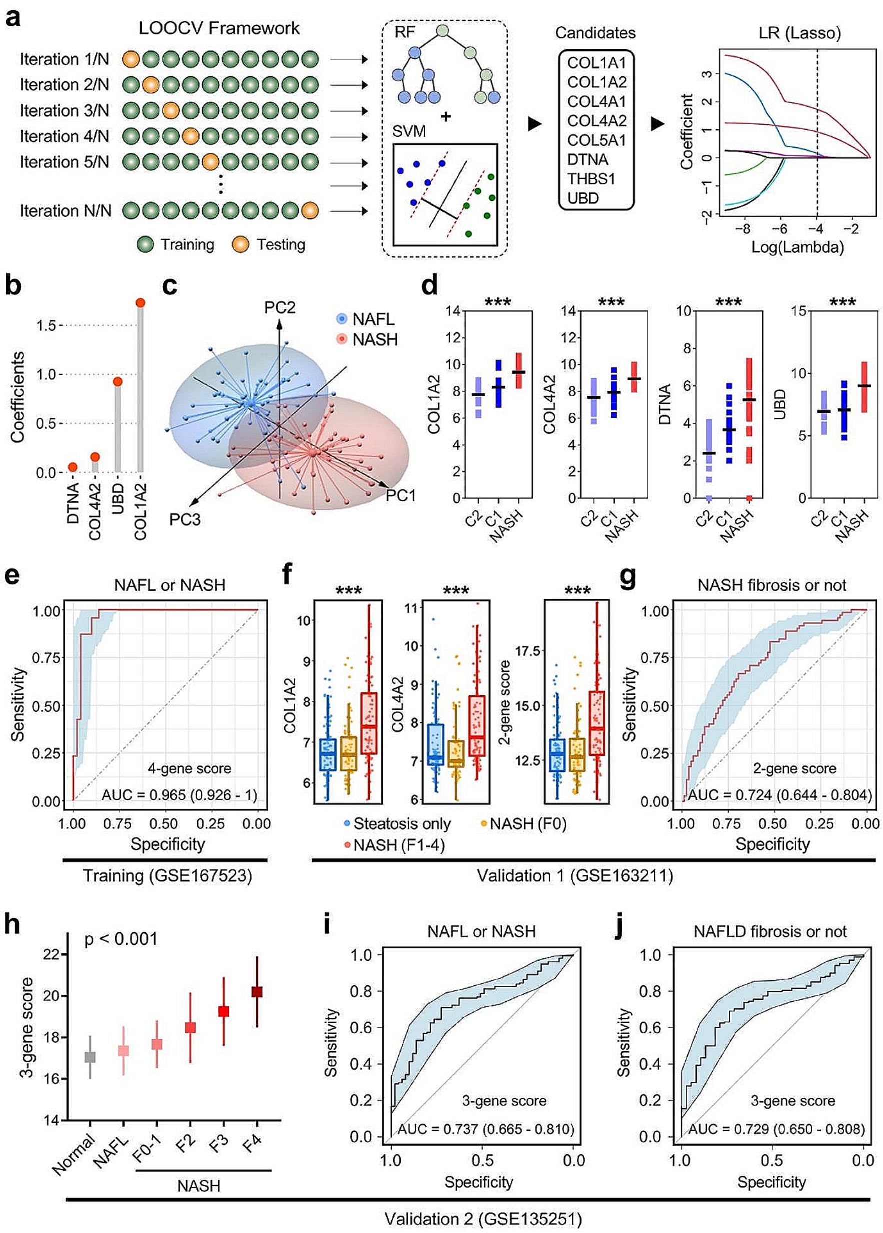

Fig. 3 Establishment and validation of a discriminative gene signature for risk stratification in NAFLD. **(a)** The training method of the LOOCV framework was introduced, and the iteration number in our study is 98 (the sample number in the training cohort). The 182 genes regarding NAFLD progression were trained in the RF and SVM algorithms with feature selection of recursive feature elimination (RFE) respectively, and 8 overlapping genes remained in the outputs of the two machine learning approaches. Subsequently, Lasso logistic regression (LR) analysis was applied on the 8 genes, and only 4 genes retained their coefficients, **(b)** and COL1A2 exhibited the highest coefficient. **(c)** PCA analysis showed a clear separation of NAFL and NASH samples with the expression matrix of the four genes. **(d)** Dotplots showed that all four genes were significantly upregulated in NASH compared to NAFL samples. *** $p<0.001$. **(e)** In the training cohort, the ROC analysis indicated that the gene signature-derived score could discriminate NASH from NAFL accurately. **(f)** In GSE163211, COL1A2 and COL4A2 were significantly upregulated in NASH with fibrosis, and the 2-gene discriminative score was also significantly higher than steatosis and NASH samples without fibrosis. **(g)** The ROC analysis demonstrated that the 2-gene (COL1A2 and COL4A2) discriminative score could identify advanced samples with a favorable performance (AUC = 0.724, 95% CI = 0.644–0.804). **(h)** In GSE135251, the "3-gene score" was significantly and stepwisely elevated from normal to NAFL to NASH with advanced fibrosis levels, and it exhibited favorable performances in discriminating **(i)** advanced stages (AUC = 0.737, 95% CI = 0.665–0.810) and **(j)** advanced fibrosis levels (AUC = 0.729, 95% CI = 0.650–0.808)

The risk-stratification gene signature was significantly correlated with malignant progression

To evaluate the biological role of the four discriminative genes in NAFLD and malignant progression, we investigated the expression profile of the four genes in different datasets which consist of normal liver tissues, NASH, and HCC samples. In the GSE164760 microarray dataset, all the four genes were significantly upregulated in HCC samples derived from NASH when compared to healthy liver tissues and NASH samples (Fig. 4a; *** $p<0.001$). In the combination of GTEx and TCGA-HCC RNA-seq database, all the

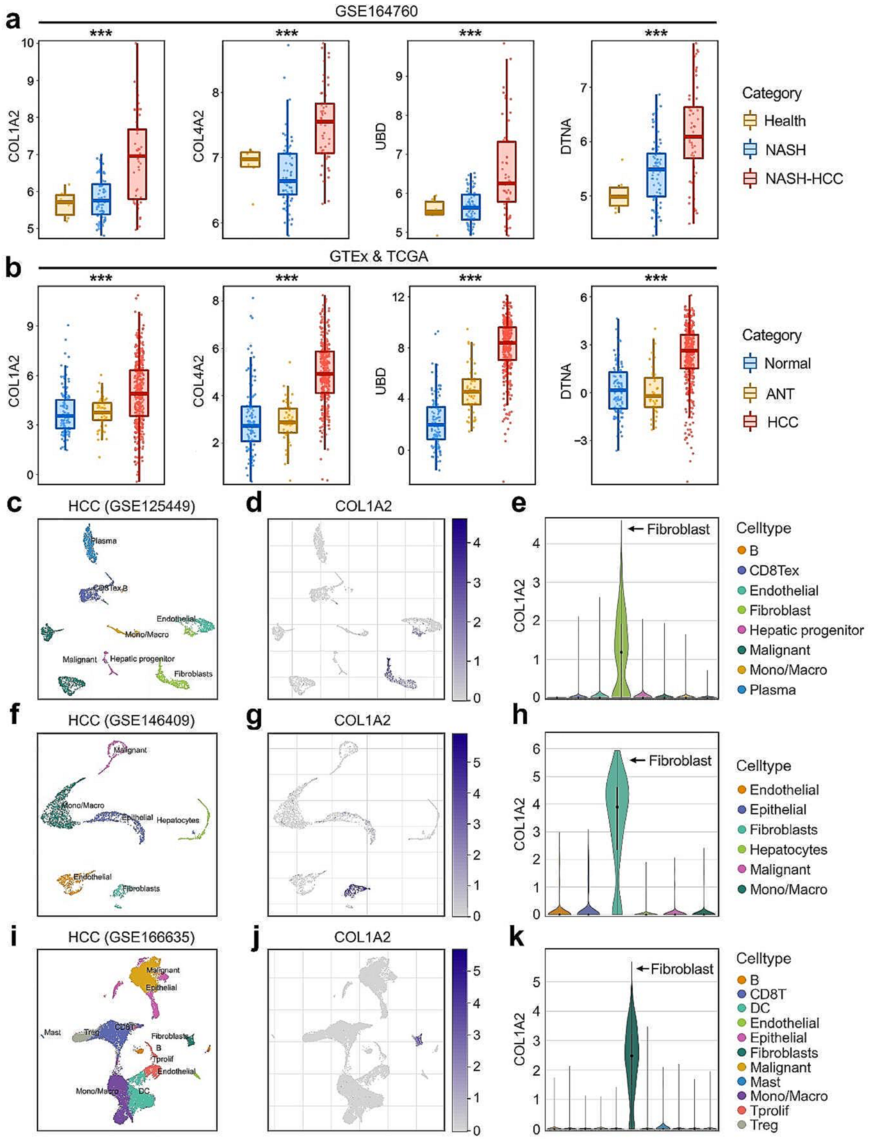

Fig. 4 The risk-stratification gene signature was significantly correlated with malignant progression. **(a)** In the GSE164760 microarray dataset, all the four genes (COL1A2, COL4A2, UBD, and DTNA) were significantly upregulated in HCC samples derived from NASH when compared to healthy liver tissues and NASH samples. **(b)** In the combination of GTEx and TCGA-HCC RNA-seq database, all the four genes were also significantly upregulated in HCC samples compared to donated normal tissues and adjacent normal tissues. **(c, f and i)** Using UMAP dimensionality reduction, three scRNA-seq datasets including GSE125449, GSE146409 and GSE166635 were used to reveal the components of HCC tumor microenvironment (TME) and **(d, g and j)** the expression profile of COL1A2 in different cell types, respectively. **(e, h and k)** Violin plots showed that COL1A2 is specifically expressed in fibroblasts, almost not expressed in other cells within HCC TME. *** $p<0.001$

four genes were also significantly upregulated in HCC samples compared to donated normal tissues and adjacent normal tissues (Fig. 4b; *** $p<0.001$).

In addition, three scRNA-seq datasets including GSE125449, GSE146409 and GSE166635 were used to reveal the components of HCC tumor microenvironment (TME) and the expression characteristics of hub genes. UMAP dimensionality reduction was used to show the distribution and dissimilarity of the cell types involved in HCC TME (Fig. 4c, f, and i). Considering COL1A2 possesses the highest discriminative coefficient among the four genes, we assessed the expression characteristics of COL1A2 (Fig. 4d, g, and j) in different cell types, and all the three violin plots showed that COL1A2 is specifically expressed in fibroblasts, almost not expressed in other cell types (Fig. 4e, h, and k). These findings demonstrated that the risk-stratification gene signature was closely relevant to the progression of NAFLD and HCC, and COL1A2 might play a specific role in fibroblast activation and fibrosis severity.

Analysis of mutational patterns and CTNNB1/COL1A2 axis in NAFLD and HCC

Using the WES data of TCGA-HCC and NMF algorithm, we attempted to decipher the mutational patterns of HCC and explore the relationship between COL1A2 and mutation patterns in HCC. The optimal k factorization of 5 was selected (Supplementary Fig. 2), and five mutational signatures were identified and matched in the COSMIC database (Fig. 5a). The three major mutational signatures were annotated with "Defective DNA mismatch repair", "Aflatoxin exposure", and "Aristolochic acid exposure", and the abundance of each mutational signature in TCGA-HCC was shown in a pie chart (Fig. 5b). We reviewed the medical records of the TCGA-HCC cohort and extracted 184 HCC samples with history of alcohol consumption, hepatitis or NAFLD, and the distribution of each mutational signature was depicted in a stacked barplot (Fig. 5c). Using chi-square test, we found that NAFLD-HCC is characterized with high COL1A2 expression (Fig. 5d). With quantiles of COL1A2 mRNA expression, 46 HCC samples were assigned to the COL1A2-lowest and -highest group, respectively. Oncoplots of the two groups demonstrated that CTNNB1 acts as the most frequently mutated gene in the COL1A2-low cohort, with the mutation frequency up to 48% (left panel, Fig. 5e). In contrast, CTNNB1 is rarely mutated in the COL1A2-high cohort (right panel, Fig. 5e). In the integrated analyses of TCGA-HCC, MSK-HCC and INSERM-HCC cohorts, the gene-pair of TP53 and CTNNB1 is shown significantly mutually exclusive (Fig. 5f). Furthermore, COL1A2 mRNA expression is significantly elevated in CTNNB1-wild type HCC samples compared to CTNNB1-mutated ones (Fig. 5g). Among nine representative oncogenic pathways, the WNT signaling pathway is the most frequently affected one in the COL1A2-low cohort (Fig. 5h). GOBP analysis was used to further explore the altered pathways in CTNNB1-WT/COL1A2-high samples. A functional network showed the five most important pathways in CTNNB1-WT/COL1A2-high HCC samples were termed "vasculature development", "chemotaxis", "ECM organization", "positive regulation of locomotion", and "positive regulation of cell adhesion" (Fig. 5i). These evidences demonstrated that the CTNNB1/COL1A2 axis might play a role in the fibrosis and inflammation severity during NAFLD-HCC progression.

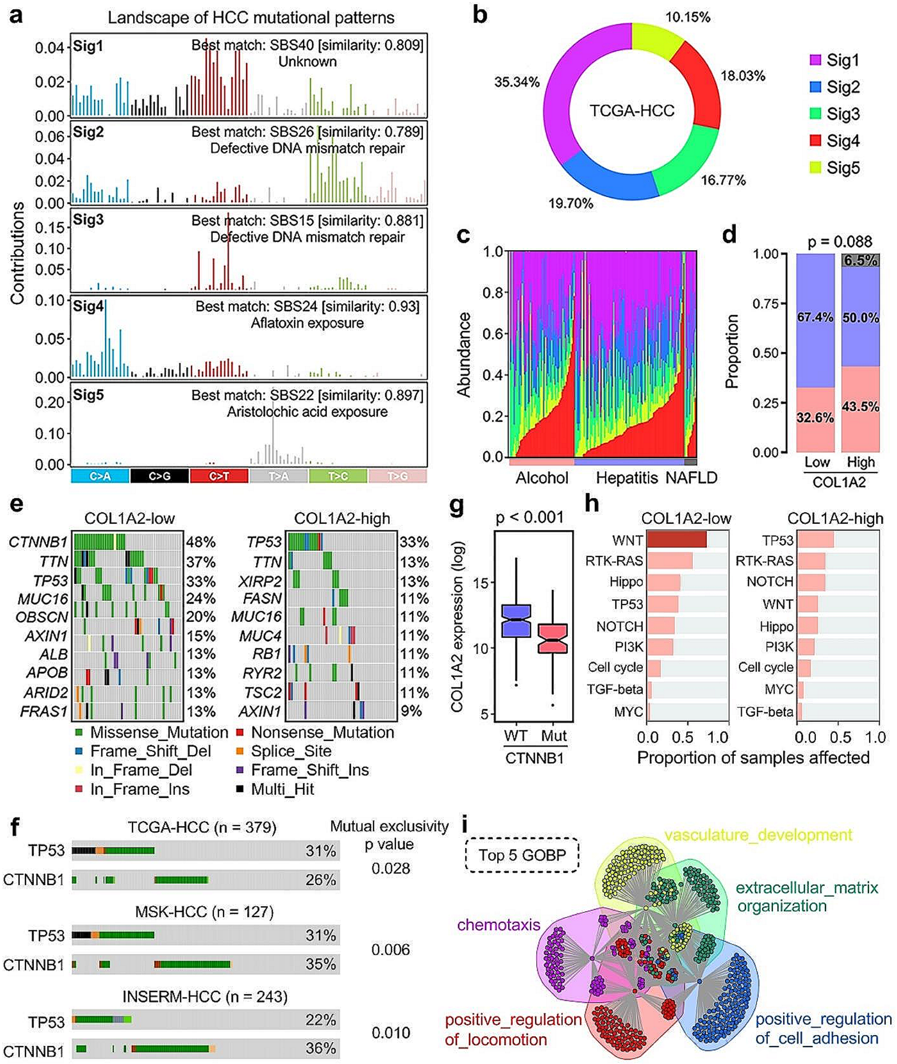

Fig. 5 CTNNB1/COL1A2 axis correlates with fibrosis severity during NAFLD-HCC progression. **(a)** Five mutational signatures were identified using TCGA-HCC WES data and NMF algorithm. The three major mutational signatures were annotated with "Defective DNA mismatch repair", "Aflatoxin exposure", and "Aristolochic acid exposure". **(b)** The abundance of each mutational signature in TCGA-HCC was shown in a pie chart. **(c)** 184 HCC samples with history of alcohol consumption, hepatitis or NAFLD were extracted, and the distribution of each mutational signature was depicted in a stacked barplot. **(d)** NAFLD-HCC is characterized with high COL1A2 expression. **(e)** Oncoplot demonstrated that CTNNB1 acts as the most frequently mutated gene in the COL1A2-low cohort, with the mutation frequency up to 48% (left panel). In contrast, CTNNB1 is observed rarely mutated in the COL1A2-high samples (right panel). **(f)** In the integrated analyses of TCGA-HCC, MSK-HCC and INSERM-HCC cohorts, the gene-pair of TP53 and CTNNB1 is shown significantly mutually exclusive. **(g)** COL1A2 mRNA expression is significantly elevated in CTNNB1-wild type HCC samples compared to CTNNB1-mutated ones. **(h)** Among nine representative oncogenic pathways, the WNT signaling pathway is the most frequently affected one in the COL1A2-low cohort. **(i)** A functional network showed the top five important pathways in CTNNB1-WT/COL1A2-high HCC samples were termed "vasculature development", "chemotaxis", "ECM organization", "positive regulation of locomotion", and "positive regulation of cell adhesion"

Distinct immune and stromal patterns were observed among different CTNNB1/COL1A2 groups

The infiltrating immune and stromal cells were estimated by TIMER, Cibersort, quanTIseq and MCP-counter, and the differences of distribution were also investigated in distinct CTNNB1/COL1A2 groups. As shown in a comprehensive heatmap, most of the immune and stromal cells are enriched in the CTNNB1-WT/COL1A2-high samples, regardless of the TMB level (Fig. 6a). Furthermore, the xCell algorithm inferred the absolute infiltration of a total of 36 cell types involved in the TME, and the chord diagram showed that the infiltrating abundance is significantly higher in the CTNNB1-WT/COL1A2-high and CTNNB1-Mut/COL1A2-high samples than the others (Fig. 2b). Infiltrating scores of CAFs were inferred using EPIC, MCP-counter, and xCell, and we observed that CAFs are significantly enriched in the CTNNB1-WT/COL1A2-high samples, suggesting the distinct fibrosis severity resulted from different classification of CTNNB1-WT and high COL1A2 expression (Fig. 6c - e). On the other hand, we

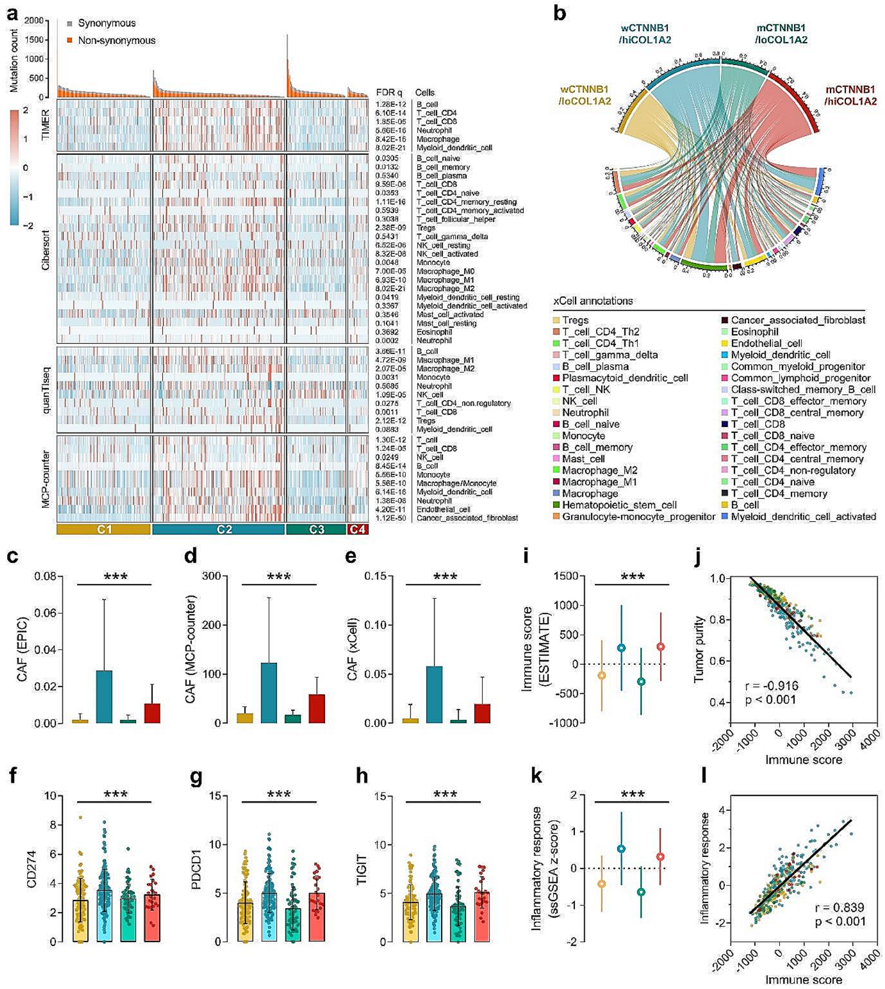

Fig. 6 Distinct immune and stromal patterns were observed among different CTNNB1/COL1A2 groups. **(a)** A comprehensive heatmap showed that most of the immune and stromal cells are enriched in the CTNNB1-WT/COL1A2-high samples, regardless of the TMB level. **(b)** The xCell algorithm inferred the absolute infiltration of a total of 36 cell types involved in the TME, and the chord diagram showed that the infiltrating abundance is significantly higher in the CTNNB1-WT/COL1A2-high and CTNNB1-Mut/COL1A2-high samples than the other groups. **(c - e)** CAFs are significantly enriched in the CTNNB1-WT/COL1A2-high samples. **(f - h)** Different levels of representative immune checkpoints including CD274, PDCD1 and TIGIT indicate heterogenous tumor immunogenicity and distinct potential response to immunotherapy among different groups. **(i)** The ESTIMATE algorithm was applied to infer the immune infiltration and tumor purity for each sample, and the CTNNB1-WT/COL1A2-high and CTNNB1-Mut/COL1A2-high groups are labelled with high immune infiltration, and **(j)** a significantly negative correlation (r = -0.916, $p < 0.001$) was observed between the immune score and tumor purity across the four groups. Similarly, **(k)** the inflammatory response activity exhibits distinct distributions among the four groups, and **(l)** a significantly positive correlation ($r = 0.839$, $p < 0.001$) between immune score and inflammatory response was also observed across the four groups. *** $p < 0.001$

evaluated the levels of representative immune checkpoints including CD274, PDCD1 and TIGIT, and significant differences were observed among different groups, indicating heterogenous tumor immunogenicity and distinct potential response to immunotherapy (Fig. 6f - h). Moreover, the ESTIMATE algorithm was applied to infer the immune infiltration and tumor purity of each sample, and we observed that the CTNNB1-WT/COL1A2-high and CTNNB1-Mut/COL1A2-high groups are labelled with high immune infiltration (Fig. 6i), and a significantly negative correlation (r = -0.916, p < 0.001) between the immune score and tumor purity across the four groups was shown in Fig. 6j. Similarly, the inflammatory response activity exhibits distinct distributions among the four groups (Fig. 6k), and a significantly positive correlation (r=0.839, p < 0.001; Fig. 6l) with immune score was also observed across the four groups.

COL1A2 significantly correlates with EMT and angiogenesis in pan-cancer

To investigate the connection between the hub gene COL1A2 and malignant features in pan-cancer, we quantified the ability of ten cancer hallmarks including EMT, angiogenesis, apoptosis, inflammation, hypoxia, glycolysis, cell cycle progression (CCP), senescence, DNA repair and oxidative phosphorylation using ssGSEA algorithm, and calculated the correlations between COL1A2 expression and the ten hallmarks of cancer (Fig. 7a). Among all the cancer hallmarks, COL1A2 mostly correlated with EMT and angiogenesis in 32 malignant solid cancers (EMT: r=0.86, p<2.2e-16; Angiogenesis: r=0.73, p<2.2e-16) in the overall TCGA pan-cancer cohort (Fig. 7b & c), or in individual tumor types (Fig. 7d & e). Considering EMT and angiogenesis are two critical biological features which contribute to the initial development of cancer, we reasonably speculated that COL1A2 serves a significant role in malignant progression in pan-cancer.

Discussion

NAFLD is a major public health problem worldwide and is becoming a leading cause of chronic liver disease [2]. As a complex heterogeneous disease, NAFLD results from both intrinsic susceptibility and environmental factors, and it is characterized by distinctive pathological features including hepatic steatosis, liver fibrosis, and chronic inflammation [3]. As the advanced stage of NAFLD, NASH acts as a key step that induces cirrhosis even HCC. However, not every NAFL patient develops to NASH, and the conversion rate is about 25% [3]. Therefore, there is an urgent need to classify NAFL patients into different risk subgroups and tailor personalized therapeutic strategies. Obviously, significantly altered signaling pathways and key regulators concerning NAFLD progression determine whether patients with simple NAFL develop to NASH or not. In this study, we attempted to identify pathways and genes closely connected with NAFLD progression and applied them to risk stratification of NAFL samples using a series of bioinformatic methods and machine learning approaches.

Firstly, we investigated the significantly dysregulated pathways in NASH using NAFL as the comparative baseline. Notably, "ECM organization" was identified as the most crucial biological process that drives NAFL to NASH. In a training cohort that contains 51 NAFL and 47 NASH, a total of 182 candidate genes involved in NAFLD progression were identified using an integrated strategy of WGCNA and DEGs screening. GO enrichment analysis confirmed that the 182 genes were mainly enriched in "ECM organization" etc., indicating their explicit role in NAFLD progression. Subsequently, the

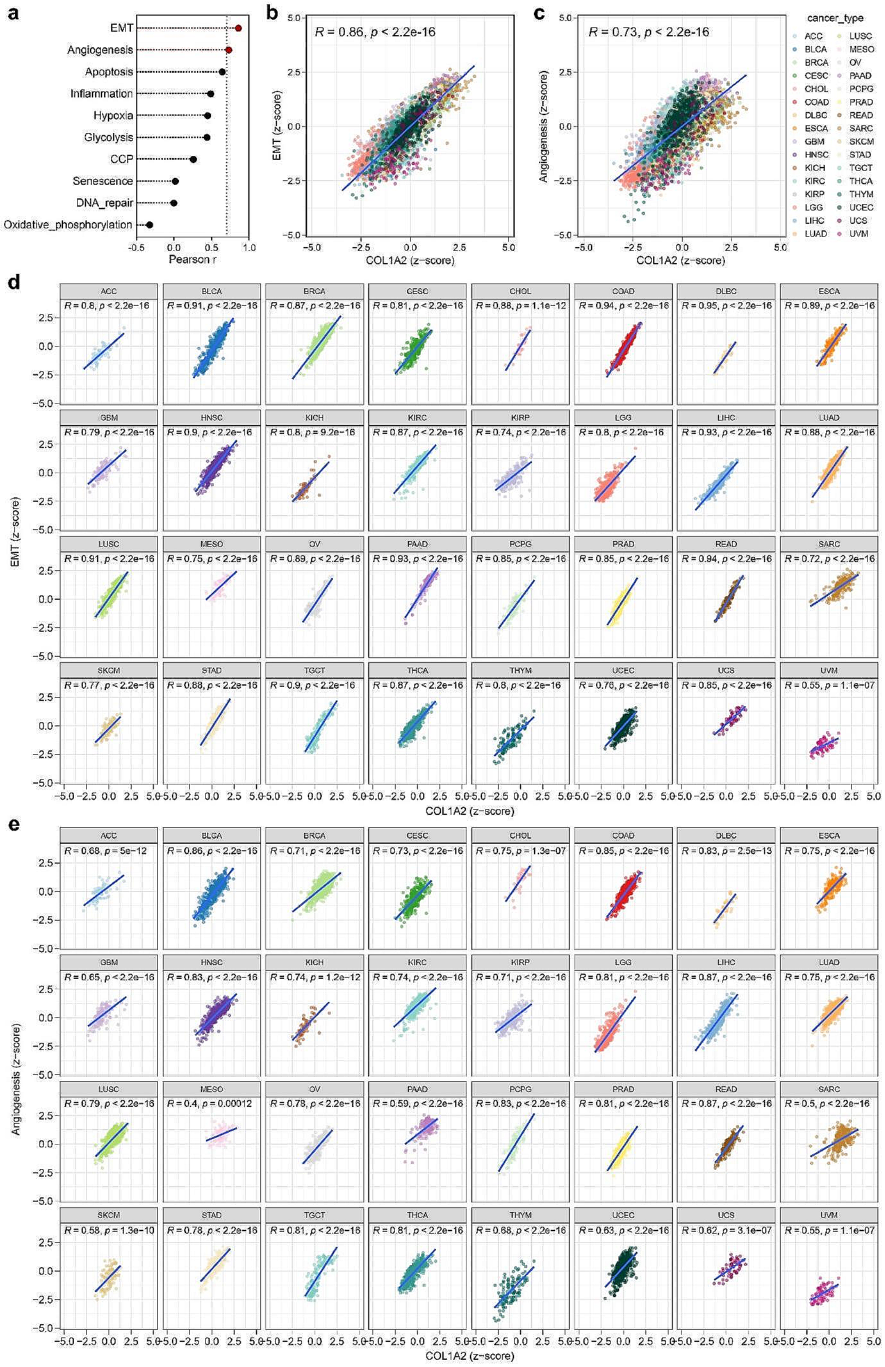

Fig. 7 COL1A2 significantly correlates with EMT and angiogenesis in pan-cancer. **(a)** Correlation landscape of COL1A2 expression and ten cancer hallmarks including EMT, angiogenesis, apoptosis, inflammation, hypoxia, glycolysis, cell cycle progression (CCP), senescence, DNA repair and oxidative phosphorylation in pan-cancer. **(b & c)** Among all the cancer hallmarks, COL1A2 mostly correlated with EMT and angiogenesis in 32 solid malignancies (EMT: $r=0.86$, $p<2.2e\text{-}16$; Angiogenesis: $r=0.73$, $p<2.2e\text{-}16$) in the overall TCGA pan-cancer cohort. **(d & e)** COL1A2 exhibits significantly positive correlations with EMT and angiogenesis in individual cancer type

182 candidate genes were submitted to the CMap algorithm for drug repositioning, and HDACi exhibited the highest predictive score as a potential drug for the treatment of NAFLD progression. Interestingly, a recent study reported that HDACi givinostat could

attenuate NAFLD and liver fibrosis [40], which raises the possibility that HDACi could be an actionable strategy for NAFLD.

Using NMF algorithm with the expression profile of the 182 genes, the 51 NAFL patients were divided into two groups (NAFL-C1 and C2), and the two groups exhibited distinct molecular and pathological features. In detail, significantly higher levels of inflammation, ECM organization and cell-cell adhesion were observed in C1, and the expression patterns of inflammatory factors in C1 were more similar to NASH. No significant difference of Th2 infiltration was observed among NAFL-C1, C2 and NASH. However, the distribution patterns of fibroblast abundance, Th1 infiltration and stroma score of C1 samples were extremely similar to NASH. It is widely acknowledged that Th1/2 balance affects the progression of fibrosis in various tissues and organs including hepatology [41], and activated fibroblasts are classically associated with fibrosis severity in disease progression [42]. Thus, we reasonably speculated that although NAFL-C1 samples appear in an early stage of NAFLD and diagnosed as NAFL, they actually have a high similarity of inflammatory response and fibrotic potential with NASH and possess an intrinsic tendency to develop to an advanced stage. Compared with NAFL-C2 patients, C1 patients may benefit from more intensive surveillance and preventive intervene.

To evaluate the risk level of NAFLD in a quantitative method, we combined different machine learning approaches including RF, SVM, and Lasso LR to screen for robust biomarkers to discriminate more severe NAFLD cases. Four genes named DTNA, COL4A2, UBD, and COL1A2 were finally screened out after the rigorous selective procedure, and COL1A2 held the highest logistic regression coefficient among the four genes. COL1A2 belongs to the type I collagen gene family and was reported to mediate ECM deposition and promote fibrotic diseases [43]. In independent validation cohorts, the established risk-stratification gene signature could discriminate advanced NAFLD samples. In addition, all the four genes are significantly upregulated in HCC samples compared to normal liver. In particular, COL1A2 is specifically expressed in fibroblasts involved in HCC TME, indicating that COL1A2 might play a critical role in fibroblast activation and fibrosis severity during NAFLD progression and development to HCC. Furthermore, COL1A2 also significantly correlates with EMT and angiogenesis in pan-cancer, suggesting its significant role in the initial development and malignant progression of solid cancers.

In addition, NAFLD-HCC samples are characterized by high expression of COL1A2 with wild-type CTNNB1. By analyzing mutational patterns of TCGA-HCC samples, we found that CTNNB1 is much more frequently mutated in low-COL1A2 HCC samples, while the mutation frequency of TP53 has no significant correlation with COL1A2 expression as a comparison. These findings indicated that Wnt/β-catenin/COL1A2 axis might play an important role in fibrosis severity in NAFLD-HCC progression, while inactivation of β-catenin might attenuate the progression of fibrosis through downregulation of fibrotic progression-related key genes such as COL1A2. Previous bioinformatic studies have shown that COL1A2 has a diagnostic value and may play an important role in NAFLD progression [44, 45], which is consistent with our results. Govaere et al. [13] previously reported a 25-gene signature for steatohepatitis and fibrosis in NAFLD using pairwise comparison and logistic regression analysis, and we observed our genes COL1A2 and DTNA were overlapped with Govaere's gene signature. In comparison, we integrated multiple machine learning approaches to screen for robust biomarkers, and our results showed that four genes are enough to discriminate advanced stages and fibrosis levels for NAFLD with favorable performances. We believe the advantage of our gene signature is simpler and more workable in clinical practice.

In summary, our study provided evidence for the necessity of molecular classification for NAFLD, and we established a risk-stratification gene signature to quantify risk assessment, aiming to identify the high-risk subset and to guide personalized treatment. We hope this work could facilitate the personalized therapeutic strategy for NAFLD, especially those patients appear in early stage but actually with high-risk NAFL.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40537-024-00899-5.

Supplementary Material 2

Supplementary Material 2

Supplementary Material 3

Acknowledgements

We acknowledge all authors and colleagues participating in this study for their work and helpful comments.

Author contributions

X.Z., Z.F., and J.Y. conceived and planned the study design; J.S., R.S., Y.W., and Y.L. performed formal analysis and data interpretation; J.S., R.S., and L.N. wrote the original draft; Y.C., Q.Y., L.Y., S.Z., C.Z., X.W., Q.W., and X.J. provided critical revisions and contributed to the editing of the paper. All authors read and approved the final manuscript.

Funding

This study was supported by National Natural Science Foundation of China (for Jing Sun, Grant No. 82203850), Jiangsu Funding Program for Excellent Postdoctoral Talent (for Run Shi, Grant No. 2022ZB427), Postdoctoral International Exchange Program (for Run Shi, Grant No. YJ20220433), the Chinese Medicine Project of Health Commission of Hubei Province (for Xuanbin Wang, Grant No. ZY2021Z010), the Advantages Discipline Group (Medicine) Project in Higher Education of Hubei Province (for Xuanbin Wang, Grant No. 2022XKQY3), the Science and Technology Development Fund, Macau SAR (for Qibiao Wu, Grant No. 0098/2021/A2), and a project funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

Data availability

All presented data in this study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

This study is based on published or public datasets and does not require ethical approval and consent.

Competing interest

The authors declare no potential conflicts of interest.

Author details

[1]Department of Endocrinology, Jiangsu Province Hospital of Chinese Medicine, Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China
[2]The First School of Clinical Medicine, Nanjing University of Chinese Medicine, Nanjing, China
[3]Department of Oncology, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China
[4]Pancreas Center, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China
[5]Department of Gastroenterology, Jiangsu Province Hospital of Chinese Medicine, Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China
[6]Laboratory of Chinese Herbal Pharmacology, Department of Pharmacology, Hubei Key Laboratory of Wudang Local Chinese Medicine Research, Renmin Hospital, Hubei University of Medicine, Shiyan, Hubei, China
[7]State Key Laboratory of Quality Research in Chinese Medicine, Faculty of Chinese Medicine, Macau University of Science and Technology, Taipa, Macau, China
[8]Department of Endocrinology, Suzhou Ninth Hospital Affiliated to Soochow University, Suzhou, China
[9]Institute of Hypertension, Jiangsu Province Hospital of Chinese Medicine, Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China

Received: 25 March 2023 / Accepted: 4 March 2024

Published online: 15 March 2024

References

1. Targher G, Tilg H, Byrne CD. Non-alcoholic fatty liver disease: a multisystem disease requiring a multidisciplinary and holistic approach. Lancet Gastroenterol Hepatol. 2021;6(7):578–88.
2. Younossi ZM. Non-alcoholic fatty liver disease - A global public health perspective. J Hepatol. 2019;70(3):531–44.
3. Diehl AM, Day C. Cause, Pathogenesis, and treatment of Nonalcoholic Steatohepatitis. N Engl J Med. 2017;377(21):2063–72.
4. Huang DQ, El-Serag HB, Loomba R. Global epidemiology of NAFLD-related HCC: trends, predictions, risk factors and prevention. Nat Rev Gastroenterol Hepatol. 2021;18(4):223–38.
5. Stefan N, Haring HU, Cusi K. Non-alcoholic fatty liver disease: causes, diagnosis, cardiometabolic consequences, and treatment strategies. Lancet Diabetes Endocrinol. 2019;7(4):313–24.
6. Buzzetti E, Pinzani M, Tsochatzis EA. The multiple-hit pathogenesis of non-alcoholic fatty liver disease (NAFLD). Metabolism. 2016;65(8):1038–48.
7. Ipsen DH, Lykkesfeldt J, Tveden-Nyborg P. Molecular mechanisms of hepatic lipid accumulation in non-alcoholic fatty liver disease. Cell Mol Life Sci. 2018;75(18):3313–27.
8. Tilg H, Adolph TE, Moschen AR. Multiple parallel hits hypothesis in nonalcoholic fatty liver disease: Revisited after a Decade. Hepatology. 2021;73(2):833–42.
9. Haukeland JW, Damas JK, Konopski Z, Loberg EM, Haaland T, Goverud I, Torjesen PA, Birkeland K, Bjoro K, Aukrust P. Systemic inflammation in nonalcoholic fatty liver disease is characterized by elevated levels of CCL2. J Hepatol. 2006;44(6):1167–74.
10. Miura K, Yang L, van Rooijen N, Ohnishi H, Seki E. Hepatic recruitment of macrophages promotes nonalcoholic steatohepatitis through CCR2. Am J Physiol Gastrointest Liver Physiol. 2012;302(11):G1310–1321.
11. Kozumi K, Kodama T, Murai H, Sakane S, Govaere O, Cockell S, Motooka D, Kakita N, Yamada Y, Kondo Y, et al. Transcriptomics identify Thrombospondin-2 as a biomarker for NASH and Advanced Liver Fibrosis. Hepatology. 2021;74(5):2452–66.
12. Subudhi S, Drescher HK, Dichtel LE, Bartsch LM, Chung RT, Hutter MM, Gee DW, Meireles OR, Witkowski ER, Gelrud L, et al. Distinct hepatic gene-expression patterns of NAFLD in patients with obesity. Hepatol Commun. 2022;6(1):77–89.
13. Govaere O, Cockell S, Tiniakos D, Queen R, Younes R, Vacca M, Alexander L, Ravaioli F, Palmer J, Petta S et al. Transcriptomic profiling across the nonalcoholic fatty liver disease spectrum reveals gene signatures for steatohepatitis and fibrosis. Sci Transl Med 2020, 12(572).
14. Pinyol R, Torrecilla S, Wang H, Montironi C, Pique-Gili M, Torres-Martin M, Wei-Qiang L, Willoughby CE, Ramadori P, Andreu-Oller C, et al. Molecular characterisation of hepatocellular carcinoma in patients with non-alcoholic steatohepatitis. J Hepatol. 2021;75(4):865–78.
15. Cancer Genome Atlas Research Network. Electronic address wbe, Cancer Genome Atlas Research N: Comprehensive and Integrative genomic characterization of Hepatocellular Carcinoma. Cell. 2017;169(7):1327–41. e1323.
16. Consortium GT. The genotype-tissue expression (GTEx) project. Nat Genet. 2013;45(6):580–5.
17. Ma L, Hernandez MO, Zhao Y, Mehta M, Tran B, Kelly M, Rae Z, Hernandez JM, Davis JL, Martin SP, et al. Tumor Cell Biodiversity drives Microenvironmental Reprogramming in Liver Cancer. Cancer Cell. 2019;36(4):418–e430416.
18. Massalha H, Bahar Halpern K, Abu-Gazala S, Jana T, Massasa EE, Moor AE, Buchauer L, Rozenberg M, Pikarsky E, Amit I, et al. A single cell atlas of the human liver tumor microenvironment. Mol Syst Biol. 2020;16(12):e9682.
19. Meng Y, Zhao Q, An L, Jiao S, Li R, Sang Y, Liao J, Nie P, Wen F, Ju J, et al. A TNFR2-hnRNPK Axis promotes primary Liver Cancer Development via activation of YAP Signaling in hepatic progenitor cells. Cancer Res. 2021;81(11):3036–50.
20. Sun J, Zhao T, Zhao D, Qi X, Bao X, Shi R, Su C. Development and validation of a hypoxia-related gene signature to predict overall survival in early-stage lung adenocarcinoma patients. Ther Adv Med Oncol. 2020;12:1758835920937904.
21. Sun J, Shi R, Zhang X, Fang D, Rauch J, Lu S, Wang X, Kasmann L, Ma J, Belka C, et al. Characterization of immune landscape in papillary thyroid cancer reveals distinct tumor immunogenicity and implications for immunotherapy. Oncoimmunology. 2021;10(1):e1964189.
22. Shi R, Bao X, Unger K, Sun J, Lu S, Manapov F, Wang X, Belka C, Li M. Identification and validation of hypoxia-derived gene signatures to predict clinical outcomes and therapeutic responses in stage I lung adenocarcinoma patients. Theranostics. 2021;11(10):5061–76.
23. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
24. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313(5795):1929–35.
25. Li T, Fan J, Wang B, Traugh N, Chen Q, Liu JS, Li B, Liu XS. TIMER: a web server for Comprehensive Analysis of Tumor-infiltrating Immune cells. Cancer Res. 2017;77(21):e108–10.
26. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, Khodadoust MS, Esfahani MS, Luca BA, Steiner D, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. 2019;37(7):773–82.
27. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, Krogsdam A, Loncova Z, Posch W, Wilflingseder D, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Med. 2019;11(1):34.
28. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, Selves J, Laurent-Puig P, Sautes-Fridman C, Fridman WH, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 2016;17(1):218.
29. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017;18(1):220.
30. Racle J, Gfeller D. EPIC: a Tool to Estimate the proportions of different cell types from bulk gene expression data. Methods Mol Biol. 2020;2120:233–48.
31. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, Trevino V, Shen H, Laird PW, Levine DA, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4:2612.
32. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018;28(11):1747–56.
33. Wang S, Li H, Song M, Tao Z, Wu T, He Z, Zhao X, Wu K, Liu XS. Copy number signature analysis tool and its application in prostate cancer reveals distinct mutational processes and clinical outcomes. PLoS Genet. 2021;17(5):e1009557.

34. Shi R, Wang X, Wu Y, Xu B, Zhao T, Trapp C, Wang X, Unger K, Zhou C, Lu S, et al. APOBEC-mediated mutagenesis is a favorable predictor of prognosis and immunotherapy for bladder cancer patients: evidence from pan-cancer analysis and multiple databases. Theranostics. 2022;12(9):4181–99.
35. Harding JJ, Nandakumar S, Armenia J, Khalil DN, Albano M, Ly M, Shia J, Hechtman JF, Kundra R, El Dika I, et al. Prospective genotyping of Hepatocellular Carcinoma: clinical implications of Next-Generation sequencing for matching patients to targeted and Immune therapies. Clin Cancer Res. 2019;25(7):2116–26.
36. Schulze K, Imbeaud S, Letouze E, Alexandrov LB, Calderaro J, Rebouissou S, Couchy G, Meiller C, Shinde J, Soysouvanh F, et al. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nat Genet. 2015;47(5):505–11.
37. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics. 2010;26(7):976–8.
38. Yu G, Lam TT, Zhu H, Guan Y. Two methods for mapping and visualizing Associated Data on Phylogeny using Ggtree. Mol Biol Evol. 2018;35(12):3041–3.
39. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular signatures database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417–25.
40. Huang HM, Fan SJ, Zhou XR, Liu YJ, Li X, Liao LP, Huang J, Shi CC, Yu L, Fu R, et al. Histone deacetylase inhibitor givinostat attenuates nonalcoholic steatohepatitis and liver fibrosis. Acta Pharmacol Sin. 2022;43(4):941–53.
41. Wynn TA. Fibrotic disease and the T(H)1/T(H)2 paradigm. Nat Rev Immunol. 2004;4(8):583–94.
42. Kendall RT, Feghali-Bostwick CA. Fibroblasts in fibrosis: novel roles and mediators. Front Pharmacol. 2014;5:123.
43. Ramirez F, Tanaka S, Bou-Gharios G. Transcriptional regulation of the human alpha2(I) collagen gene (COL1A2), an informative model system to study fibrotic diseases. Matrix Biol. 2006;25(6):365–72.
44. Gao R, Wang J, He X, Wang T, Zhou L, Ren Z, Yang J, Xiang X, Wen S, Yu Z, et al. Comprehensive analysis of endoplasmic reticulum-related and secretome gene expression profiles in the progression of non-alcoholic fatty liver disease. Front Endocrinol (Lausanne). 2022;13:967016.
45. Zheng J, Wu H, Zhang Z, Yao S. Dynamic co-expression modular network analysis in nonalcoholic fatty liver disease. Hereditas. 2021;158(1):31.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.