Single cell transcriptomics of hepatic stellate cells uncover crucial pathways and key regulators involved in non-alcoholic steatohepatitis

Weiwei He 1,2,3, Caoxin Huang 2,3, Xiulin Shi 2,3, Menghua Wu 1,2,3, Han Li 1-4, Qihong Liu 1,2, Xiaofang Zhang 2,3, Yan Zhao 2,3*, Xuejun Li 2,3,4,5*

Affiliations:
1. School of Medicine, Xiamen University, Xiamen, China;
2. Xiamen Diabetes Institute, The First Affiliated Hospital of Xiamen University, Xiamen, China;
3. Fujian Provincial Key Laboratory of Translational Medicine for Diabetes, China;
4. Department of Endocrinology and Diabetes, The First Affiliated Hospital of Xiamen University, Xiamen, China;
5. Xiamen Clinical Medical Center for Endocrine and Metabolic Diseases, Xiamen Diabetes Prevention and Treatment Center, Xiamen, China.

*Corresponding author: Prof. Xuejun Li, Department of Endocrinology and Diabetes, The First Affiliated Hospital, Xiamen University; No.55 Zhenhai Road, Xaimen 361003, China; Phone: 86-592-2137610; Fax: 86-592-2137557; Email: lixuejun@xmu.edu.cn. Prof. Yan Zhao, Xiamen Diabetes Institute, The First Affiliated Hospital of Xiamen University, Xiamen, China; Phone: 86-592-2137610; Fax: 86-592-2137557; Email: zhaoyan1976@126.com.

These two authors contributed to the article equally.
Abstract

Background: Fibrosis is an important pathological process in the development of non-alcoholic steatohepatitis (NASH), and the activation of hepatic stellate cell (HSC) is a central event in liver fibrosis. However, the transcriptomic change of activated HSCs (aHSCs) and resting HSCs (rHSCs) in NASH patients has not been assessed. This study aimed to identify transcriptomic signature of HSCs during the development of NASH and the underlying key functional pathways.

Methods: NASH-associated transcriptomic change of HSCs was defined by single cell RNA-sequencing (scRNA-seq) analysis, and those top up-regulated genes were identified as NASH-associated transcriptomic signatures. Those functional pathways involved in the NASH-associated transcriptomic change of aHSCs were explored by weighted gene co-expression network analysis (WGCNA) and functional enrichment analyses. Key regulators were explored by upstream regulator analysis and transcription factor enrichment analysis.

Results: scRNA-seq analysis identified numerous differentially expressed genes in both rHSCs and aHSCs between NASH patients and healthy controls. Both scRNA-seq analysis and in-vivo experiments showed the existence of rHSCs (mainly expressing α-SMA) in the normal liver and the increased aHSCs (mainly expressing Collagen 1) in the fibrosis liver tissues. NASH-associated transcriptomic signature of rHSC (NASHrHSCsignature) and NASH-associated transcriptomic signature of aHSC (NASHaHSCsignature) were identified. WGCNA revealed the main pathways correlated with transcriptomic change of aHSCs. Several key upstream regulators and transcription factors of determining the functional change of aHSCs in NASH were identified.

Conclusion: This study developed a useful transcriptomic signature with the potential in assessing the fibrosis severity in the development of NASH. This study also identified the main pathways in the activation of HSCs during the development of NASH.
Keywords: Non-alcoholic steatohepatitis; Hepatic stellate cell; Fibrosis

1. Introduction

Non-alcoholic fatty liver disease (NAFLD), as the main liver disease manifestation of metabolic diseases, is closely related to obesity, insulin resistance, type 2 diabetes, and dyslipidemia [1, 2]. With the improvement of living standards and changes in lifestyles, the incidence of NAFLD has been increasing worldwide, and it has become one of the most important causes of chronic liver dysfunction [2, 3]. NAFLD ranges from simple steatosis to nonalcoholic steatohepatitis (NASH), which is characterized by ballooning of hepatocytes and inflammation [4-6]. NAFLD is now the most common chronic liver disease with a global prevalence of nearly 25%, while the global prevalence of NASH is estimated at 3% to 5% and NASH is the second leading cause of liver transplantation and poses a heavy financial burden [7-9].

Fibrosis is a major determinant of clinical outcome in patients with NASH and is associated with an increased risk of cirrhosis and liver cancer [10-12]. In addition, the occurrence of liver fibrosis is a significant predictor of all-cause and liver-related disease mortality in NAFLD, and the risk of liver-related mortality increases exponentially with higher fibrosis stages [13, 14]. However, there is lack of effective treatments for fibrosis in NASH [15]. Therefore, it is important to explore the mechanisms of liver fibrosis of NASH and uncover the key treatment targets. Liver fibrosis is mainly characterized by overproduction and deposition of extracellular matrix (ECM) [16]. Under normal condition, ECM produced by hepatic stellate cells (HSCs) during liver injury and tissue repair can be hydrolyzed by matrix metalloproteinases (MMPs). However, an imbalance between ECM synthesis and degradation can lead to the development of liver fibrosis [16-18]. Activation of the HSCs is currently regarded as a central event in liver fibrosis [19, 20]. Studies over the last two decades have shown that myofibroblasts (MFBs) are the main cells producing ECM during various chronic liver injuries, while HSCs remain the main source of MFBs in various
clinical and experimental liver fibrosis models [21-23]. HSCs can transform into MFBs and express α-SMA in large quantities, and this transformation is a key event in the progression of liver fibrosis [24].

NASH-related fibrosis is an important pathological process in the progression of NAFLD, but the risk of developing fibrosis cannot be predicted by liver tissue biopsy. Fibrosis is an important pathological process in the development of NASH, and the activation of HSC is a central event in liver fibrosis. However, the transcriptomic change of activated HSCs (aHSCs) and resting HSCs (rHSCs) in NASH patients has not been assessed. In this study, we analyzed the single cell RNA-sequencing (scRNA-seq) data of liver tissues from NASH patients and healthy controls. This study aimed to identify transcriptomic signature of HSCs during the development of NASH and the underlying key functional pathways.

2. Methods

2.1 scRNA-seq data of NASH and healthy control liver tissues

scRNA-seq data of NASH and healthy control liver tissues from Gene Expression Omnibus (GEO) were analyzed in our study. The data were from GSE136103 (Supplementary table 1). The Seurat package was used in the scRNA-seq analyses, and SCTtransform approach was used in integrating data from multiple samples [25]. Cells were clustered with both uniform manifold approximation and projection (UMAP) and t-distributed stochastic neighbor embedding (t-SNE) methods, and the types of cell clusters were defined with previously defined makers of distinct cells. The signature genes highly expressed in each cluster but not in the other clusters were identified with Seurat package, and the differentially expressed genes (DEGs) of HSCs between NASH patients and controls were further calculated.

2.2 Transcriptomic profile datasets

We searched the transcriptomic profile datasets of liver tissues from NAFLD patients in GEO. Ultimately, GSE49541, GSE126848 and GSE130970 were included as validation datasets of our study, and all those 3 datasets were transcriptomic profile datasets of liver tissues from NAFLD patients. In addition,
we used GSE148849, a transcriptomic dataset of HSC stimulated by TGF-β in vitro, to assess the changes of NASH-related signature gene sets during HSCs activation. DESeq2 of R package was used to determine DEGs for RNA sequencing (RNA-seq) datasets, and limma package was used to determine DEGs for microarray datasets. We further validated the major findings from scRNA-seq analyses by bulk RNA-seq data containing 206 patients with NAFLD and further evaluated the clinical significance of key transcriptomic signatures (GSE135251). The characteristics of those publicly available transcriptomic datasets used in the analysis are shown supplementary table 2.

2.3 Gene set variation analysis (GSVA)

GSVA is a non-parametric unsupervised analysis method mainly used to evaluate the results of gene set enrichment in transcriptomes [26]. To verify the enrichment of the HSC-related signature gene sets in scRNA-seq data, we used GSM4041162 for GSVA. In addition, to verify the enrichment of the signature gene set for TGF-β stimulated HSC activation, we used GSE148849 for GSVA.

2.4 Gene set enrichment analysis (GSEA)

GSEA is a computational method that determines whether an a priori defined set of genes shows statistically significant and consistent differences between two biological states [27]. To determine the enrichment of the signature gene set for TGF-β stimulated HSC activation, we used GSE148849 for GSEA. In addition, to explore the enrichment of the signature gene sets developed above in the Differentially expressed genes (DEGs) list for the progression from NAFL to NASH or fibrosis progression, we also performed GSEA using data from our previous study [28].

2.5 Weighted gene co-expression network analysis (WGCNA)

To further clarify the possible functional pathways related to the enrichment of signature gene sets, WGCNA was further performed using the R package "WGCNA". WGCNA is a systematic method used to describe the gene co-expression patterns between different samples that can be used to identify highly synergistic co-expression modules [29]. GSE49541 was used in this study.

Accepted Manuscript published as EC-22-0502. Accepted for publication: 23-Dec-2022
including 72 patients with NAFLD.

2.6 Upstream regulator and transcription factor analyses

In this study, we further explored those key regulators or transcription factor analyses of determining the functional change of aHSCs in NASH. To infer the upstream regulators from the DEGs in aHSCs between NASH and controls, Quaternary test statistical analyses were performed with the R package of QuaternaryProd [30]. To identify those transcription factors of determining the functional change of aHSCs in NASH, transcriptional regulatory networks of those DEGs in aHSCs between NASH and controls were explored with the Cytoscape plugin of iRegulon, which could detect enriched transcription factor motifs and their direct targets [31].

2.7 Experimental animals

C57L/J male mice were obtained from Xiamen University Experimental Animal Center. Mice were fed standard rodent chow ad libitum and housed on woodchip beds. Mice were randomly divided into NASH group and Control group at 12 weeks of age. The NASH mice model group was fed a high-fat and choline methionine-deficient diet (HFMCD) (A06071301B16; Research Diets, New Brunswick NJ) for 10 weeks. The control mice were fed standard rodent chow ad libitum.

DBA/2J mice were purchased from Beijing Vital River Laboratory Animal Technology Co., Ltd. and the mice were placed in the Xiamen University Experimental Animal Center. Mice were administered intraperitoneally at a dose of 5 ul of 10% CCl4 olive oil solution per gram of body weight. CCl4 injections were started at 12 weeks of age, twice a week for 6 weeks. Olive oil was used in control mice.

2.8 Histological Examination

The liver specimens of each group were fixed in 10% buffered neutral formalin for 24 h. The fixed specimens were routinely processed to obtain 4 ~ 5 μm thick paraffin sections for histological and immunohistochemical evaluation. Sections of each group were stained with Masson trichrome and hematoxylin and eosin (H&E) stain. Collagen 1 (1:200, ab5694, abcam) and anti-a-SMA
(1:200, 72026T, CST) were used for immunohistochemical staining. The samples were then observed and photographed under a microscope. For immunofluorescence staining, liver sections were stained with FITC fluorescent anti-a-SMA antibody (72026T, CST) and cy3 fluorescent anti-Collagen 1 antibody (ab5694, abcam). Nuclei were then stained with DAPI, and the liver sections were then observed and photographed using a fluorescence microscope.

2.9 Statistical analysis

Expressions of key genes in liver tissues were presented with median and interquartile range (IQR), and difference between advanced fibrosis liver tissues and mild fibrosis liver tissues was determined with Mann-Whitney U test. Difference in the enrichment scores from GSVA between groups was determined with unpaired t-test. The performance of enrichment scores from GSVA in diagnosing fibrosis or NASH among NAFLD patients was explored by receiver operating characteristic (ROC) analysis and area under the curve (AUC) was calculated. R software (version 3.6.1) was used in statistical analyses. P < 0.05 was considered to be significant.

3. Results

3.1 Single cell transcriptomics in liver identified HSCs subset-specific transcriptomic signatures

A total of 6 liver samples of single cell transcriptomics were included into our study, and they were GSM4041151, GSM4041156, GSM4041157, GSM4041159, GSM4041162 and GSM4041163, respectively. The main characteristics of these samples were shown in supplementary table 1. These 6 samples contained 4 healthy controls and 2 NAFLD samples with fibrosis. After integration by SCTtransform approach, scRNA-seq analyses were subsequently performed. Those cells were classified into 19 clusters by the tSNE or UMAP algorithm (Figure 1-A and B). Two main HSCs subsets including aHSC and rHSC were identified, and they were both identified in the liver tissues of NASH patients and healthy controls (Figure 1-C).

Through scRNA-seq analyses, we further identified the signature genes expressed by each cell cluster. The signature genes of aHSC and rHSC were
shown in Table 1. For instance, rHSCs highly expressed RGS5, NDUFA4L2, MYH11, RERGL, etc., and aHSC highly expressed LUM, COL3A1, DPT, PCOLCE, etc. To facilitate the analysis of HSCs alterations using bulk transcriptomic data, we constructed the transcriptomic signature of activated HSC (aHSCsignature) and transcriptomic signature of resting HSC (rHSCsignature) based on the above results. The aHSCsignature consisted of the top 54 genes that were highly expressed in aHSCs. rHSCsignature consisted of the top 79 genes that highly expressed in rHSCs (Table 1). Supplementary figure 1 shows the genes that were highly expressed in each cell subpopulation in the scRNA-seq analyses (Supplementary figure 1).

To verify whether the above-mentioned transcriptomic signatures can be used to define the corresponding HSC subsets, we evaluated the enrichment of the above two transcriptomic signatures in liver tissue of GSM4041162 using GSVA (Supplementary figure 2). As shown in Supplementary figure 2, rHSCsignature significantly enriched in rHSCs cluster, but not in other cell clusters, indicating that the rHSCsignature can be used to represent rHSCs. aHSCsignature significantly enriched in aHSCs cluster but not in other cell clusters, indicating that the aHSCsignature can be used to represent aHSCs (Supplementary figure 2).

3.2 Single cell transcriptomics identified NASH-specific transcriptomic signatures in both rHSC and aHSCs

We identified the genes abnormally changed in each HSC subset of NASH patients through scRNA-seq analysis (Figure 2-A and B). As shown in Figure 2, aHSCs in NASH patients expressed higher levels of IFITM1, MIF, COL1A1, COL1A2, COL3A1, etc. than aHSCs in controls, and rHSCs in NASH patients expressed higher levels of ID3, NOTCH3, CRIP1, etc. than rHSC in controls (Table 1). Single cell transcriptomics of HSCs reveled that aHSCs and rHSCs both existed in the liver tissues of NASH patients and healthy controls, but there was an obviously increased transition from rHSCs to aHSCs in NASH patients (Figure 3-A). Based on the above findings, we established NASH-associated transcriptomic signature of rHSC (NASHrHSCsignature) and NASH-associated
transcriptomic signature of activated HSC (NASHaHSCsignature), respectively (Table 1). GSVA validated that NASHaHSCsignature obviously enriched in the aHSCs cluster of NASH patients, and NASHrHSCsignature obviously enriched in the rHSCs cluster of NASH patients, suggesting that the above two transcriptomic signatures could represent the respective cell subpopulations well (Figure 3-B and C). For instance, NASHaHSCsignature was significantly enriched in aHSCs in NASH patients but not in other cell subpopulations, suggesting that this transcriptomic signature can be used to represent NASH-associated aHSCs (Figure 3-B and C).

In addition, we also used transcriptomic data of HSCs stimulated by TGF-β (GSE148849) to assess the changes of those transcriptomic signatures above in HSCs activation. GSEA validated that NASHaHSCsignature was significantly increased in TGF-β-induced HSCs activation (Figure 4-A and B). Moreover, the enrichment score of NASHaHSCsignature calculated by GSVA was significantly increased in TGF-β-induced HSCs activation (Figure 4-C). The outcomes above suggested that NASHaHSCsignature could represent the activation state of HSCs.

We further validated the enrichment of NASHaHSCsignature in liver tissues of NASH patients or NAFLD patients with advanced fibrosis using the results of our previous study. We found that NASHaHSCsignature was significantly enriched in the liver of NASH patients and NAFLD patients with advanced fibrosis (Figure 4-D and E). It is further illustrated that the enrichment score of NASHaHSCsignature could represent the severity of fibrosis in NAFLD. In addition, aHSCsignature was significantly enriched in NASH patients and NAFLD with advanced fibrosis patients, indicating an increased number of aHSCs in the liver of NASH patients and NAFLD patients with advanced fibrosis (Figure 4-D and E). As shown in Figure 4-F, GSEA enrichment plots showing the significantly increased enrichment of NASHaHSCsignature in the livers of patients with NASH and advanced fibrosis of NAFLD (Figure 4-F).

We next validated the increased enrichment of NASHaHSCsignature in the livers of NAFLD patients with advanced fibrosis in bulk transcriptomic data of
GSE49541 and GSE130970, which included patients with NAFLD with or without fibrosis. As shown in Figure 5, the enrichment scores of NASHaHSCsignature calculated by GSVA in NAFLD patients with advanced fibrosis were significantly higher than that in patients without or with mild fibrosis (P<0.05; Figure 5-A). ROC analysis revealed enrichment score of NASHaHSCsignature in patients with NAFLD could help to diagnose advanced fibrosis (Figure 5-B). The outcomes showed that NASHaHSCsignature could effectively predict the advanced fibrosis in NAFLD patients. However, NASHaHSCsignature had limited performance in predicting NASH progression among NAFLD patients (Figure 5-C and D). Moreover, using a NASH mice model and CCl4-included fibrosis mice model, we validated the existence of rHSCs (mainly expressing α-SMA) in the normal liver and the increased aHSCs (mainly expressing Collagen 1) in the fibrosis liver tissues (Figure 6 and 7, Supplementary figure 3 and 4).

3.3 Crucial functional pathways involved in the increased transition from rHSCs to aHSCs in NAFLD patients

To uncover those key genes and functional pathways involved in the increased transition from rHSCs to aHSCs in NAFLD patients, we performed WGCNA analyses of liver transcriptome data in NAFLD patients, in which the enrichment scores of HSCs-relevant transcriptomic signatures were used as clinical traits. Crucial co-expression modules correlated with the enrichment scores of HSCs-relevant transcriptomic signatures were thus explored by WGCNA analyses. As shown in Figure 8, the co-expression pattern of genes in the liver transcriptome data of NAFLD patients were successfully constructed. The most significant co-expression module correlated with the enrichment score of NASHaHSCsignature was Memagenta module (Correlation coefficient = 0.76, P = 2.0E-11), followed by Meturquoise (correlation coefficient = 0.44, P = 0.001) and Megreenyellow module (correlation coefficient = 0.43, P = 0.001). The MEmagenta co-expression module was also associated with advanced fibrosis (correlation coefficient = 0.85, P = 5.0E-16) and aHSCsignature enrichment score (correlation coefficient = 0.87, P = 5.0E-18) (Figure 8-B). Functional
annotation analysis of Memagenta module showed that its functions were characterized by multiple ECM-related pathways such as extracellular matrix, extracellular matrix assembly and extracellular structural organization (Figure 8-D). Functional annotation analysis of MEgreenyellow and MEturquoise co-expression modules showed that their functions were both characterized by immune response-related pathways (Supplementary figure 5). Thus, WGCNA analysis confirmed that ECM-related pathways and immune-related pathways were the main functional pathways involved in the activation of HSCs during NAFLD progression.

3.4 Key regulators involved in the transcriptomic change of aHSCs in NASH

We further explored those key regulators or transcription factor analyses of determining the functional change of aHSCs in NASH. As shown Supplementary figure 6-A, several key upstream regulators of determining the functional change of aHSCs in NASH were identified such as SDC1, GRP, SDC4 and MUC1. In the transcription factor enrichment analysis, some enriched transcription factor motifs, transcription factors and their corresponding target genes were identified (Supplementary table 3). For example, JAZF1 and FOBL1 could regulate the expressions of more than 40 genes up-regulated in the aHSCs of NASH patients (Supplementary figure 6-B). In addition, we found that JunB was one of the most critical transcription factors in NASH-associated aHSC and could regulate the expressions of many signature genes in NASHaHSCsignature (Supplementary table 3). We used GSE148849, a transcriptomic dataset of HSC stimulated by TGF-β in vitro, to assess the change of JunB during HSCs activation, and we found that the expression of JunB in primary HSCs was significantly increased after TGF-β stimulation (P=0.008) (Supplementary figure 7).

3.5 Correlation between key signatures and fibrosis severity of NAFLD patients

We validated the findings with a liver sequencing dataset containing 206 patients with NAFLD and further evaluated the clinical significance of key transcriptomic signatures (GSE135251). The results showed that the GSVA
enrichment scores of aHSCsignature and NASHaHSCsignature were significantly increased in NASH patients (P<0.05) (Supplementary figure 8). GSVA enrichment scores of aHSCsignature and NASHaHSCsignature were significantly increased in the high NAS group (NAS 5-8) than the low NAS group (NAS 1-4) (P<0.001) (Supplementary figure 9), and both signature enrichment scores increased with the increase of the severity of NAFLD patients (P<0.05) (Supplementary figure 10). The GSVA enrichment scores of aHSCsignature and NASHaHSCsignature both increased significantly with the fibrosis severity of NAFLD patients (Supplementary figure 11 and 12). The findings suggested that changes in these HSCs-related signatures were able to show a progressive increase with the progression of liver fibrosis as well as NASH progression in common NAFLD patients. In addition, we analyzed the correlation between the expression of the top 10 up-regulated genes in aHSCsignature and NASHaHSCsignature and the severity of liver fibrosis in NAFLD patients (Supplementary table 4). The results showed that the expressions of many genes were significantly correlated with the severity of liver fibrosis in NAFLD patients, but several genes were not significantly correlated with the severity of liver fibrosis. Therefore, the transcriptomic signatures had better performance in assessing the severity of fibrosis in NAFLD patients compared to single gene (Supplementary table 4).

**Discussion**

Fibrosis in NASH is driven by the activation of HSCs, which transform from quiescent HSCs to myofibroblasts that produce collagen and other types of extracellular matrix [32, 33]. Activated HSCs, as the source of hepatic myofibroblasts in NASH, are the most critical factor of producing excessive ECM and causing advanced fibrosis in the liver of NASH patients [22]. This study was performed to identify transcriptomic signature of HSCs during the development of NASH and the underlying key functional pathways. This study defined the NASH-associated transcriptomic change of aHSCs, and developed a useful transcriptomic signature with the potential in assessing the fibrosis severity in the development of NASH. This study also identified extracellular
matrix-related pathways and immune-related pathways as key players in the activation of HSCs during the development of NASH. Therefore, this study identified NASH-associated transcriptomic signature of HSC and provided new insights into fibrosis progression of NAFLD.

Our study identified differentially expressed genes in rHSC and aHSCs based on single cell sequencing data and successfully established rHSCsignature and aHSCsignature, respectively. We then defined the NASH-associated transcriptomic changes of aHSCs, and developed a useful transcriptomic signature with the potential of representing the degree of HSCs activation and fibrosis severity in NASH. NASHaHSCsignature was significantly enriched during TGF-β stimulated HSCs activation in vitro, indicating that NASHaHSCsignature could represent the degree of HSCs activation. NASHaHSCsignature also had a strong diagnostic significance in assessing the severity of fibrosis of patients with NAFLD.

Many of those genes in aHSCsignature and NASHaHSCsignature have been reported to be closely associated with extracellular matrix and fibrosis. For instance, Lumican, encoded by LUM, has a key role in collagen assembly [34], and the formation of collagen fibers in the extracellular matrix of several tissues is regulated by lumican [35-37]. Data presented by Anuradha Krishnan et al. suggest that lumican plays an important role in the progression of liver fibrosis by maintaining the stability of collagen fibers during fibrosis [38]. Dermatopontin (DPT) has been shown to modulate collagen and fibrin fiber formation, induce cell adhesion and promote wound healing [39-41]. There is a study found that the expression of DPT is positively correlated with the severity of liver fibrosis [42]. ASPN plays a key role in tissue injury and regeneration [43]. It has been found that ASPN expression is increased in the mouse model of pulmonary fibrosis, and ASPN is mainly localized in α-SMA⁺ myofibroblasts. In vitro experiments demonstrated that ASPN knockout inhibited myofibroblast differentiation [43]. However, the mechanisms of other signature genes in the fibrosis of NAFLD patients are unclear and still need to be investigated by more studies.
Apart from ECM-related pathways, we found immune-related pathways were also key functional pathways involved in the activation of HSCs during NASH progression. There is emerging evidence supporting that innate and adaptive immune activation is the driving force in establishing liver inflammation and fibrosis in NASH [44]. Some previous studies have found that immune cells or cytokines have important roles in the development of inflammation and liver fibrosis in NASH [45]. In the liver tissue of NAFLD patients, immune cells such as monocytes and macrophages are involved in inflammation, thus promoting the progression of NAFLD to NASH [46-48]. Therefore, immune-related pathways have key roles in the activation of HSCs during NASH progression and may be promising treatment targets.

Our study provides useful information for clinical evaluation of activated HSCs in NASH patients and may help to improve the diagnosis and risk stratification of fibrosis. Activated HSCs are the main effector cells of liver fibrosis and can produce excess extracellular matrix through chronic liver injury. The NASHaHSCsignature developed in our study can represent the degree of HSCs activation and has a strong diagnostic significance for the severity of NAFLD with fibrosis. It may help clinicians to determine the risk of developing liver fibrosis in the future and help to take early interventions to inhibit or delay the development of fibrosis among NAFLD patients.

Key upstream regulators or transcription factors related to the functional changes of HSCs are potential treatment targets for liver fibrosis of NASH patients. In this study, we tried to uncover those key regulators or transcription factor analyses of determining the functional change of aHSCs in NASH by 2 bioinformatic methods. Several key upstream regulators related to the functional changes of HSCs in NASH were identified such as SDC1, GRP, SDC4 and MUC1. SDC1 encodes heparan sulfate proteoglycan syndecan-1, which has been reported to be positively associated with liver fibrosis in patients with chronic liver diseases [49, 50]. Knockdown of syndecan-1 could reduce the proliferation of keloid fibroblasts and the production of ECM [51]. Another study by Parimon et al. found that syndecan-1 was a pro-fibrotic signal and could promotes lung
fibrosis by reprogramming the phenotypes of alveolar type II cells via augmenting TGF-β and Wnt signaling [52]. The findings above suggest that syndecan-1 encoded by SDC1 may be a key regulator involving in liver fibrosis and a potential treatment target. In addition, JunB was found to be one of the most critical transcription factors in NASH-associated aHSC and could regulate the expressions of many signature genes in NASHaHSCsignature (Supplementary table 3). JunB belongs to the JUN transcription factor family, including JunD and c-Jun, and can bind with Fos family and other transcription factors to form AP-1 dimer, which is a key transcription factor regulating cell survival and death pathways [53]. It is also involved in various cellular processes such as proliferation, differentiation, apoptosis, transformation, cell migration, inflammation and wound healing [54]. Increasingly, JunB has also been found to play an important role in fibrogenesis, and JunB activates the TGF-β pathway and promotes COL1A2 deposition [55-57]. In this study, we found that the expression of JunB in primary HSCs was significantly increased after TGF-β stimulation. Therefore, JunB may be involved in the progression of liver fibrosis in NASH patients. The roles of other predicted upstream regulators or transcription factors in liver fibrosis of NASH patients are largely unclear. Further studies are recommended to explore the roles of those predicted upstream regulators or transcription factors in NASH, and determine whether they are potential treatment targets against liver fibrosis in NASH.

Compared with previous literature, our study added novel insights into the fibrosis of NAFLD. We found that the transcriptomic signatures had better performance in assessing the severity of fibrosis in NAFLD patients compared to single gene. Moreover, the findings from a validation study suggested that changes in these HSCs-related signatures were able to show a progressive increase with the progression of liver fibrosis as well as NASH progression in NAFLD patients, suggesting the potential role of those transcriptional signatures in the evaluation and risk stratification for liver fibrosis in NAFLD.

There are some limitations in our study. First, the sample size of scRNA-seq studies is usually small because of the high cost of scRNA-seq, and it is the same
with our study. The findings from this scRNA-seq study of small sample size can be validated by scRNA-seq studies with more samples. Second, for the potential upstream regulators, no causality could be confirmed by bioinformatic outcomes, and further experiments need to performed to provide empirical evidence validating their involvement. Finally, the clinical significance of those transcriptional signatures in the diagnosis or management of NAFLD need be explored with further prospective cohort studies or randomized controlled trials.

In summary, this study defined the NASH-associated transcriptomic changes of activated HSCs, and developed a useful transcriptomic signature with the potential of representing the degree of HSCs activation and fibrosis severity in NASH. This study also identified extracellular matrix-related pathways and immune-related pathways as possible key players in the activation of HSCs during the development of NASH.

**Ethics approval**

The study was approved by the Ethics Committee of the First Affiliated Hospital of Xiamen University (Ethics Number: 2021J011344).

**Conflict of Interest**

The authors declare that they have no conflict of interest.

**Acknowledgement**

We would like to thank all the researchers who have shared their data in GEO database.

**Financial Support**

This study was supported by the National Natural Science Foundation of China (No. 81870606) and the Natural Science Foundation of Fujian Province (2019J01569).

**Abbreviations**

NAFLD: Non-alcoholic fatty liver disease
NASH: nonalcoholic steatohepatitis
MMPs: matrix metalloproteinases
HSCs: hepatic stellate cells
MFBs: myofibroblasts
α-SMA: α-smooth muscle actin
scRNA-seq: single cell RNA-sequencing
aHSCs: activated HSCs
DEGs: differentially expressed genes
GEO: Gene Expression Omnibus
GSVA: Gene Set Variation Analysis
GSEA: Gene set enrichment analysis
WGCNA: Weighted gene co-expression network analysis
ROC: receiver operating characteristic
rHSCs: resting HSCs
NASHrHSCsignature: NASH-associated transcriptomic signature of resting HSC
NASHaHSCsignature: NASH-associated transcriptomic signature of activated HSC
H&E: Hematoxylin and eosin
IHC: Immunohistochemical
IFC: Immunofluorescence

Authors’ contributions
Xuejun Li and Weiwei He conceived of the paper; Weiwei He and Caoxin Huang wrote the original draft; Xiulin Shi, Menghua Wu and Han Li generated the figures; Qiuhong Liu, Xiaofang Zhang and Yan Zhao reviewed and edited the paper. All the authors agreed to the published version of the manuscript.
References

2. Younossi ZM. Non-alcoholic fatty liver disease - A global public health perspective. *J Hepatol* 2019 **70** 531-544.
5. Farrell GC, Larter CZ. Nonalcoholic fatty liver disease: from steatosis to cirrhosis. *Hepatology* 2006 **43** S99-S112.


steatohepatitis. *Endocrine* 2021 **73** 52-64.


Choudhuri S, Garg NJ. Trypanosoma cruzi Induces the PARP1/AP-1 Pathway for
Upregulation of Metalloproteinases and Transforming Growth Factor beta in Macrophages: Role in Cardiac Fibroblast Differentiation and Fibrosis in Chagas Disease. *mBio* 2020 11.

**Figure legends**

**Figure 1** Single cell transcriptomics of liver tissues identified HSCs subsets and the transcriptomic signatures (A, t-SNE plot of cell clusters in liver tissues; B, UMAP plot of cell clusters in liver tissues; C, Comparative analysis cell clusters in liver tissues between NASH and healthy controls.)

**Figure 2** Establishment of NASH-associated transcriptomic signatures of activated HSCs and resting HSCs by single cell RNA-sequencing analysis (A, Heat map showing the expressions of key signature genes in activated HSCs and resting HSCs from NASH patients or healthy controls; B, Violin plots showing the expressions of key signature genes in activated HSCs and resting HSCs from NASH patients or healthy controls.)

**Figure 3** Validation of NASH-associated transcriptomic signatures of activated HSCs and resting HSCs by single cell RNA-sequencing analysis (A, UMAP plot showing comparative analysis of the clusters in activated HSCs and resting HSCs between NASH and healthy controls; B, GSVA revealed that NASHaHSCsignature obviously enriched in the aHSCs cluster of NASH patients, and NASHrHSCsignature obviously enriched in the rHSCs cluster of NASH patients; C, Comparison of the enrichment scores of NASHaHSCsignature and NASHrHSCsignature in each HSCs cluster of NASH patients and healthy controls by violin plots.)

**Figure 4** NASHaHSCsignature was significantly increased in TGF-β-induced
HSCs activation and NASH progression and fibrosis of NAFLD patients (A, GSEA analyses validated the significantly increased enrichment of NASHaHSCsignature in TGF-β-induced HSCs activation; B, Bubble plot showing the enrichment of those signature gene sets during TGF-β-induced HSCs activation; C, Difference in the enrichment scores of the above signature gene sets in each HSC sample calculated by GSVA between TGF-β and control groups; D, Bubble plot showing the enrichment of those signature gene sets in the livers of patients with NASH progression; E, Bubble plot showing the enrichment of those signatures in the livers of NAFLD patients with advanced fibrosis; F, GSEA enrichment plots showing the significantly increased enrichment of NASHaHSCsignature in the livers of patients with NASH progression (left) or advanced fibrosis (right).)

**Figure 5** Validation of the increased enrichment of NASHaHSCsignature in the livers of NAFLD patients with advanced fibrosis (A, Difference in the enrichment scores of NASHaHSCsignature calculated by GSVA between patients with advanced fibrosis and controls; B, ROC analysis revealed enrichment score of NASHaHSCsignature in patients with NAFLD could help to diagnose advanced fibrosis; C, Difference in the enrichment scores of NASHaHSCsignature calculated by GSVA between NAFL patients and NASH patients; D, ROC analysis of GSVA enrichment scores of NASHaHSCsignature in patients with NAFL and NASH.)

**Figure 6** Expression of Collagen 1 was significantly increased in areas of advanced fibrosis in liver tissue of CCl4-induced DBA/2J mice (A. Representative images of liver sections stained by H&E and Masson staining from control and mice of CCl4-induced liver fibrosis (original magnification, ×200); B. Immunohistochemistry of a-SMA and Collagen 1 expression in each group.)

**Figure 7** Expression changing of a-SMA and Collagen 1 in control and mice of
CCl4-induced liver fibrosis by immunofluorescence in DBA/2J mice (A. Representative images of liver sections stained by immunofluorescence from control and mice of CCl4-induced liver fibrosis (original magnification, ×50); B. Representative images of liver sections stained by immunofluorescence from control and mice of CCl4-induced liver fibrosis (original magnification, ×200).)

Figure 8 WGCNA identified crucial functional pathways involved in the increased transition from rHSCs to aHSCs in NAFLD patients (A, Clustering dendrogram showed the co-expression pattern of genes in the liver transcriptome data of NAFLD patients, in which each co-expression gene module was marked with one specific color; B, The heatmap showed the module-trait relationship identified in the WGCNA analysis, in which the coefficient and P values were presented and the transition from green to red indicated the increase in statistical significance; C, The hub genes in Memagenta module were intensively correlated with each other; D, Bubble plot showed those enriched pathways of genes in the MEmagenta co-expression module.)

Supplementary figure 1 Heatmap showed the expression of key genes in different cell clusters of liver tissues.

Supplementary figure 2 Establishment and validation of transcriptomic signatures for aHSCs and rHSCs by single cell RNA-sequencing analysis (A, GSVA revealed that rHSCs transcriptomic signature obviously enriched in the rHSCs cluster, and aHSCs transcriptomic signature obviously enriched in the aHSCs cluster; B, Comparison of the enrichment scores of HSCs transcriptomic signatures in each cell clusters by violin plots.)

Supplementary figure 3 Expression of Collagen 1 was significantly increased in areas of advanced fibrosis in liver tissue of C57/6J mice induced by HFMCD (A. Representative images of liver sections stained by H&E and Masson staining from control and NASH mice (original magnification, ×200); B.
Immunohistochemistry of α-SMA and Collagen 1 expression in each group.

**Supplementary figure 4** Expression changing of α-SMA and Collagen 1 in control and HFMCD-induced NASH mice by immunofluorescence in C57/6J mice (A. Representative images of liver sections stained by immunofluorescence from control and NASH mice (original magnification, ×50); B. Representative images of liver sections stained by immunofluorescence from control and NASH mice (original magnification, ×200).)

**Supplementary figure 5** Enriched pathways of the key co-expression modules in WGCNA of liver transcriptome data (A, Enriched pathways of genes in the MEgreenyellow co-expression module; B, Enriched pathways of genes in the MEturquoise co-expression module.)

**Supplementary figure 6** Upstream regulator and transcription factor enrichment analysis identified key regulators involved in the transcriptomic change of aHSCs in NASH (A, Summary of the results in the upstream regulator analysis; B, Two key transcription factors, JAZF1 and FOBL1, and the corresponding target genes in the transcription factor enrichment analysis.)

**Supplementary figure 7** Expression of JunB was significantly increased in TGF-β-stimulated primary HSCs.

**Supplementary figure 8** GSVA enrichment scores of aHSCsignature, NASHaHSCsignature, rHSCsignature and NASHrHSCsignature between NAFL and NASH patients.

**Supplementary figure 9** GSVA enrichment scores of aHSCsignature, NASHaHSCsignature, rHSCsignature and NASHrHSCsignature between high NAS group (NAS 5-8) and low NAS group (NAS 1-4).
Supplementary figure 10 GSVA enrichment scores of aHSCsignature and NASHaHSCsignature increased with the increase of the severity of NAFLD patients.

Supplementary figure 11 GSVA enrichment scores of aHSCsignature and NASHaHSCsignature among Fibrosis 0, Fibrosis (1-2) and Fibrosis (3-4) group (up). GSVA enrichment scores of aHSCsignature and NASHaHSCsignature between Fibrosis (0-1) and Fibrosis (3-4) group (down).

Supplementary figure 12 GSVA enrichment scores for aHSCsignature and NASHaHSCsignature in all liver fibrosis subgroups in patients with NAFLD (up) and NASH (down).
Table 1 Gene lists of the signature gene sets developed in this study

<table>
<thead>
<tr>
<th>Gene signatures</th>
<th>Gene list</th>
</tr>
</thead>
<tbody>
<tr>
<td>rHSCsignature</td>
<td>RGS5, NDUFA4L2, MYH11, RERGL, COX4I2, MAP3K7CL, RCAN2, NET1, PLN, NOTCH3, LMOD1, SERPIN11, PGF, C2orf40, RASD1, NTRK2, CNN1, KLHL23, TBX2, FOXS1, KCNK17, PPP1R12B, HIGD1B, EFHD1, RERG, NRGN, GPR20, GUCY1B3, RRAD, RBPSM2, LGI4, MRGPRF, MRV11, HRC, CAP2, GRIP2, ACTG2, HEY2, FAM162B, MYOZ1, DMPK, EDNRA, ENPEP, VASN, ITGA7, NT5DC2, CDH6, PRKG1, ARHGEF25, SYNM, RASL12, NTRK3, PDGFA, FAM129A, CLMN, ARPC1A, SMIM10, PCSK7, AOC3, PTK2, MTHFD2, TTL7, UBA2, GPRC5C, PPP1R12A, TMEM51, PHLDA2, HACD1, KCNM1, SPECC1, HOXB2, GRK5, OAT, TLE1, ILK, PACSIN2, CCNI, PPP1R15A, ROCK1</td>
</tr>
<tr>
<td>aHSCsignature</td>
<td>LUM, COL3A1, DPT, PCOLCE, GG5T, ASPN, FBLN1, OLFML3, COL14A1, ITGBL1, COLEC10, ISLR, ADAMTSL2, PDGFRB, COL5A1, SFRP4, OGN, PODN, THBS2, GPC3, LOXL1, STMN2, LAMA2, ABCA8, COL5A2, CH25H, SAMD11, LAMC3, THY1, ADAMTS13, CXCL14, CYGB, ANGPTL6, CTSL, ADAMTS2, CLEC11A, GPC6, AGTR1, RCN3, COL6A3, CBBE1, CTHRC1, CRABP2, CCL9, PDGFL, FGF7, FSTL3, FKBPI0, CERCAM, QSOX1, CCDC146, CYBRD1, CTSF</td>
</tr>
<tr>
<td>NASHrHSCsignature</td>
<td>ID3, MALAT1, CRIP1, IGFBP7, MTRNR2L12, NDUFS5, PHLDA1, PLAC9, CRISPLD2, CSRP2, NOTCH3, CCDC102B, HIGD1B, RASD1, CRIP2, NR4A1, NDUFA4L2, UBN2, MTRNR2L8, ANGPTL4, MT2A, ITM2C, PDGFRB, SNCG, COX4I2, FABP5, ADIRF, COL1A2, CCL2</td>
</tr>
<tr>
<td>NASHaHSCsignature</td>
<td>IFITM1, MIF, COL1A1, S100A11, COL1A2, LGALS1, PDLIM3, LGALS3, SPON2, NMTN, NBEAL1, CTHRC1, VCAN, SERPINF1, COL3A1, SELM, MGP, CLEC11A, PPBP, PARK7, MTRNR2L12, TIMP1, FN1, FHL2, EMP3, COL5A1, IGFBP7, SERPINE1, NBL1, FAM127A, CRABP2, REXO2, DK, MXRA8, LMNA, LRP1, SPARC, COL4A2, INAFM1, SPRY1, VIMP, PRDX4, TNFRSF12A, COX7A1, COL5A2, PHPT1, LNX, HCFC1R1, IGFBP6, PFDN4, PDLIM2, ARID5B, COL6A3, CD81, ISLR, ST3GAL4, PDGFRB, TMEM204, PYCR1, LTBP3, THY1</td>
</tr>
</tbody>
</table>

(Note: The criteria for inclusion of genes in rHSCsignature or aHSCsignature were the log2 of fold change >0.5 and adjusted P value <0.05. The criteria for inclusion of genes in NASHrHSCsignature or NASHaHSCsignature were the log2 of fold change >0.3 and P value <0.05. Genes with significant findings in other relevant cell subsets were not included. Full names of those gene symbols were shown in the supplementary document.)
Single cell transcriptomics of liver tissues identified HSCs subsets and the transcriptomic signatures

331x247mm (300 x 300 DPI)
Establishment of NASH-associated transcriptomic signatures of activated HSCs and resting HSCs by single cell RNA-sequencing analysis

350x238mm (300 x 300 DPI)
Validation of NASH-associated transcriptomic signatures of activated HSCs and resting HSCs by single cell RNA-sequencing analysis

363x163mm (300 x 300 DPI)
NASHaHSCsignature was significantly increased in TGF-β-induced HSCs activation and NASH progression and fibrosis of NAFLD patients

365x280mm (300 x 300 DPI)
Validation of the increased enrichment of NASHaHSCsignature in the livers of NAFLD patients with advanced fibrosis

312x198mm (300 x 300 DPI)
Expression of Collagen 1 was significantly increased in areas of advanced fibrosis in liver tissue of CCl4-induced DBA/2J mice.
Expression changing of α-SMA and Collagen 1 in control and mice of CCl4-induced liver fibrosis by immunofluorescence in DBA/2J mice

480x324mm (300 x 300 DPI)
WGCNA identified crucial functional pathways involved in the increased transition from rHSCs to aHSCs in NAFLD patients

297x341mm (300 x 300 DPI)
Supplementary figures

**Supplementary figure 1** Heatmap showed the expression of key genes in different cell clusters of liver tissues.
Supplementary figure 2 Establishment and validation of transcriptomic signatures for aHSCs and rHSCs by single cell RNA-sequencing analysis (A, GSVA revealed that rHSCs transcriptomic signature obviously enriched in the rHSCs cluster, and aHSCs transcriptomic signature obviously enriched in the aHSCs cluster; B, Comparison of the enrichment scores of HSCs transcriptomic signatures in each cell clusters by violin plots.)
**Supplementary figure 3** Expression of Collagen 1 was significantly increased in areas of advanced fibrosis in liver tissue of C57/6J mice induced by HFMCD (A. Representative images of liver sections stained by H&E and Masson staining from control and NASH mice (original magnification, ×200); B. Immunohistochemistry of α-SMA and Collagen 1 expression in each group.)
Supplementary figure 4 Expression changing of α-SMA and Collagen 1 in control and HFMCD-induced NASH mice by immunofluorescence in C57/6J mice (A. Representative images of liver sections stained by immunofluorescence from control and NASH mice (original magnification, ×50); B. Representative images of liver sections stained by immunofluorescence from control and NASH mice (original magnification, ×200).)
**Supplementary figure 5** Enriched pathways of the key co-expression modules in WGCNA of liver transcriptome data (A, Enriched pathways of genes in the MEgreenyellow co-expression module; B, Enriched pathways of genes in the MEturquoise co-expression module.)
Supplementary figure 6 Upstream regulator and transcription factor enrichment analysis identified key regulators involved in the transcriptomic change of aHSCs in NASH (A, Summary of the results in the upstream regulator analysis; B, Two key transcription factors, JAZF1 and FOBL1, and the corresponding target genes in the transcription factor enrichment analysis.)
Supplementary figure 7 Expression of JunB was significantly increased in TGF-β-stimulated primary HSCs.
Supplementary figure 8 GSVA enrichment scores of aHSCsignature, NASHaHSCsignature, rHSCsignature and NASHrHSCsignature between NAFL and NASH patients.
Supplementary figure 9 GSVA enrichment scores of aHSCsignature, NASHaHSCsignature, rHSCsignature and NASHrHSCsignature between high NAS group (NAS 5-8) and low NAS group (NAS 1-4).
Supplementary figure 10 GSVA enrichment scores of aHSCsignature and NASHaHSCsignature increased with the increase of the severity of NAFLD patients.
Supplementary figure 11 GSVA enrichment scores of aHSCsignature and NASHaHSCsignature among Fibrosis 0, Fibrosis (1-2) and Fibrosis (3-4) group (up). GSVA enrichment scores of aHSCsignature and NASHaHSCsignature between Fibrosis (0-1) and Fibrosis (3-4) group (down).
Supplementary figure 12 GSVA enrichment scores for aHSCsignature and NASHaHSCsignature in all liver fibrosis subgroups in patients with NAFLD (up) and NASH (down).
### Supplementary table 1 Samples involving in single-cell RNAseq of GSE136103

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Type of cells</th>
<th>Cell number *</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM4041151</td>
<td>CD45− cells of healthy control</td>
<td>750 cells</td>
</tr>
<tr>
<td>GSM4041156</td>
<td>CD45− cells of healthy control</td>
<td>3,003 cells</td>
</tr>
<tr>
<td>GSM4041157</td>
<td>CD45− cells of healthy control</td>
<td>1,068 cells</td>
</tr>
<tr>
<td>GSM4041159</td>
<td>CD45− cells of healthy control</td>
<td>3,265 cells</td>
</tr>
<tr>
<td>GSM4041162</td>
<td>CD45− cells of NAFLD patient with fibrosis</td>
<td>1,477 cells</td>
</tr>
<tr>
<td>GSM4041163</td>
<td>CD45− cells of NAFLD patient with fibrosis</td>
<td>1,632 cells</td>
</tr>
</tbody>
</table>

(*Cells with feature genes no less than 800 and the percentage of MT genes less than 15% were included.)*
**Supplementary table 2 Characteristics of 4 publicly available transcriptomic dataset used in the analyses**

<table>
<thead>
<tr>
<th>GSE ID</th>
<th>Participants</th>
<th>Tissues</th>
<th>Methods</th>
<th>Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE49541</td>
<td>32 patients with advanced fibrosis and 40 patients with mild fibrosis</td>
<td>Liver</td>
<td>Microarray</td>
<td>GPL570</td>
</tr>
<tr>
<td>GSE126848</td>
<td>16 NASH patients, 15 NAFL patients, 14 healthy subjects and 12 obese subjects</td>
<td>Liver</td>
<td>RNA-seq</td>
<td>GPL18573</td>
</tr>
<tr>
<td>GSE130970</td>
<td>16 patients with advanced fibrosis and 53 patients with mild fibrosis</td>
<td>Liver</td>
<td>RNA-seq</td>
<td>GPL16791</td>
</tr>
<tr>
<td>GSE148849</td>
<td>6 TGF-β stimulated primary hepatic stellate cells and 6 control primary hepatic stellate cells</td>
<td>Primary hepatic stellate cells</td>
<td>RNA-seq</td>
<td>GPL17021</td>
</tr>
<tr>
<td>GSE135251</td>
<td>51 NAFL patients and 155 NASH patients with different stage of fibrosis</td>
<td>Liver</td>
<td>RNA-seq</td>
<td>GPL18573</td>
</tr>
</tbody>
</table>

(RNA-seq, RNA sequencing)
## Supplementary table 3 Top 20 enriched transcription factor motifs, transcription factors and corresponding target genes

<table>
<thead>
<tr>
<th>Motif id</th>
<th>AUC</th>
<th>NES</th>
<th>Transcription factor</th>
<th>Target genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>yetfasco-2175</td>
<td>0.08</td>
<td>4.98</td>
<td>GATA5,GATA3,GATA2,GATA4,ELK1, GATA1,GATA6,PAX4,NR2F2,NR2F1,NF1,FOX3A,TCF3,JAZF1,SRF,TFPL2,TFPL1,MEF2A,MEF2C,MEF2D,MEF2BNB-MEF2B,MYF2,MYBL2,MYBL1,SNAPC4,ZNF384,ZNF362,KDM4E,KDM4D,KDM4C,KDM4A,KDM4B</td>
<td>LRP1,NR1D1,SERPINE1,PAPPA,LMNA,MMP19,COL1A1,NFIX,CRABP2,PAPN1,GNB2,KLF9</td>
</tr>
<tr>
<td>flyfactorsurvcyk_Jra_SA_ANGER_5_FB_gtn001297</td>
<td>0.07</td>
<td>4.79</td>
<td>FOSL1,FOSB,FOSL2,FOS,JUN,SP1,BA TF,JUND,JUNB,ARNT,NFE2,BACH2,JD P2,SRF</td>
<td>EMP3,LMNA,LGALS3,CALM1,TNFRSF12A,ZFAND2A,PAPPA,GPX1,S100A10,NR1D1,FHL2,MMP19,CRABP2,PFN1,COL1A1,SPHK1,COL3A1</td>
</tr>
<tr>
<td>yetfasco-2167</td>
<td>0.07</td>
<td>4.67</td>
<td>ELK1,GATA5,GATA2,GATA3,GATA6,GATA4,GATA1,NF1,FOX3A,NR2F2,PAX4,NR2F1,TCF3,JAZF1,SRF,TFP,TBP,TFPL2,TFPL1,MEF2A,MEF2C,MEF2D,MEF2BNB-MEF2B,MYF2,MYBL2,MYBL1,SNAPC4,ZNF384,ZNF362,KDM4E,KDM4D,KDM4C,KDM4A,KDM4B</td>
<td>NR1D1,LRP1,PAPPA,SERPINE1,MMP19,NFIX,COL1A1,LMNA,CRABP2,PFN1,KLF9</td>
</tr>
<tr>
<td>transfac_pro-M02658</td>
<td>0.07</td>
<td>4.60</td>
<td>GABPA,ETS1</td>
<td></td>
</tr>
<tr>
<td>yetfasco-2244</td>
<td>0.07</td>
<td>4.58</td>
<td>GATA1,GATA3,ELK1,GATA4,GATA6,GATA2,GATA5,NR2F1,NR2F2,PAX4,F OX3,NF1,TCF3,JAZF1,SRF,TFP,TBP,TFPL2,TFPL1,MEF2A,MEF2C,MEF2D,MEF2BNB-MEF2B,ZNF384,ZNF362,MYF2,MYBL2,MYBL1,SNAPC4,TCF7L1</td>
<td>PAPPA,SERPINE1,LRP1,NR1D1,MMP19,NFIX,PABPN1,COL1A1,KLF9,CRA BP2,GNB2,FN1,CCND1,LMNA,FHL2</td>
</tr>
<tr>
<td>yetfasco-2199</td>
<td>0.07</td>
<td>4.57</td>
<td>GATA1,GATA6,GATA4,GATA3,ELK1,GATA2,GATA5,PAX4,NR2F2,NR2F1,F OX3,NF1,NFE2L1,TFAP4,JAZF1,SRF,TFB,TFPL2,TFPL1,ZNF384,ZNF362,MEF2A,MEF2D,MEF2C,MEF2B,MF2BNB-MEF2B,MYF2,MYBL2,MYBL1,SNAPC4,TCF7L1</td>
<td>NR1D1,LRP1,SERPINE1,MMP19,PAPPACOL1A1,CRABP2,LMNA,FN1,KLF9,NFIX,PABPN1,GNB2</td>
</tr>
<tr>
<td>yetfasco-1781</td>
<td>0.07</td>
<td>4.44</td>
<td>JUN,JUNB,BATF,FOS,FOSL1,FOSL2,JUND,FOSB,BACH2,SP1,ARNT,SRF</td>
<td></td>
</tr>
<tr>
<td>yetfasco-2119</td>
<td>0.07</td>
<td>4.43</td>
<td>GATA5,GATA1,GATA4,GATA3,ELK1,GATA2,GATA6,FOX3A,NF1,NR2F2,PAX4,NR2F1,TCF3,SRF,TFP,TBP,TFPL1,JAZF1,MEF2A,MEF2C,MEF2D,MEF2BNB-MEF2B,MYF2,MYBL2,MYBL1,SNAPC4,ZNF384,ZNF362,MYF2,MYBL1,SNAPC4,TCF7L1</td>
<td>NFIX,LRP1,SERPINE1,MMP19,PAPPA,KLF9,GNB2,COL1A1,LMNA,CCND1,NR1D1,CRABP2,NAA20,FHL2,SERPINE1</td>
</tr>
</tbody>
</table>

Copyright © 2022 the authors
<table>
<thead>
<tr>
<th>ID</th>
<th>Transcription Factors</th>
<th>Regulation</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>transfac_publ-M00497</td>
<td>GATA6, GATA2, ELK1, GATA1, GATA3, GATA5, GATA4, NR2F1, PAX4, NR2F2, NF1, FOXA3, NFE2L1, TFAP4, CBF, RUNX2, SMAD4, JAZF1, SRF, TBP, TBPL1, MEF2A, MEF2C, MEF2D, MEF2BNB, MEF2B, ZNF384, ZNF362, MYBL1, SNAPC4, MYBL2, PAX1, PAX9</td>
<td>STAT3</td>
<td>0.07 4.43</td>
</tr>
<tr>
<td>yetfasco-2163</td>
<td>NR2F1, NR2F2, PAX4, GATA4, GATA2, GATA5, ELK1, GATA1, GATA3, NR2F1, NF1, FOXA3, TCF3, SRF, TBPL1, MEF2A, MEF2C, MEF2D, MEF2BNB, MEF2B, MYBL2, MYBL1, SNAPC4, ZNF384, ZNF362</td>
<td>SRPN1A1, LP1, MMP19, COL1A1, PAPPA, SERPINE1, CRABP2, SERPINE1, CNND1</td>
<td>0.07 4.43</td>
</tr>
<tr>
<td>yetfasco-2187</td>
<td>GATA4, GATA6, ELK1, GATA3, GATA2, GATA5, GATA1, NR2F2, PAX4, NR2F1, STAT1, E2F4, E2F6, NF1, FOXA3, TCF3, SRF, TBPL1, MEF2B, MEF2A, MEF2D, MEF2BNB, MEF2C, MYBL2, MYBL1, SNAPC4, ZNF384, ZNF362</td>
<td>NF1, GNB2, NFIX</td>
<td>0.07 4.38</td>
</tr>
<tr>
<td>yetfasco-2229</td>
<td>GATA1, GATA3, ELK1, GATA6, GATA5, GATA4, GATA2, NR2F2, NR2F1, PAX4, FOXA3, NFE2L1, RUNX2, CBF, TCF3, JAZF1, SRF, TBP, TBPL1, MEF2A, MEF2C, MEF2D, MEF2B, MEF2BNB, ZNF384, ZNF362, MYBL1, SNAPC4</td>
<td>PAPPA, LR1, NR1D1, COL1A1, MMP19, CRABP2, COL1A2, KLF9, F1, GNB2, SERPINE1, CNND1, LMNA</td>
<td>0.07 4.37</td>
</tr>
<tr>
<td>swissregulon-FOS_FOSL2_B_L1__JUN_B-D_p2</td>
<td>JUND, FOS, FOSL1, FOSB, JUNB, FOSL2, BATF, JUN</td>
<td>EMP3, LMNA, TCF7L1, LMNA</td>
<td>0.07 4.36</td>
</tr>
<tr>
<td>hdpi-LUZP1</td>
<td>COL1A1, LP1, GNB2, PAPPA, CRABP2, COL1A2, KLF9, F1, CRABP2, COL1A2, CRABP2, CNND1, LMNA</td>
<td>PAPPA, LR1, NR1D1, COL1A1, MMP19, CRABP2, COL1A2, KLF9, F1, GNB2, SERPINE1, CNND1, LMNA</td>
<td>0.07 4.35</td>
</tr>
<tr>
<td>yetfasco-2200</td>
<td>ELK1, GATA2, GATA3, GATA5, GATA6, GATA4, GATA1, PAX4, NR2F2, NR2F1, TCF3, STAT1, NF1, FOXA3, NEUROD1, E2F4, E2F6, CBF, RUNX2, NFE2L1, TFAP4, SRF, TBPL2, MEF2A, MEF2D, MEF2BNB, MEF2B, ZNF384, ZNF362, MYBL1, SNAPC4, ZNF384, ZNF362, KD M4E, KDM4D, KDM4C, KDM4A, KDM4B</td>
<td>NR1D1, LR1, MMP19, SERPINE1, CNND1, LP1, COL1A1, CRABP2, PAPPA</td>
<td>0.07 4.32</td>
</tr>
<tr>
<td>yetfasco-2171</td>
<td>MEF2A, MEF2BNB, MEF2B, MEF2C, MEF2D, MEF2B, NF1, FOXA3, GATA4, GATA1, GATA3, GATA5, G</td>
<td>LMNA, LP1, SERPINE1, NR1D1, MMP19, CRABP2, COL1A1, GNB2, CRABP2, NFIX, KLF9, PAPPA, NAA20, PABPN1</td>
<td>0.07 4.23</td>
</tr>
<tr>
<td>TransfacPublication</td>
<td>P-value</td>
<td>Fold Change</td>
<td>Gene Names</td>
</tr>
<tr>
<td>---------------------</td>
<td>---------</td>
<td>-------------</td>
<td>------------</td>
</tr>
<tr>
<td>transfac_publ-M00378</td>
<td>0.07</td>
<td>4.21</td>
<td>ATA6, ELK1, GATA2, NR2F2, NR2F1, PAX4, NFIA, NFIC, NEUROD1, TCF3, SRF, TBPL2, TBPL1, JAZF1, MYBL2, MYBL1, MYB, SNACPC4, KDM4E, KDM4D, KDM4C, KDM4B, KDM4A, ZNF384, ZNF362</td>
</tr>
<tr>
<td>yetfasco-2094</td>
<td>0.07</td>
<td>4.20</td>
<td>PAX4, GATA3, GATA6, GATA5, GATA2, GATA4, GATA1, LRP1, CRABP2, GNB2, LMNA, COL1A1, EMP3, KLF9, NR4A1, NR1D1</td>
</tr>
</tbody>
</table>

Supplementary table 4 Correlation between up-regulated expression of the top
**10 genes and severity of fibrosis**

<table>
<thead>
<tr>
<th>aHSCsignature</th>
<th>r</th>
<th>P value</th>
<th>NASHaHSCsignature</th>
<th>r</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUM</td>
<td>0.5</td>
<td>&lt;0.001</td>
<td>IFITM1</td>
<td>-0.06</td>
<td>0.355</td>
</tr>
<tr>
<td>DPT</td>
<td>0.52</td>
<td>&lt;0.001</td>
<td>MIF</td>
<td>-0.24</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>PCOLCE</td>
<td>-0.14</td>
<td>0.039</td>
<td>COL1A1</td>
<td>0.41</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>GGT5</td>
<td>0.15</td>
<td>0.042</td>
<td>S100A11</td>
<td>0.29</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ASPN</td>
<td>0.18</td>
<td>0.011</td>
<td>COL1A2</td>
<td>0.49</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>FBLN1</td>
<td>0.38</td>
<td>&lt;0.001</td>
<td>PDLIM3</td>
<td>0.32</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>OLFML3</td>
<td>0.1</td>
<td>0.149</td>
<td>LGALS3</td>
<td>0.42</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>COL14A1</td>
<td>0.33</td>
<td>&lt;0.001</td>
<td>SPON2</td>
<td>0.23</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ITGBL1</td>
<td>0.66</td>
<td>&lt;0.001</td>
<td>NNMT</td>
<td>0.17</td>
<td>0.012</td>
</tr>
<tr>
<td>ISLR</td>
<td>0.27</td>
<td>&lt;0.001</td>
<td>CTHRC1</td>
<td>0.38</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
**rHSCsignature:** RGS5: regulator of G protein signaling 5; NDUFA4L2: NDUFA4 mitochondrial complex associated like 2; MYH11: myosin heavy chain 11; RERGL: RERG like; COX4I2: cytochrome c oxidase subunit 4I2; MAP3K7CL: MAP3K7 C-terminal like; RCAN2: regulator of calcineurin 2; NET1: neuroepithelial cell transforming 1; PLN: phospholamban; NOTCH3: notch receptor 3; LMOD1: leiomodin 1; SERPINI1: serpin family I member 1; PGF: placental growth factor; C2orf40: ECRG4 augurin precursor; RASD1: ras related dexamethasone induced 1; NTRK2: neurotrophic receptor tyrosine kinase 2; CNN1: calponin 1; KLHL23: kelch like family member 23; TBX2: T-box transcription factor 2; FOXS1: forkhead box S1; KCNK17: potassium two pore domain channel subfamily K member 17; PPP1R12B: protein phosphatase 1 regulatory subunit 12B; HIGD1B: HIG1 hypoxia inducible domain family member 1B; EFHD1: EF-hand domain family member D1; RERG: RAS like estrogen regulated growth inhibitor; NRGN: neurogranin; GPR20: G protein-coupled receptor 20; GUCY1B3: guanylate cyclase 1 soluble subunit beta 1; RRAD: Ras related glycolysis inhibitor and calcium channel regulator; RBPMS2: RNA binding protein, mRNA processing factor 2; LGI4: leucine rich repeat LGI family member 4; MRGPRF: MAS related GPR family member F; HRC: histidine rich calcium binding protein; CAP2: cyclase associated actin cytoskeleton regulatory protein 2; GRIP2: glutamate receptor interacting protein 2; ACTG2: actin gamma 2, smooth muscle; HEY2: hes related family bHLH transcription factor with YRPW motif 2; FAM162B: family with sequence similarity 162 member B; MYOZ1: myozenin 1; DMPK: DM1 protein kinase; EDNRA: endothelin receptor type A; ENPEP: glutamyl aminopeptidase; VASN: vasorin; ITGA7: integrin subunit alpha 7; NT5DC2: 5'-nucleotidase domain containing 2; CDH6: cadherin 6; PRKG1: protein kinase cGMP-dependent 1; ARHGEF25: Rho guanine nucleotide exchange factor 25; SYNM: synemin; RASL12: RAS like family 12; NTRK3: neurotrophic receptor tyrosine kinase 3; PDGFA: platelet derived growth factor subunit A; FAM129A: niban apoptosis regulator 1; CLMN: calmin; ARPC1A: actin related protein 2/3 complex subunit 1A; SMIM10: small integral membrane protein 10; PCSK7: proprotein
convertase subtilisin/kexin type 7; AOC3: amine oxidase copper containing 3; PTK2: protein tyrosine kinase 2; MTHFD2: methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2, methenyltetrahydrofolate cyclohydrolase; TTLL7: tubulin tyrosine ligase like 7; UBA2: ubiquitin like modifier activating enzyme 2; GPRC5C: G protein-coupled receptor class C group 5 member C; PPP1R12A: protein phosphatase 1 regulatory subunit 12A; TMEM51: transmembrane protein 51; PHLDA2: pleckstrin homology like domain family A member 2; HACD1: 3-hydroxyacyl-CoA dehydratase 1; KCNMB1: potassium calcium-activated channel subfamily M regulatory beta subunit 1; PPP1R15A: protein phosphatase 1 regulatory subunit 15A; ROCK1: Rho associated coiled-coil containing protein kinase 1

**aHSCsignature:** LUM: lumican; COL3A1: collagen type III alpha 1 chain; DPT: dermatopontin; PCOLCE: procollagen C-endopeptidase enhancer; GGT5: gamma-glutamyltransferase 5; ASPN: asporin; FBLN1: fibulin 1; OLFML3: olfactomedin like 3; COL14A1: collagen type XIV alpha 1 chain; ITGBL1: integrin subunit beta like 1; COLEC10: collectin subfamily member 10; ISLR: immunoglobulin superfamily containing leucine rich repeat; ADAMTSL2: ADAMTS like 2; PDGFRA: platelet derived growth factor receptor alpha; COL5A1: collagen type V alpha 1 chain; SFRP4: secreted frizzled related protein 4; OGN: osteoglycin; PODN: podocin; THBS2: thrombospondin 2; GPC3: glypican 3; LOXL1: lysyl oxidase like 1; STMN2: stathmin 2; LAMA2: laminin subunit alpha 2; ABCA8: ATP binding cassette subfamily A member 8; COL5A2: collagen type V alpha 2 chain; CH25H: cholesterol 25-hydroxylase; SAMD11: sterile alpha motif domain containing 11; LAMC3: laminin subunit gamma 3; THY1: Thy-1 cell surface antigen; ADAMTS13: ADAM
metallopeptidase with thrombospondin type 1 motif 13; CXCL14: C-X-C motif chemokine ligand 14; CYGB: cytoglobin; ANGPTL6: angiopoietin like 6; CTSK: cathepsin K; ADAMTS2: ADAM metallopeptidase with thrombospondin type 1 motif 2; CLEC11A: C-type lectin domain containing 11A; GPC6: glycan 6; AGTR1: angiotensin II receptor type 1; RCN3: reticulocalbin 3; COL6A3: collagen type VI alpha 3 chain; CCBE1: collagen and calcium binding EGF domains 1; CTHRC1: collagen triple helix repeat containing 1; CRABP2: cellular retinoic acid binding protein 2; CCL19: C-C motif chemokine ligand 19; PDLIM4: PDZ and LIM domain 4; LAMB1: laminin subunit beta 1; FGF7: fibroblast growth factor 7; FSTL3: follistatin like 3; FKBPT10: FKBP prolyl isomerase 10; CERCAM: cerebral endothelial cell adhesion molecule; QSOX1: quiescin sulphydryl oxidase 1; CCDC146: coiled-coil domain containing 146; CYBRD1: cytochrome b reductase 1; CTSF: cathepsin F

NASHrHScSignature: ID3: inhibitor of DNA binding 3; MALAT1: metastasis associated lung adenocarcinoma transcript 1; CRIP1: cysteine rich protein 1; IGFBP7: insulin like growth factor binding protein 7; MTRNR2L12: MT-RNR2 like 12; NDUFS5: NADH:ubiquinone oxidoreductase subunit S5; PHLDA1: pleckstrin homology like domain family A member 1; PLAC9: placenta associated 9; CRISPLD2: cysteine rich secretory protein LCCL domain containing 2; CSRP2: cysteine and glycine rich protein 2; NOTCH3: notch receptor 3; CCDC102B: coiled-coil domain containing 102B; HIGD1B: HIG1 hypoxia inducible domain family member 1B; RASD1: ras related dexamethasone induced 1; CRIP2: cysteine rich protein 2; NR4A1P: nuclear receptor subfamily 4 group A member 1; NDUFA4L2: NDUFA4 mitochondrial complex associated like 2; UBN2: ubinuclein 2; MTRNR2L8: MT-RNR2 like 8; ANGPTL4: angiopoietin like 4; MT2A: metallothionein 2A; ITM2C: integral membrane protein 2C; PDGFRB: platelet derived growth factor receptor beta; SNCG: synuclein gamma; COX4I2: cytochrome c oxidase subunit 4I2; FABP5: fatty acid binding protein 5; ADIRF: adipogenesis regulatory factor; COL1A2:
collagen type I alpha 2 chain; CCL2: C-C motif chemokine ligand 2

**NASHaHSCsignature:** IFITM1: interferon induced transmembrane protein 1; MIF: macrophage migration inhibitory factor; COL1A1: collagen type I alpha 1 chain; S100A11: S100 calcium binding protein A11; COL1A2: collagen type I alpha 2 chain; LGALS1: galectin 1; PDLIM3: PDZ and LIM domain 3; LGALS3: galectin 3; SPON2: spondin 2; NNMT: nicotinamide N-methyltransferase; NBEAL1: neurobeachin like 1; CTHRC1: collagen triple helix repeat containing 1; VCAN: versican; SERPINF1: serpin family F member 1; COL3A1: collagen type III alpha 1 chain; SELM: selenoprotein M; MGP: matrix Gla protein; CLEC11A: C-type lectin domain containing 11A; PPIB: peptidylprolyl isomerase B; PARK7: Parkinsonism associated deglycase; MTRNR2L12: MT-RNR2 like 12; TIMP1: TIMP metallopeptidase inhibitor 1; FN1: fibronectin 1; FHL2: four and a half LIM domains 2; EMP3: epithelial membrane protein 3; COL5A1: collagen type V alpha 1 chain; IGFBP7: insulin like growth factor binding protein 7; SERPINE1: serpin family E member 1; NBL1: NBL1, DAN family BMP antagonist; FAM127A: retrotransposon Gag like 8C; CRABP2: cellular retinoic acid binding protein 2; REXO2: RNA exonuclease 2; MDK: midkine; MXRA8: matrix remodeling associated 8; LMNA: lamin A/C; LRP1: LDL receptor related protein 1; SPARC: secreted protein acidic and cysteine rich; COL4A2: collagen type IV alpha 2 chain; INAFM1: InaF motif containing 1; SPRY1: sprouty RTK signaling antagonist 1; VIMP: selenoprotein S; PRDX4: peroxiredoxin 4; TNFRSF12A: TNF receptor superfamily member 12A; COX7A1: cytochrome c oxidase subunit 7A1; COL5A2: collagen type V alpha 2 chain; PHPT1: phosphohistidine phosphatase 1; LXN: latexin; HCFC1R1: host cell factor C1 regulator 1; IGFBP6: insulin like growth factor binding protein 6; PFDN4: prefoldin subunit 4; PDLIM2: PDZ and LIM domain 2; ARID5B: AT-rich interaction domain 5B; COL6A3: collagen type VI alpha 3 chain; CD81: CD81 molecule; ISLR: immunoglobulin superfamily containing leucine rich repeat; ST3GAL4: ST3 beta-galactoside alpha-2,3-sialyltransferase 4; PDGFRA: platelet derived growth factor receptor alpha; TMEM204: transmembrane
protein 204; PYCR1: pyrroline-5-carboxylate reductase 1; LTBP3: latent transforming growth factor beta binding protein 3; THY1: Thy-1 cell surface antigen