Transcriptomic profiling across the nonalcoholic fatty liver disease spectrum reveals gene signatures for steatohepatitis and fibrosis

TRANSCRIPTOMIC PROFILING ACROSS THE NON-ALCOHOLIC FATTY LIVER DISEASE SPECTRUM REVEALS GENE SIGNATURES FOR STEATOHEPATITIS AND FIBROSIS

Olivier Govaere[1], Simon Cockell[2], Dina Tiniakos[1,3], Rachel Queen[2], Ramy Younes[1,4], Michele Vacca[5], Leigh Alexander[6], Federico Ravaioli[1,7], Jeremy Palmer[1], Salvatore Petta[8], Jerome Boursier[9], Chiara Rosso[4], Katherine Johnson[1], Kristy Wonders[1], Christopher P. Day[1], Mattias Ekstedt[10], Matej Orešič[11], Rebecca Darlay[12], Heather J. Cordell[12], Fabio Marra[13], Antonio Vidal-Puig[5], Pierre Bedossa[1,14], Jörn M. Schattenberg[15], Karine Clément[16], Michael Allison[17], Elisabetta Bugianesi[4], Vlad Ratziu[14], Ann K. Daly[1*], Quentin M. Anstee[1,18*] *on behalf of the EPoS & LITMUS Investigators*[†]

Affiliations:

[1] Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom.

[2] Bioinformatics Support Unit, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom.

[3] Dept of Pathology, Aretaieio Hospital, National & Kapodistrian University of Athens, Athens, Greece.

[4] Department of Medical Sciences, Division of Gastro-Hepatology, A.O. Città della Salute e della Scienza di Torino, University of Turin, Turin, Italy.

[5] University of Cambridge Metabolic Research Laboratories, Wellcome-MRC Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, United Kingdom.

[6] SomaLogic, Inc., Boulder, CO, USA.

[7] Department of Medical and Surgical Sciences, University of Bologna, Italy.

[8] Sezione di Gastroenterologia, Dipartimento Biomedico di Medicina Interna e Specialistica, Università di Palermo, Palermo, Italy.

[9] Hepatology Department, Angers University Hospital, Angers, France.

[10] Department of Health, Medicine and Caring Sciences, Linköping University, Linköping, Sweden.

[11] School of Medical Sciences, Örebro University, 702 81 Örebro, Sweden.

[12] Population Health Sciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom.

[13] Dipartimento di Medicina Sperimentale e Clinica, University of Florence, Florence, Italy.

[14] Assistance Publique-Hôpitaux de Paris, hôpital Pitié Salpêtrière, Sorbonne University, ICAN (Institute of Cardiometabolism and Nutrition), Paris, France.

[15] I. Department of Medicine, University Hospital Mainz, Mainz, Germany.

[16] Nutrition and obesity: systemic approaches, Inserm, Sorbonne University, Paris, France.

[17] Liver Unit, Department of Medicine, Cambridge NIHR Biomedical Research Centre, Cambridge University NHS Foundation Trust, United Kingdom.

[18] Newcastle NIHR Biomedical Research Centre, Newcastle upon Tyne Hospitals NHS Trust, Newcastle upon Tyne, United Kingdom.

*Correspondence:

Prof Ann K. Daly PhD

Professor of Pharmacogenetics,

Translational and Clinical Research Institute,

The Medical School, Newcastle University,

4th Floor, William Leech Building,

Framlington Place,

Newcastle-upon-Tyne, NE2 4HH,

United Kingdom.

Telephone: + 44 (0) 191 208 7031

Email: a.k.daly@ncl.ac.uk

Prof Quentin M. Anstee PhD, FRCP

Professor of Experimental Hepatology & Honorary Consultant Hepatologist,

Translational and Clinical Research Institute,

The Medical School, Newcastle University,

4th Floor, William Leech Building,

Framlington Place,

Newcastle-upon-Tyne, NE2 4HH,

United Kingdom.

Telephone: + 44 (0) 191 208 7012

Email: quentin.anstee@ncl.ac.uk

One Sentence Summary:

A detailed transcriptomic analysis across the NAFLD spectrum elucidates pathophysiological processes and identifies tractable serum biomarkers.

Abstract:

The mechanisms that drive non-alcoholic fatty liver disease (NAFLD) remain incompletely understood. This large multicenter study characterized the transcriptional changes that occur in liver tissue across the NAFLD spectrum as disease progresses to cirrhosis, using these data to identify novel circulating biomarkers. A discovery cohort comprising 206 histologically characterized NAFLD samples underwent high-throughput RNA sequencing. Unsupervised clustering stratified NAFLD based on disease activity and fibrosis stage with differences in age, ALT, type 2 diabetes mellitus and carriage of the PNPLA3 *rs738409* variant. Relative to early disease, 25 differentially expressed genes were consistently identified as fibrosing-steatohepatitis progressed through stages F2-4. These findings were independently validated by logistic modelling as robust indicators of disease stage in a separate replication cohort (n=175) and an integrative analysis with single-cell RNA sequencing data elucidated the relative contribution of specific intrahepatic cell populations. Translating these findings to the protein-level, SomaScan™ analysis in over 300 NAFLD serum samples confirmed that circulating concentrations of the encoded proteins (AKR1B10 and GDF15) were strongly associated with disease activity and fibrosis stage: serving as putative biomarkers for fibrosing-steatohepatitis. Supporting the biological plausibility of these data, *in vitro* functional studies determined that endoplasmic reticulum stress upregulated expression of genes *AKR1B10*, *GDF15* and *PDGFA*, whereas GDF15 supplementation tempered the inflammatory response in macrophages upon lipid loading and lipopolysaccharide stimulation. This detailed study provides novel insights into the pathophysiology of progressive fibrosing-steatohepatitis, as well as proof of principle that transcriptomic changes represent potentially tractable and clinically relevant disease biomarkers.

Introduction

Non-alcoholic fatty liver disease (NAFLD) is an increasingly common progressive disease characterized by excessive hepatic accumulation of triglyceride/reactive lipid species and is strongly associated with the metabolic syndrome, i.e. central obesity, type 2 diabetes mellitus (T2DM), hypertension and dyslipidemia (*1*). With a global increase in sedentary behavior and obesity, the prevalence of NAFLD is rising rapidly, and now affects approximately 25% of the adult population worldwide (*1*). NAFLD is subdivided into 'simple' steatosis (non-alcoholic fatty liver, NAFL), and non-alcoholic steatohepatitis (NASH), defined by the presence of necro-inflammation and hepatocyte ballooning. If NASH persists, fibrosis occurs and may progress to cirrhosis and, ultimately, end-stage liver disease (*2*). NAFL was traditionally considered a stable and relatively benign disease state that lacked the capacity to progress. However, recent data from serial biopsy studies have demonstrated that NAFL may transit into NASH and onwards to advanced fibrosis (*3*). NAFLD is therefore best considered a dynamic disease with steatohepatitic activity waxing and waning, and fibrosis stage also progressing and regressing subject to the actions of a variety of genetic, epigenetic and environmental modifiers (*2*). Key management challenges for NAFLD include the lack of effective biomarkers that may be used to risk stratify patients and the lack of approved pharmacological therapies.

To date, efforts to examine the transcriptomic changes that occur as NAFLD progresses have largely employed microarray-based techniques and so have lacked the comprehensive approach provided by global RNA sequencing (*4-10*), or have been confined to relatively small patient cohorts that do not adequately represent the full spectrum of disease from normal liver through NAFL, to NASH exhibiting progressive stages of fibrosis and cirrhosis (*11-13*). Consequently, they have been limited to dichotomous comparisons between mild and advanced disease, which do not provide the granularity needed to fully appreciate the complex transcriptomic changes as NAFLD evolves, and lack independent validation.

We report a comprehensive transcriptomic analysis conducted using RNA sequencing technology across the full histological spectrum from healthy controls through to NASH-associated cirrhosis in a large cohort of European patients. Adopting an integrative transcriptomic approach to unravel pathways responsible for the stepwise progression of NAFLD, we identified and validate the gene expression signatures associated with early stages of disease, subsequent progression and specific histological features. Translating our findings from the hepatic transcriptome to the protein level, we further validate selected gene expression changes associated with disease activity and fibrosis stage using immunohistochemistry and measurement of circulating protein as exemplar biomarkers.

Results

Unsupervised clustering stratifies NAFLD based on fibrosis and disease activity

The current study incorporates transcriptomic data on 403 individuals, 381 NAFLD samples and 22 control samples, representing the full histological range from normal liver tissue to NASH-cirrhosis. 206 frozen tissue samples from patients with NAFLD were included in a discovery

cohort and processed for high-throughput RNA sequencing (Fig. 1). All samples were histologically scored by two expert liver pathologists (DT & PB) according to the widely accepted, semi-quantitative, NASH-CRN NASH Activity Score (NAS) and the FLIP steatosis (S), activity (A), and fibrosis (F) scoring systems (*14, 15*) and grouped according to histopathological disease grade and stage, i.e. NAFL, NASH with different fibrosis stages F0, -F1, F2, F3 and F4. Detailed phenotypic description and demographics are reported in Table 1. As indicated by the principal component analysis, potential confounding factors sex, batch and center were corrected for in the analysis (Fig. S1).

Unsupervised clustering based on gene expression stratified the 206 NAFLD samples from the discovery cohort into two distinct groups, annotated as cluster A and B (Fig. 2). Histologically, cluster A was characterized by more advanced fibrosis (Mann-Whitney U test, p=5.15E-10), a higher grade of hepatic ballooning (p=1.67E-05) and lobular inflammation (Kleiner scoring 0-3, p=1.70E-03; SAF scoring 0-2, p=1.36E-04), with no differences in steatosis grade when compared to cluster B (p=4.78E-01) (Table 1). This translated into a higher number of patients diagnosed with NASH in cluster A (84.62%) compared to cluster B (69.5%; Chi-square p=2.11E-02). Moreover, when stratifying based on a high disease activity using a NAS score ≥ 4 (sum of steatosis, ballooning and Kleiner inflammation) or a SAF activity score ≥ 2 (sum of ballooning and SAF inflammation), cluster A showed an enrichment compared to cluster B (NAS≥ 4 83.08% vs 66.67%, Chi-square p= 1.49E-02; SAF activity score ≥ 2 86.15% vs 68.79%, p=8.08E-03) (Table 1). Compared to cluster B, patients in cluster A were slightly older (57.08 vs 52.57 years; Mann-Whitney U test p=1.58E-02), more overweight (BMI 32.51 vs 30.8; Mann-Whitney U test p=1.99E-02), were more likely to have T2DM (70.77 vs 45.39%; Chi-square p= 6.90E-04), and exhibited higher HbA1c concentrations (52.16 vs 45.8 mmol/mol; Mann-Whitney U test p= 4.17E-02), higher serum AST (53.23 vs 40.67 u/L; Mann-Whitney U test p= 4.62E-04) and reduced platelet count (212.53 vs 237.78 x 10^9, Mann-Whitney U test p= 2.05E-02) (Table 1). Although carriage of well described genetic variants that have been associated with NAFLD severity (*16*) (*GCKR* rs1260326, *HSD17B13* rs72613567, *PNPLA3* rs738409 and *TM6SF2* rs58542926) did not confer discrete gene expression profiles within the overall RNAseq dataset (Fig. S2), cluster A showed an increase in carriers for the *PNPLA3* rs738409 polymorphism when compared to cluster B (Chi-square p= 3.06E-02).

Between cluster A and B, 1,292 differentially expressed genes (DEGs) were found (Table S1). KEGG pathway analysis indicated an enrichment in genes correlating to pathways including 'Extracellular Matrix interaction', 'Focal adhesion', 'PI3k-Akt signaling' and 'Wnt signaling'(Fig.S3A-C). Moreover, cluster A showed an increased expression in cytokine genes such as *CCL2*, *CCL20*, *CCL19* and *CCL28* (Fig. S3B) and also in hepatic progenitor cell marker genes (*TACSD2*/*TROP2*, *EPCAM*, *SOX9*, *KRT19*, *KRT7*, *CD24*, *JAG1*), suggesting progenitor cell-mediated regeneration. Taken together, unsupervised clustering stratified NAFLD not only based on fibrosis stage but also on disease activity and carriage of the *PNPLA3* rs738409 variant, transcriptionally defining a patient subgroup that shows high Wnt signaling, active tissue remodeling and progenitor cell-mediated regeneration.

Nested within cluster B are further subgroups that are not accounted for by strong differences in histological features alone but are rather characterized by changes in gene expression associated with tissue remodeling and PI3K-Akt signaling pathway including mTOR signaling-

related genes (e.g. *RICTOR*, *LAMTOR4*), and differences in nuclear factors including *SP3*, *NR2C2* and *LCOR* (Table S2 and Fig. S3D). This suggests that NAFLD is a more heterogeneous condition than has previously been clinically defined.

Transcriptional changes apparent in NAFL are sustained as fibrosing-steatohepatitis progresses

To further understand the pathogenesis and progression of NAFLD, we performed supervised clustering comparing the different stages of NAFLD with control samples. Principal component analysis showed a distinct separation between the NAFLD cases (n=206) and 'healthy obese' controls (histologically normal liver tissue samples obtained from obese patients, n=10) (Fig. S4A). Similarly, using publicly available control RNAseq data sets, a gradual shift was observed from healthy non-obese individuals to healthy obese patients and to NAFLD patients (Fig. S4B) (*12*). A total of 2,603 DEGs were identified between NAFLD and the control samples, and between 1,875 - 3,578 DEGs when comparing the individual stages, showing enrichment for pathways including 'Metabolic pathways', 'Ribosome' and 'TNF signaling'(q-value<0.05, Fig. S4C) with certain pathways such as 'Platelet activation'-, 'PPAR signaling'- and 'Extracellular matrix'-related genes only exhibiting enrichment at later stages of disease progression, i.e. NASH F3/4 (Fig. S5A).

Considering *a priori* pathophysiologically relevant candidates (*17*), we found expression of several genes to be increased in early disease, i.e. NAFL, including the cellular senescence marker *CDKN1A/p21,* the inflammatory marker *IKBKG/NEMO* and *CYP7A1*, the rate-limiting enzyme in the classical bile acid synthesis pathway that is subject to FXR-mediated regulation (*18*), which peaked as NASH developed and remained elevated but became less so as fibrosis progressed (Fig. S5B). This shows that transcriptomic changes related to initiation of inflammation, cellular senescence and bile secretion begin to occur soon after NAFL inception, and prior to histologically evident steatohepatitis.

Defining differentially expressed gene-sets associated with steatohepatitis and fibrosis

In order to study mechanisms associated with disease progression, we performed pairwise analyses between disease phenotype categories using the RNAseq data from the 206 NAFLD patients in the discovery cohort. To identify modifiers of steatohepatitis we compared the different NASH stages with NAFL, whereas to discover modifiers of fibrosis we looked at different fibrosis stages within NASH. Taking NAFL as a baseline, no statistically significant DEGs were identified compared with NASH F0/1, whilst 50 DEGs, 907 DEGs and 1,369 DEGs were observed when comparing NAFL with NASH F2, NASH F3 and NASH F4 respectively (Fig. 3A and Table S3-5). Similarly, when using NASH F0/1 as a baseline, no genes were differentially expressed compared with NASH F2, whereas the comparison with NASH F3 identified 434 DEGs and 1,194 DEGs when comparing with NASH F4, with 393 DEGs in common between these two stages (Fig.3A and Table S6-7). GO annotation analyses of the different comparisons are described in Fig. S6.

To identify a core gene-set 'signature' associated with the progression from NAFL or NASH F0/1 to more advanced disease, we focused on DEG commonality between the different pairwise comparisons (Fig. 3A). The two intersections using either NAFL or NASH F0/1 as a baseline shared 25 DEGs, with 24 of them showing a gradual increase with disease progression (i.e. *AKR1B10*, *ANKRD29*, *CCL20*, *CFAP221*, *CLIC6*, *COL1A1*, *COL1A2*, *DTNA*, *DUSP8*, *EPB41L4A*, *FERMT1*, *GDF15*, *HECW1*, *IL32*, *ITGBL1*, *LTBP2*, *PDGFA*, *PPAPDC1A*, *RGS4*, *SCTR*, *STMN2*, *THY1*, *TNFRSF12A*, *TYMS*) and one gene showing a gradual decrease (i.e. *HSD17B14*) (Fig. 3B). These 25 genes were all differentially expressed when comparing NASH F2-4 to NAFL+NASH F0/1 combined and all apart from *TYMS* (fold change below 1.5) were differentially expressed between unsupervised cluster A and cluster B (Fig. 3C and 3D). Supporting this, when we used the widely accepted histological thresholds for high-probability of NASH (NAS≥4 or SAF activity ≥ 2) to stratify the discovery cohort, we identified 369 and 320 DEGs respectively (Table S8 and S9). All 25 genes were present within the DEGs when the cohort was stratified purely by histological activity (inflammation and hepatocyte ballooning) with SAF activity ≥ 2. However, *ANKRD29*, *GDF15* and *TYMS* were omitted (and so 22 of the 25 genes were captured) when using the NAS≥4 threshold, which conflates degree of steatosis and grade of histological activity into a single index.

Findings from the discovery analysis were replicated using nanoString® analysis in an independent cohort of 175 NAFLD liver biopsies (Table S10). The breakdown of relative expression in the different disease stages compared to the controls is presented in Fig. S7. When comparing NAFL and NASH F0/1 vs. NASH F2-4, we found 21 of the initial 25 gene-set to be significantly differentially expressed (i.e. *AKR1B10*, *CCL20*, *CFAP221*, *CLIC6*, *COL1A1*, *COL1A2*, *DTNA*, *DUSP8*, *FERMT1*, *GDF15*, *HECW1*, *IL32*, *ITGBL1*, *LTBP2*, *PDGFA*, *PPAPDC1A*, *RGS4*, *SCTR*, *STMN2*, *THY1* and *TNFRSF12A*) at p<0.001 (Fig. S7). Taken together, these data identified a consistently differentially expressed gene signature of more advanced disease, associated with disease activity and correlating to the unsupervised clustering.

A 25 gene-set of differentially expressed genes independently predicts features of NAFLD

To investigate further the relationship between the changes in gene expression and the components of the histological phenotype, and to dissect apart the collinearity between those features, we performed regression analysis using the RNAseq data to generate gene signatures linearly associated with the severity of each specific histological feature (steatosis, inflammation, hepatocyte ballooning and fibrosis) (Fig. 4A). Hierarchal clustering of genes significantly (adj p<0.05) associated with histological criteria revealed a stronger overlap between inflammation, ballooning and fibrosis than with steatosis (Fig.4B). Each of the set of 25 genes correlated strongly with increasing severity of the histological components inflammation, ballooning and fibrosis, with 12 genes showing additional overlap with steatosis (*AKR1B10*, *CCL20*, *COL1A1*, *COL1A2*, *DTNA*, *DUSP8*, *GDF15, PDGFA*, *PPAPDC1A*, *STMN2*, *THY1* and *TNFRSF12A*) (Fig. 4A and Table S11).

To understand the effect of the individual components of the 25 DEG gene-set on histological features, we evaluated the predictive value in the NAFLD discovery cohort (n=206) using univariate and multivariate logistic regression analysis. Table S12 summarizes the results from

the univariate analyses. In the multivariate models predicting steatohepatitis grade, *HSD17B14* (OR=0.5049, p =0.001), *PPAPDC1A* (OR=3.2079, p <0.0001), *SCTR* (OR=0.6891, p<0.05) and *TNFRSF12A* (OR=2.7367, p<0.0001) expression were independent factors significantly predictive of NAS≥4 with an AUROC of 0.85; AST level (OR= 1.1079, p<0.01) together with *EPB41L4A* (OR=2.6634, p<0.01), *GDF15* (OR=1.0253, p=0.001) and *ITGBL1* (OR=2.022, p<0.01) expression which independently predicted a SAF activity score ≥ 2 with an AUROC of 0.86 (Fig. 4C-D and Table S13). The presence of advanced fibrosis F3-4 was independently predicted by T2DM (OR=2.7481, p<0.05) combined with the expression of *ANKRD29* (OR=2.5776, p<0.01), *CLIC6* (OR=2.2211, p<0.0001) and *STMN2* (OR=1.6397, p =0.001) with an AUROC of 0.93 (Fig. 4E and Table S13). This combined model to predict fibrosis had a significantly higher AUROC than the FIB4 score (*19*) (FIB4 AUROC 0.73, DeLong test p<0.00001, Fig. 4F). In addition, the presence of fibrosing steatohepatitis, defined as NAS≥4 and a fibrosis stage ≥F2 (NASH+NAS≥4+F≥2) was predicted by the expression of *EPB41L4A*, *HSD17B14*, *PPAPDC1A* and *TNFRSF12A* (Table S13). The independent variables predicting the presence of lobular inflammation, hepatocyte ballooning and portal inflammation are summarized in Table S13. The multivariate models established using the discovery RNAseq data were tested in the replication cohort (n=175). Similar AUROCs were found in both cohorts with no significant differences (Hosmer-Lemeshow chi^2 test) for the presence of lobular inflammation, ballooning, portal inflammation, NAS ≥ 4, SAF activity score ≥ 2, advanced fibrosis and NASH+NAS≥4+F≥2 (Table S14).

Integrated single-cell RNA sequencing analysis identifies cell clusters based on the 25-gene signature

To explore how the 25 DEG gene-set changes within specific cell populations during NAFLD progression, we performed an integrated single-cell RNA sequencing (scRNAseq) analysis by projecting publicly available scRNAseq data from healthy and cirrhotic liver disease onto the RNAseq data grouped by fibrosis stage (discovery cohort n=206) (*20*). Epithelial cells, macrophages, mesenchymal cells, endothelial/sinusoidal cells and lymphocytes were identified within the scRNAseq data based on the expression of lineage specific markers as annotated in Table S15 and Fig. S8. Uniform Manifold Approximation and Projection (UMAP) plots with a resolution of 0.7 identified eight distinct clusters within the epithelial cells from the healthy and cirrhotic scRNAseq data combined (Fig. 5A and S9). Looking at the expression of our 25-gene signature, one specific cluster within cirrhosis showed concurrent expression of ten markers including *AKR1B10*, *ANKRD29*, *CLIC6*, *DTNA*, *GDF15*, *IL32*, *PDGFA*, *RGS4*, *SCTR* and *TNFRSF12A* (epithelial cluster 5, Fig. 5A). Projecting the signatures of the different epithelial cell populations onto the NAFLD RNAseq data using the CIBERSORT analytical tool, the epithelial cluster capturing ten of our 25-gene signature showed a high expression in advanced NASH with the strongest enrichment in NASH F4 (Fig. 5B).

When assessing the different macrophage populations, *HSD17B14* expression was observed in one cluster, annotated as macrophage cluster 5, present in both healthy and end-stage liver disease (Fig. 5C). To further characterize this cluster, we looked at the expression of *CCR2*, *CD163* and *TREM2*, markers for macrophages that have been reported to be relevant in fibrosis and NAFLD (*20-23*). The *HSD17B14*+ cluster showed expression for *CD163* but was negative for

CCR2 and *TREM2*, and was enriched in NAFL within the NAFLD RNAseq data (Fig. 5D). *TREM2* expression was observed in two clusters within the macrophage populations which also showed positivity for *CD163*. The first cluster, macrophage cluster 0, was enriched in NASH F2, F3 and F4, while the second cluster, macrophage cluster 6, was mainly expressed in NASH F3 and F4 (Fig. 5C and 5D). Moreover, the *CD163*+/*TREM2*- macrophage population identified only within the cirrhotic scRNAseq data (macrophage cluster 7) showed an enrichment in NASH F3 and NASH F4 stages (Fig. 5D). *CCR2* expression was only observed in macrophage cluster 8, although this signature was not enriched within the NAFLD RNAseq data (Fig. 5D). Additionally, when looking at the expression of *COL1A1*, *COL1A2*, *ITGBL1*, *LTBP2*, *PDGFA*, *PPAPDC1A*, *RGS4*, *STMN2* and *TNFRSF12A* different mesenchymal populations could be identified which showed distinct enrichment depending on the stage of the NAFLD disease (Fig. S9). Taken together, integrated scRNAseq analysis identified distinct populations based on the expression of our 25-gene signature and suggests that dynamic changes within intrahepatic different macrophage populations occur during NAFLD progression.

Validation of hepatic gene expression in liver tissue by immunohistochemistry

To validate the findings from the integrated RNAseq/scRNAseq analysis, three markers were selected based on availability of antibodies for immunohistochemistry staining of 33 formalin-fixed paraffin-embedded (FFPE) liver tissue cases derived from the NAFLD replication cohort (i.e. AKR1B10, GDF15 and STMN2). AKR1B10 protein expression was seen focally in hepatocytes in NAFL, showing a cytoplasmic and nuclear immunostaining pattern (Fig. S10). In NASH, AKR1B10 positivity was more prominent in ballooned hepatocytes and in hepatocytes located neighboring necro-inflammatory foci and periportal/periseptal areas. Additional, weaker immunostaining of sinusoidal lining cells was seen in the majority of the NAFLD cases. The number of AKR1B10 immunopositive hepatocytes increased with disease stage, peaking at F4. GDF15 staining showed a granular cytoplasmic positivity in the hepatocytes of the NAFLD samples (Fig. S10). Moreover, GDF15 expression was focally observed in parenchymal immune cells. Hepatocyte immune-positivity for GDF15 was non-zonal. In addition, STMN2 immunopositivity was seen in macrophages in the portal inflammatory infiltrate with an increasing number of positive cells towards end-stage cirrhosis. Weak STMN2 expression was observed in sinusoidal lining cells in all the stages of NAFLD (Fig. S10).

Serum AKR1B10 and GDF15 correlate with disease stage and stratify patients based on activity

To determine whether evidence of the hepatic transcriptomic changes could also be detected peripherally, we assessed whether circulating protein concentrations of the 25 DEG gene-set accurately reflected histological disease severity as an exemplar for future potential biomarker development. Proteomics analysis was performed on 305 serum samples from patients with histologically characterized NAFLD using SomaScan® technology. 13 proteins were detectable in the serum samples, reflecting 14 out of the 25 genes, with COL1A being the protein for the genes *COL1A1* and *COL1A2*. Assessing the different histological scores, AKR1B10 was the only circulating protein showing a significant increment reflecting the increase in steatosis grade

(Kruskal-Wallis test, p=9.66E-06; Fig. 6 and Table S16). Serum AKR1B10, COL1A and GDF15 concentrations showed significant differences with an increase in the score for ballooning (Kruskal-Wallis test p=3.80E-14, p= 1.02E-02, p= 2.05E-05 respectively) whilst both AKR1B10 and GDF15 were associated with high Kleiner and SAF inflammation scores. Conversely, HECW1 showed a gradual decrease (Kruskal-Wallis test, p<0.05) (Fig.6 and Table S16). Serum AKR1B10 and GDF15 were also significantly increased with the rise in fibrosis stage (Kruskal-Wallis test, p= 6.22E-13 and p= 3.44E-15 respectively), with AKR1B10 showing a 2.19 fold increase when comparing fibrosis stage F4 with F0 and GDF15 a 2.75 fold increase (post-hoc corrected p<0.001; Fig. 6). Furthermore, serum AKR1B10 and GDF15 were significantly increased in patients with NASH (1.9 and 1.35 fold change respectively), in patients with NAS≥4 (which conflates steatosis with activity, 2.14 and 1.35 fold change respectively) and in patients with SAF activity ≥2 (activity of steatohepatitis, 1.91 and 1.38 fold change respectively) (Mann-Whitney p<0.0001), while COL1A only captured NASH and SAF activity ≥2 (Mann-Whitney p<0.05) (Fig.6 and Table S16). When the more severe subset of patients that would meet current enrolment criteria for therapeutic trials in NASH were considered (defined as NASH+NAS>=4+F>=2), a 2.19 fold increase in serum AKR1B10 and a 1.51 fold increase in serum GDF15 were observed (Mann-Whitney, p<0.05). When stratifying patients based on the unsupervised clustering (n=59), amongst the detectable proteins relating to the 25-gene signature, serum AKR1B10 and GDF15 were significantly increased in patients belonging to cluster A compared to the patients from cluster B (Mann-Whitney, p<0.05) (Fig. 6).

Endoplasmic reticulum stress-induced GDF15 reduces the inflammatory response

To study the functional basis of the 25 DEGs with respect to established mechanistic processes that underpin NAFLD progression, we first assessed gene expression changes *in vitro* after Hep G2 cells were exposed to endoplasmic reticulum (ER) stress (tunicamycin or thapsigargin) or lipid loading (oleic and/or palmitic acid) (n=3 per group) (Fig. 7A). After 24h thapsigargin treatment, a significant increase in mRNA of *AKR1B10* (unpaired Student's t-test, p<0.001), *CCL20* (p<0.05), *DUSP8* (p<0.05), *GDF15* (p<0.01), *PDGFA* (p<0.01) and *TNFRSF12A* (p<0.01) was observed, while *FERMT1* (p<0.05) and *TYMS* (p<0.001) were reduced. Tunicamycin treatment significantly induced the expression of *AKR1B10* (unpaired Student's t-test, p<0.01) and *PDGFA* (p<0.05), and reduced the expression of *CCL20* (p<0.05) and *TYMS* (p<0.05). No significant differences in mRNA expression were observed after treatment with palmitic, oleic or combined palmitic/oleic acid. Western blotting confirmed the increase in AKR1B10 and GDF15 protein expression, together with an increase in the ER stress marker CHOP in Hep G2 cells after tunicamycin and thapsigargin treatment but not after lipid loading (Fig. 7B).

To investigate the potential role of GDF15 in the inflammatory response, THP-1 monocytes were differentiated into macrophages with or without recombinant human GDF15 for 48h, followed by either a 6h lipid or lipopolysaccharide (LPS) treatment. Supplementing GDF15 significantly reduced the release interleukin 6 (IL6) by the macrophages into the cell culture medium upon palmitic acid (unpaired Student's t-test, p<0.0001) and palmitic/oleic acid treatment (p<0.001) (Fig. 7C). Furthermore, GDF15 reduced the release of Tumor Necrosis Factor (TNFA) in untreated cells (p<0.01), palmitic acid (p<0.01) and palmitic/oleic acid

(p<0.001) loaded cells, and in cells treated with LPS (p<0.0001) (Fig. 7D). Additionally, a reduced release of C-C Motif Chemokine Ligand 2 (CCL2) was observed in LPS treated (p<0.001) and untreated cells (p<0.001) when conditioned with GDF15, whereas a slight increase was seen when the cells were loaded with lipids (p<0.05) (Fig. 7E). In sum, our functional results showed that ER stress is a strong inducer of several of our core signature genes and that GDF15 tempered the inflammatory response in macrophages *in vitro* upon lipid loading and LPS stimulation.

Relevance of HSD17B14 to advanced NAFLD

Genetic polymorphisms in *HSD17B13* have previously been associated with protection against more advanced steatohepatitis possibly due to modifying retinol metabolism (*24, 25*). As *HSD17B14* expression was inversely related to an increased risk of NAS≥4 using multivariate predictive models in both the discovery and replication cohorts and showed expression in a specific hepatic macrophage population, we explored the role of *HSD17B14* in retinol metabolism. To characterize the oxidoreductase activity of HSD17B14, we used NAD+/NADH luminescent assays. The recombinant protein HSD17B14 showed a 1.98-fold increase of NAD+ conversion into NADH in the presence of retinol as a substrate (p<0.001) and a 3.38-fold increase in the presence of the known substrate estradiol (as control) (p<0.001, Fig. S11) compared to the enzyme in the presence of NAD+ alone. Similar results were observed using the recombinant protein HSD17B13 as a positive control (Fig. S11).

DISCUSSION

The patterns of hepatic gene expression during NAFLD progression provide novel insights into disease mechanism and may help to identify tractable therapeutic targets. Several previous studies have attempted to address transcriptomic changes in NAFLD, however, whilst some interesting findings have emerged, many previous studies have been limited by use of expression microarrays that restrict gene coverage, small overall sample sizes that include very few cases with advanced disease, and frequently, the absence of a replication cohort (*4-12, 26*). The current, more comprehensive, transcriptomic analysis used a large independent cohort representing the full histological range of NAFLD to detect the changes in hepatic gene expression as the disease progresses by RNAseq, which has a wider dynamic range to detect gene expression changes and sets no *a priori* restrictions on gene coverage. The results of this study share some commonality with the existing literature, for example, highlighting the relevance of bile acid metabolism and the FXR/ CYP7A1 axis in NASH pathogenesis, as well as differential expression of *AKR1B10*, *CCL20*, *COL1A1*, *COL1A2*, *DUSP8*, *IL32*, *ITGBL1, STMN2*, *THY1* and *TYMS* (see Table S17). Crucially, these new data substantially extend knowledge of the transcriptomic profile of NAFL-NASH and provide greater granularity across intermediate grades and stages of disease. Adopting a hypothesis-generating pairwise analysis strategy, we have identified a 25 gene-set associated with the transition from NAFL to NASH and onward progression to fibrosis and cirrhosis. We also demonstrated that 21 of these 25 genes discriminated mild from advanced disease (NAFL-NASH F0/1 vs. NASH F≥2) in an independent

replication cohort, and so have transcriptionally defined a key group of high-risk patients that are most likely to progress to advanced disease or experience clinical events and so should plausibly be targeted for therapy (*27*).

Interestingly, only a few of the 25 genes can be considered as true extracellular matrix genes (e.g. *COL1A1* and *COL1A2*), or markers for hepatic stellate cell activation and fibrogenesis (*ITGBL1* and *STMN2*) (*28, 29*). The inflammatory genes *GDF15*, *CCL20* and *IL32* were found to be increased with disease progression, even though these ligands/chemokines have been reported to be protective against features of advanced liver disease in animal models (*30-32*). In high fat diet-fed mice, ectopic expression of Gdf15 has been reported to reduce lipid accumulation by enhancing hepatic fatty-acid oxidation, whereas overexpression of Il32 ameliorates steatosis and inflammation in a NAFLD mouse model (*30, 32*). Moreover, parenchymal expression of Ccl20 improves hepatic fibrosis through the recruitment of gamma-delta T cells in chronic carbon tetrachloride mouse models (*31*). Similarly, our data suggest that endoplasmic reticulum stress-induced release of GDF15 may ameliorate the inflammatory response in macrophages.

Other genes such as *AKR1B10* or *HSD17B14* have primarily been described to have a metabolic function (*26, 33*). AKR1B10 is an aldo/keto reductase that converts retinal into retinol (Vitamin A1) and has been reported to be critical for cell survival *in vitro* through modulation of lipid metabolism and mitochondrial function (*33*). Moreover, AKR1B10 expression is induced by the transcription factor NRF2 which activates protective pathways in response to oxidative stress (*34*). ER stress has also been reported to induce the transcriptional activity of NRF2 to promote cell survival (*35*). Our *in vitro* data showed that ER stress not only induced the expression of AKR1B10 but also GDF15, which is in line with previous reports (*36, 37*). This would suggest that some of our core signature genes are expressed by epithelial cells to deal with oxidative and/or ER stress, and to help resolve the inflammatory response in advanced NAFLD. Controversially, ER stress has been described to also induce steatosis in NAFLD, meaning that how hepatocytes deal with lipid-induced stress could actually further induce steatosis (*38*). Though the release of retinol during stellate cell activation has been well documented, its effect in other hepatic cells is less clear (*39*). Recently, Ma and colleagues reported that the polymorphism rs6834314, which is protective against NAFLD, confers a loss of enzymatic activity of HSD17B13 towards retinol (*25*). Our results showed a gradual decrease in expression of *HSD17B14*, another member of the 17-beta-hydroxysteroid dehydrogenase family as NAFLD progresses which was expressed in macrophages. This novel finding could indicate a role for retinol metabolism in hepatic macrophages during the progression of NAFLD. Importantly, increased retinol and decreased retinal means that less conversion of retinal to the biologically active retinoic acid isomers (all-*trans* and 9-*cis* retinoic acid) is possible which may affect nuclear receptor RAR/RXR signaling. A number of different enzymes contribute to these processes in addition to AKR1B10, HSD17B13 and HSD17B14 but the large increase in AKR1B10 expression as NAFLD progresses together with the decrease in HSD17B14 would tend to favor increased retinol. Whether a metabolic shift from retinal to retinol is a way for hepatocytes to survive in a background of chronic lipid-induced endoplasmic reticulum stress or whether it is a means of communicating with stellate cells and macrophages in the hepatic microenvironment, is still not clear.

Despite the wide variety in function of the 25 genes, they all showed a strong collinearity with ballooning, inflammation and fibrosis, suggesting a strong connection between cell damage,

inflammation and tissue scarring. To understand how these genes drive the disease, we used predictive multivariate models, identifying independent variables associated with specific histological features. Interestingly, the variables predicting fibrosis were different from the variables predicting a high disease activity, either based on the NAS or the SAF activity score, meaning that these variables could be regarded as proper independent drivers of disease features.

From a clinical point of view, there is a clear imperative to develop better non-invasive means to diagnose and stratify patients for treatment or enrolment into clinical trials without the need for a liver biopsy. In recent years, a great effort has been made to identify soluble serum markers to predict the presence of NASH and/or advanced liver fibrosis, such as collagen turnover biomarkers (*40*). Several of the markers we detected in our 25-gene signature proved to be candidate serum markers of advanced NAFLD, further demonstrating the robustness of this signature. Previous reports support these findings. Serum AKR1B10 has been shown to correlate with advanced disease in a small Japanese cohort of NAFLD patients, while increased serum GDF15 has been reported to associate with a greater risk of advanced fibrosis in a South Korean study (*41, 42*). In this study, the predictive univariate and multivariate models highlighted the potential of several additional genes as markers for disease activity or fibrosis grading. As a proof of concept, we showed that serum AKR1B10 and GDF15 not only stratified patients with more active steatohepatitis but also discriminated between NAFL-NASH F0/1 and NASH F≥2: the subset of NAFLD patients that are at greatest risk of future disease progression and would arguably be best targeted for therapy (*27*). Furthermore, serum AKR1B10 and GDF15 were significantly elevated in the unsupervised cluster A, characterized by high disease activity and high fibrosis ($p<0.05$).

Transcriptomic staging showed features such as senescence, DNA damage, autophagy and bile secretion/FXR signaling, that one might presume developed late in disease natural history once steatohepatitis was present, were already observed in NAFL. For example, *CYP7A1*, the rate-limiting enzyme in the classical bile acid synthesis pathway that is subject to FXR-mediated regulation, was already up-regulated in NAFL with expression peaking during the early stages of steatohepatitis (*18*). This possibly suggests that the therapeutic window for response to FXR-agonists may range from NAFL onwards, although this will require further validation. In contrast, *CCL2* expression was increased in the unsupervised cluster A characterized by high disease activity and concordant expression of ECM/fibrosis- and hepatic progenitor cell-related genes (*43*). Therapeutic inhibition of CCR2-positive monocyte-derived macrophages has been reported to reduce inflammation and fibrosis in murine NASH and fibrosis in human disease (*21, 44*). Furthermore, we found that the expression of *TREM2*, a marker for macrophages reported to be involved in hepatic fibrosis, to be associated with high disease activity and in the comparisons NAFL with NASH F2-4 (*20, 23*). Although *TREM2* is not part of our core gene signature, our integrated scRNAseq analysis indicated the importance and dynamics of specific macrophage cell populations during NAFLD progression. Different TREM2+ positive cell clusters, as well as CD163+ clusters, were enriched during different stages of NAFLD. Likewise, we identified different subpopulations of mesenchymal cells based on our core gene signature during NAFLD progression. These findings suggest that therapeutic interventions may be most efficacious if their use is targeted to the specific transcriptomic patterns that occurs as NAFLD

progresses; serum markers may be useful indicators to identify pathways susceptible to future treatments.

There are some limitations to this study. Whilst histology remains the optimum approach and accepted reference standard to accurately grade steatohepatitis and stage fibrosis, it is subject to sampling error. Four genes out of the 25 genes were not differently expressed in our replication cohort (*ANKRD29, EPB41L4A, TYMS* and *HSD17B14*). This lack of replication could reflect subtle differences in phenotype between the cohorts, for example, minor differences in histological severity of the piece of biopsy core from which RNA was extracted compared to the part examined by microscopy given that disease is not completely homogenous within the liver. However, others have previously described an increased *TYMS* expression in advanced NASH, and, using a different approach to assess risk of NAFLD progression, we demonstrated that *HSD17B14* and *EPB41L4A* expression are useful predictors of high NAS and SAF scores respectively (*7, 26*). In this study, we observed differences between the unsupervised clusters in carriage of the *PNPLA3* rs738409 variant but not for the genotypes *GCKR* rs1260326, *HSD17B13* rs72613567 or *TM6SF2* rs58542926 (*16*). Nevertheless, the carriage of these genotypes did not appear to confer distinct differences in gene expression in the RNAseq analysis. One possible explanation is that we are only looking within a relatively small NAFLD population, at least compared to GWAS studies. A more likely explanation however is that these polymorphisms contribute to the initial susceptibility to NAFLD but do not change the nature of the pathogenic processes that are in play during disease progression. Furthermore, though we confirmed that proteins encoded by many of the mRNAs in our 25-gene signature were detectable in serum and showed changes similar to those at the mRNA level during NAFLD progression, we could not confirm this for all 25 mRNA species. Detection limits of the technique used or limitations in the type of sample available could explain this.

In conclusion, the current study has identified a number of novel changes in gene expression during NAFLD progression including several that may be of diagnostic and prognostic relevance (for example, small but functionally important changes in gene expression occur from early NAFL), and has confirmed that genes such as *AKR1B10* and *GDF15* are consistent markers for NAFLD progression. AKR1B10 and HSD17B14 may contribute to retinoic acid homeostasis which, based on recent findings such as the role for HSD17B13 in NAFLD genetic susceptibility, increasingly seems relevant to NAFLD progression in addition to carcinogenesis (*41*). These data, from the largest NAFLD hepatic transcriptome dataset generated to date, provide important new insights into disease pathophysiology, identifying both stable and dynamic differences in gene expression that occur during NAFLD progression.

Materials and Methods

Study design

A total of 381 NAFLD biopsies and 305 NAFLD serum samples were included in this study covering the full histological disease spectrum. The discovery cohort of 206 NAFLD samples was processed for RNAseq while the Replication/Validation cohort including 175 NALFD cases was used for nanoString® analysis and immunohistochemistry (Fig. 1). Detailed phenotypic description and demographics are reported in Table 1 and Table S10. Both cohorts were stratified according to histopathological disease grade and stage, i.e. NAFL, NASH-F0, -F1, -F2, -F3 and -F4. Logistic modelling was used to correlate gene expression with histological features. Potentially tractable and clinically relevant disease biomarkers were tested in a cohort of 305 serum samples. To study the functional basis of core gene signatures, we used *in vitro* cell line models.

Patient selection

In this multicenter study, 436 liver biopsy samples from 403 European Caucasian patients and 305 serum samples from the same number of European Caucasian patients were included. Cases were derived from the European NAFLD Registry (NCT04442334). The discovery cohort comprised 216 snap-frozen biopsy samples from 206 patients diagnosed with NAFLD in France, Germany, Italy and the UK, and 10 "healthy obese" control cases without any biochemical or histological evidence of NAFLD from patients undergoing bariatric surgery in France. These patients were selected on the basis of both study participation and ability to isolate sufficient high quality RNA for sequencing from the liver biopsy. The replication cohort consisted of 220 formalin-fixed paraffin-embedded (FFPE) and frozen samples from 175 NAFLD patients (59 FFPE and 116 frozen) diagnosed in France and the UK, and 12 'healthy obese' control cases (frozen). All samples were centrally scored by two expert liver pathologists (DT & PB) according to the semiquantitative NASH-CRN Scoring System (NAS) and the FLIP steatosis (S), activity (A), and fibrosis (F) scoring system (*14, 15*). Fibrosis was staged from F0 through to F4 (cirrhosis). Serum samples, collected within six months of the biopsy date, were available from 305 patients with histologically proven NAFLD diagnosed in France, Germany, Italy, Sweden and the UK. 59 serum samples out of the 305 matched with patients and biopsies enrolled in the discovery cohort. Patients with alternate diagnoses and etiologies were excluded, including excessive alcohol intake (30g per day for males, 20g for females), viral hepatitis, autoimmune liver diseases and steatogenic medication use. NAFL samples with a fibrosis stage of ≥2 were not included in this study. This study was approved by the relevant Ethical Committees in the participating countries.

High-throughput RNA sequencing

Frozen tissue samples were lysed using Trizol (Sigma-Aldrich) and mRNA was extracted with the Allprep DNA/RNA Micro kit (Qiagen). Concentrations and quality were assessed using the Agilent Pico 6000 kit on the Bioanalyzer 2100 (Agilent). Samples were processed with the TruSeq RNA Library Prep Kit v2 and sequenced on the NextSeq 550 System (Illumina). Data are available on the NCBI GEO repository (GSE135251).

Bioinformatics

Fastqc (v0.11.5) and MultiQC (v1.2dev) were used to establish raw sequencing quality. Alignment to the reference genome (GRCh38, Ensembl release 76) was performed using STAR . Gene-level count tables were produced using HT-Seq. Counts were normalized using the trimmed mean of M values method (TMM) and transformed using limma's voom methodology. Normalized and transformed counts were analyzed for differential expression using linear models as implemented by limma (*45*). Statistical significance of protein-coding genes was determined by an FDR corrected q-value < 0.05 and a fold change of > 1.5 (*46*). Confounding effects were corrected for by inclusion as additive effects in the linear model used for determining differential expression. For visualization only, additive effects were subtracted from the expression data using limma's removeBatchEffects function. For each grouped comparison, a correction for batch effect and gender was applied; in the comparisons excluding the controls, an additional correction for center was implemented. *PNPLA3* rs738409, *TM6SF2* rs58542926, *GCKR* rs1260326 were determined using the RNAseq reads, *HSD17B13* rs72613567 SNP genotypes or using TaqMan® probes (Applied Biosystems) on DNA from peripheral blood mononuclear cells. If a suitable assay could not be designed, a proxy SNP was chosen (https://ldlink.nci.nih.gov/). DAVID annotation tool was used for pathway enrichment (*47, 48*). Data were visualized with GOPlot 1.0.2 and clusterProfiler (*49, 50*).

Integrated analysis was performed using single cell RNAseq data (GSE136103) from healthy liver and end-stage liver disease samples (*20*). Filtering was applied to remove any cells with greater 30% mitochondrial genes or fewer than 300 genes. The cells were normalized and clustered using Seurat as described previously (*20*). A clustering resolution of 0.2 was used resulting in 13 clusters and cell types were annotated based on the expression of specific gene markers. Cell types were then clustered at a higher resolution of 1.2. CIBERSORT analytical tool was used to determine cell type abundance within the bulk RNAseq data (*51*).

nanoString®

mRNA was isolated from FFPE samples using the High Pure FFPET RNA Isolation Kit (06650775001, Life Science Roche). Concentrations were determined using the Qubit™ RNA HS Assay kit (ThermoFisher). Frozen tissue samples were processed as described above. Custom-made assay panel (nanoString®) was used on the nanoString® nCounter system. Input was normalized to 100ng or a maximum of 6 µl volume was used. Quality control metrics were performed using the internal positive and negative control. Normalization to housekeeping genes was done using the nSolver 3.0 software (nanoString®).

Proteomics

305 serum samples (20 µl, 1 in 20 dilution) were analysed using the aptamer-based proteomic SomaScan® Platform (SomaLogic) as previously described (*52*). In brief, slow off-rate modified labeled aptamers (SOMAmer reagents) were added to each sample to form SOMAmer-protein bead complexes. After capturing of the beads and removal of nonspecifically bound reagents, the SOMAmers were quantified by hybridization to DNA microarrays. Relative quantity of SOMAmer reagents measured by the SOMAscan assay reflecting original protein concentrations

(i.e., relative fluorescent units, RFUs). RFU values were submitted to a log10-transform prior to analyses.

Additional Material and Methods can be found in the Supplementary Materials.

Supplementary Materials

Table S14. Accuracy and validation of models in the discovery and replication cohort

Table S15. Annotation clusters single-cell RNAsequencing data

Table S16. Proteomics analysis on NAFLD serum samples

Table S17. Previous reports on the relevance of 25-gene signature members to NAFLD

Data File S1. Individual–level data figure graphs

References

1. Z. Younossi, Q. M. Anstee, M. Marietti, T. Hardy, L. Henry, M. Eslam, J. George, E. Bugianesi, Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention. *Nat Rev Gastroenterol Hepatol*, (2017); published online EpubSep 20 (10.1038/nrgastro.2017.109).
2. Q. M. Anstee, H. L. Reeves, E. Kotsiliti, O. Govaere, M. Heikenwalder, From NASH to HCC: current concepts and future challenges. *Nat Rev Gastroenterol Hepatol*, (2019); published online EpubApr 26 (10.1038/s41575-019-0145-7).
3. S. McPherson, T. Hardy, E. Henderson, A. D. Burt, C. P. Day, Q. M. Anstee, Evidence of NAFLD progression from steatosis to fibrosing-steatohepatitis using paired biopsies: implications for prognosis and clinical management. *J Hepatol* **62**, 1148-1155 (2015); published online EpubMay (10.1016/j.jhep.2014.11.034).
4. J. Starmann, M. Falth, W. Spindelbock, K. L. Lanz, C. Lackner, K. Zatloukal, M. Trauner, H. Sultmann, Gene expression profiling unravels cancer-related hepatic molecular signatures in steatohepatitis but not in steatosis. *PLoS One* **7**, e46584 (2012)10.1371/journal.pone.0046584).
5. C. A. Moylan, H. Pang, A. Dellinger, A. Suzuki, M. E. Garrett, C. D. Guy, S. K. Murphy, A. E. Ashley-Koch, S. S. Choi, G. A. Michelotti, D. D. Hampton, Y. Chen, H. L. Tillmann, M. A. Hauser, M. F. Abdelmalek, A. M. Diehl, Hepatic gene expression profiles differentiate presymptomatic patients with mild versus severe nonalcoholic fatty liver disease. *Hepatology* **59**, 471-482 (2014); published online EpubFeb (10.1002/hep.26661).
6. B. M. Arendt, E. M. Comelli, D. W. Ma, W. Lou, A. Teterina, T. Kim, S. K. Fung, D. K. Wong, I. McGilvray, S. E. Fischer, J. P. Allard, Altered hepatic gene expression in nonalcoholic fatty liver disease is associated with lower hepatic n-3 and n-6 polyunsaturated fatty acids. *Hepatology* **61**, 1565-1578 (2015); published online EpubMay (10.1002/hep.27695).
7. A. Teufel, T. Itzel, W. Erhart, M. Brosch, X. Y. Wang, Y. O. Kim, W. von Schonfels, A. Herrmann, S. Bruckner, F. Stickel, J. F. Dufour, T. Chavakis, C. Hellerbrand, R. Spang, T. Maass, T. Becker, S. Schreiber, C. Schafmayer, D. Schuppan, J. Hampe, Comparison of Gene Expression Patterns Between Mouse Models of Nonalcoholic Fatty Liver Disease and Liver Tissues From Patients. *Gastroenterology* **151**, 513-525 e510 (2016); published online EpubSep (10.1053/j.gastro.2016.05.051).
8. P. Lefebvre, F. Lalloyer, E. Bauge, M. Pawlak, C. Gheeraert, H. Dehondt, J. Vanhoutte, E. Woitrain, N. Hennuyer, C. Mazuy, M. Bobowski-Gerard, F. P. Zummo, B. Derudas, A. Driessen, G. Hubens, L. Vonghia, W. J. Kwanten, P. Michielsen, T. Vanwolleghem, J. Eeckhoute, A. Verrijken, L. Van Gaal, S. Francque, B. Staels, Interspecies NASH disease activity whole-genome profiling identifies a fibrogenic role of PPARalpha-regulated dermatopontin. *JCI Insight* **2**, (2017); published online EpubJul 6 (10.1172/jci.insight.92264).
9. J. T. Haas, L. Vonghia, D. A. Mogilenko, A. Verrijken, O. Molendi-Coste, S. Fleury, A. Deprince, A. Nikitin, E. Woitrain, L. Ducrocq-Geoffroy, S. Pic, B. Derudas, H. Dehondt, C. Gheeraert, L. Van Gaal, A. Driessen, P. Lefebvre, B. Staels, S. Francque, D. Dombrowicz, Transcriptional Network Analysis Implicates Altered Hepatic Immune Function in NASH development and resolution. *Nat Metab* **1**, 604-614 (2019); published online EpubJun (10.1038/s42255-019-0076-1).
10. S. Zhao, W. P. Fung-Leung, A. Bittner, K. Ngo, X. Liu, Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One* **9**, e78644 (2014)10.1371/journal.pone.0078644).
11. G. S. Gerhard, C. Legendre, C. D. Still, X. Chu, A. Petrick, J. K. DiStefano, Transcriptomic Profiling of Obesity-Related Nonalcoholic Steatohepatitis Reveals a Core Set of Fibrosis-Specific Genes. *J Endocr Soc* **2**, 710-726 (2018); published online EpubJul 1 (10.1210/js.2018-00122).
12. M. P. Suppli, K. T. G. Rigbolt, S. S. Veidal, S. Heeboll, P. L. Eriksen, M. Demant, J. I. Bagger, J. C. Nielsen, D. Oro, S. W. Thrane, A. Lund, C. Strandberg, M. J. Konig, T. Vilsboll, N. Vrang, K. L. Thomsen, H. Gronbaek, J. Jelsing, H. H. Hansen, F. K. Knop, Hepatic transcriptome signatures in patients with varying degrees of nonalcoholic fatty liver disease compared with healthy normal-weight individuals. *Am J Physiol Gastrointest Liver Physiol* **316**, G462-G472 (2019); published online EpubApr 1 (10.1152/ajpgi.00358.2018).
13. S. A. Hoang, A. Oseini, R. E. Feaver, B. K. Cole, A. Asgharpour, R. Vincent, M. Siddiqui, M. J. Lawson, N. C. Day, J. M. Taylor, B. R. Wamhoff, F. Mirshahi, M. J. Contos, M. Idowu, A. J. Sanyal, Gene Expression

Predicts Histological Severity and Reveals Distinct Molecular Profiles of Nonalcoholic Fatty Liver Disease. *Sci Rep* **9**, 12541 (2019); published online EpubAug 29 (10.1038/s41598-019-48746-5).

14. D. E. Kleiner, E. M. Brunt, M. Van Natta, C. Behling, M. J. Contos, O. W. Cummings, L. D. Ferrell, Y. C. Liu, M. S. Torbenson, A. Unalp-Arida, M. Yeh, A. J. McCullough, A. J. Sanyal, N. Nonalcoholic Steatohepatitis Clinical Research, Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology* **41**, 1313-1321 (2005); published online EpubJun (10.1002/hep.20701).
15. P. Bedossa, C. Poitou, N. Veyrie, J. L. Bouillot, A. Basdevant, V. Paradis, J. Tordjman, K. Clement, Histopathological algorithm and scoring system for evaluation of liver lesions in morbidly obese patients. *Hepatology* **56**, 1751-1759 (2012); published online EpubNov (10.1002/hep.25889).
16. Q. M. Anstee, R. Darlay, S. Cockell, M. Meroni, O. Govaere, D. Tiniakos, A. D. Burt, P. Bedossa, J. Palmer, Y. L. Liu, G. P. Aithal, M. Allison, H. Yki-Jarvinen, M. Vacca, J. F. Dufour, P. Invernizzi, D. Prati, M. Ekstedt, S. Kechagias, S. Francque, S. Petta, E. Bugianesi, K. Clement, V. Ratziu, J. M. Schattenberg, L. Valenti, C. P. Day, H. J. Cordell, A. K. Daly, E. P. C. Investigators, Genome-wide association study of non-alcoholic fatty liver and steatohepatitis in a histologically-characterised cohort. *J Hepatol*, (2020); published online EpubApr 13 (10.1016/j.jhep.2020.04.003).
17. G. Parthasarathy, X. Revelo, H. Malhi, Pathogenesis of Nonalcoholic Steatohepatitis: An Overview. *Hepatol Commun* **4**, 478-492 (2020); published online EpubApr (10.1002/hep4.1479).
18. J. Y. Chiang, R. Kimmel, C. Weinberger, D. Stroup, Farnesoid X receptor responds to bile acids and represses cholesterol 7alpha-hydroxylase gene (CYP7A1) transcription. *J Biol Chem* **275**, 10918-10924 (2000); published online EpubApr 14 (10.1074/jbc.275.15.10918).
19. J. K. Dyson, S. McPherson, Q. M. Anstee, Non-alcoholic fatty liver disease: non-invasive investigation and risk stratification. *Journal of clinical pathology* **66**, 1033-1045 (2013); published online EpubDec (10.1136/jclinpath-2013-201620).
20. P. Ramachandran, R. Dobie, J. R. Wilson-Kanamori, E. F. Dora, B. E. P. Henderson, N. T. Luu, J. R. Portman, K. P. Matchett, M. Brice, J. A. Marwick, R. S. Taylor, M. Efremova, R. Vento-Tormo, N. O. Carragher, T. J. Kendall, J. A. Fallowfield, E. M. Harrison, D. J. Mole, S. J. Wigmore, P. N. Newsome, C. J. Weston, J. P. Iredale, F. Tacke, J. W. Pollard, C. P. Ponting, J. C. Marioni, S. A. Teichmann, N. C. Henderson, Resolving the fibrotic niche of human liver cirrhosis at single-cell level. *Nature* **575**, 512-518 (2019); published online EpubNov (10.1038/s41586-019-1631-3).
21. O. Krenkel, T. Puengel, O. Govaere, A. T. Abdallah, J. C. Mossanen, M. Kohlhepp, A. Liepelt, E. Lefebvre, T. Luedde, C. Hellerbrand, R. Weiskirchen, T. Longerich, I. G. Costa, Q. M. Anstee, C. Trautwein, F. Tacke, Therapeutic inhibition of inflammatory monocyte recruitment reduces steatohepatitis and liver fibrosis. *Hepatology* **67**, 1270-1283 (2018); published online EpubApr (10.1002/hep.29544).
22. C. Rosso, K. Kazankov, R. Younes, S. Esmaili, M. Marietti, M. Sacco, F. Carli, M. Gaggini, F. Salomone, H. J. Moller, M. L. Abate, H. Vilstrup, A. Gastaldelli, J. George, H. Gronbaek, E. Bugianesi, Crosstalk between adipose tissue insulin resistance and liver macrophages in non-alcoholic fatty liver disease. *J Hepatol* **71**, 1012-1021 (2019); published online EpubNov (10.1016/j.jhep.2019.06.031).
23. X. Xiong, H. Kuang, S. Ansari, T. Liu, J. Gong, S. Wang, X. Y. Zhao, Y. Ji, C. Li, L. Guo, L. Zhou, Z. Chen, P. Leon-Mimila, M. T. Chung, K. Kurabayashi, J. Opp, F. Campos-Perez, H. Villamil-Ramirez, S. Canizales-Quinteros, R. Lyons, C. N. Lumeng, B. Zhou, L. Qi, A. Huertas-Vazquez, A. J. Lusis, X. Z. S. Xu, S. Li, Y. Yu, J. Z. Li, J. D. Lin, Landscape of Intercellular Crosstalk in Healthy and NASH Liver Revealed by Single-Cell Secretome Gene Analysis. *Mol Cell* **75**, 644-660 e645 (2019); published online EpubAug 8 (10.1016/j.molcel.2019.07.028).
24. N. S. Abul-Husn, X. Cheng, A. H. Li, Y. Xin, C. Schurmann, P. Stevis, Y. Liu, J. Kozlitina, S. Stender, G. C. Wood, A. N. Stepanchick, M. D. Still, S. McCarthy, C. O'Dushlaine, J. S. Packer, S. Balasubramanian, N. Gosalia, D. Esopi, S. Y. Kim, S. Mukherjee, A. E. Lopez, E. D. Fuller, J. Penn, X. Chu, J. Z. Luo, U. L. Mirshahi, D. J. Carey, C. D. Still, M. D. Feldman, A. Small, S. M. Damrauer, D. J. Rader, B. Zambrowicz, W. Olson, A. J. Murphy, I. B. Borecki, A. R. Shuldiner, J. G. Reid, J. D. Overton, G. D. Yancopoulos, H. H. Hobbs, J. C. Cohen, O. Gottesman, T. M. Teslovich, A. Baras, T. Mirshahi, J. Gromada, F. E. Dewey, A Protein-Truncating HSD17B13 Variant and Protection from Chronic Liver Disease. *N Engl J Med* **378**, 1096-1106 (2018); published online EpubMar 22 (10.1056/NEJMoa1712191).
25. Y. Ma, O. V. Belyaeva, P. M. Brown, K. Fujita, K. Valles, S. Karki, Y. S. de Boer, C. Koh, Y. Chen, X. Du, S. K. Handelman, V. Chen, E. K. Speliotes, C. Nestlerode, E. Thomas, D. E. Kleiner, J. M. Zmuda, A. J. Sanyal, N. Y.

Kedishvili, T. J. Liang, Y. Rotman, 17-Beta Hydroxysteroid Dehydrogenase 13 Is a Hepatic Retinol Dehydrogenase Associated With Histological Features of Nonalcoholic Fatty Liver Disease. *Hepatology*, (2018); published online EpubNov 11 (10.1002/hep.30350).

26. M. Ryaboshapkina, M. Hammar, Human hepatic gene expression signature of non-alcoholic fatty liver disease progression, a meta-analysis. *Sci Rep* **7**, 12361 (2017); published online EpubSep 27 (10.1038/s41598-017-10930-w).
27. M. S. Siddiqui, S. A. Harrison, M. F. Abdelmalek, Q. M. Anstee, P. Bedossa, L. Castera, L. Dimick-Santos, S. L. Friedman, K. Greene, D. E. Kleiner, S. Megnien, B. A. Neuschwander-Tetri, V. Ratziu, E. Schabel, V. Miller, A. J. Sanyal, G. Liver Forum Case Definitions Working, Case definitions for inclusion and analysis of endpoints in clinical trials for nonalcoholic steatohepatitis through the lens of regulatory science. *Hepatology* **67**, 2001-2012 (2018); published online EpubMay (10.1002/hep.29607).
28. M. Wang, Q. Gong, J. Zhang, L. Chen, Z. Zhang, L. Lu, D. Yu, Y. Han, D. Zhang, P. Chen, X. Zhang, Z. Yuan, J. Huang, X. Zhang, Characterization of gene expression profiles in HBV-related liver fibrosis patients and identification of ITGBL1 as a key regulator of fibrogenesis. *Sci Rep* **7**, 43446 (2017); published online EpubMar 6 (10.1038/srep43446).
29. V. Paradis, D. Dargere, Y. Bieche, T. Asselah, P. Marcellin, M. Vidaud, P. Bedossa, SCG10 expression on activation of hepatic stellate cells promotes cell motility through interference with microtubules. *Am J Pathol* **177**, 1791-1797 (2010); published online EpubOct (10.2353/ajpath.2010.100166).
30. D. Li, H. Zhang, Y. Zhong, Hepatic GDF15 is regulated by CHOP of the unfolded protein response and alleviates NAFLD progression in obese mice. *Biochem Biophys Res Commun* **498**, 388-394 (2018); published online EpubApr 6 (10.1016/j.bbrc.2017.08.096).
31. L. Hammerich, J. M. Bangen, O. Govaere, H. W. Zimmermann, N. Gassler, S. Huss, C. Liedtke, I. Prinz, S. A. Lira, T. Luedde, T. Roskams, C. Trautwein, F. Heymann, F. Tacke, Chemokine receptor CCR6-dependent accumulation of gammadelta T cells in injured liver restricts hepatic inflammation and fibrosis. *Hepatology* **59**, 630-642 (2014); published online EpubFeb (10.1002/hep.26697).
32. D. H. Lee, J. E. Hong, H. M. Yun, C. J. Hwang, J. H. Park, S. B. Han, D. Y. Yoon, M. J. Song, J. T. Hong, Interleukin-32beta ameliorates metabolic disorder and liver damage in mice fed high-fat diet. *Obesity (Silver Spring)* **23**, 615-622 (2015); published online EpubMar (10.1002/oby.21001).
33. C. Wang, R. Yan, D. Luo, K. Watabe, D. F. Liao, D. Cao, Aldo-keto reductase family 1 member B10 promotes cell survival by regulating lipid synthesis and eliminating carbonyls. *J Biol Chem* **284**, 26742-26748 (2009); published online EpubSep 25 (10.1074/jbc.M109.022897).
34. L. E. Tebay, H. Robertson, S. T. Durant, S. R. Vitale, T. M. Penning, A. T. Dinkova-Kostova, J. D. Hayes, Mechanisms of activation of the transcription factor Nrf2 by redox stressors, nutrient cues, and energy status and the pathways through which it attenuates degenerative disease. *Free Radic Biol Med* **88**, 108-146 (2015); published online EpubNov (10.1016/j.freeradbiomed.2015.06.021).
35. S. B. Cullinan, D. Zhang, M. Hannink, E. Arvisais, R. J. Kaufman, J. A. Diehl, Nrf2 is a direct PERK substrate and effector of PERK-dependent cell survival. *Mol Cell Biol* **23**, 7198-7209 (2003); published online EpubOct (10.1128/mcb.23.20.7198-7209.2003).
36. M. Parafati, R. J. Kirby, S. Khorasanizadeh, F. Rastinejad, S. Malany, A nonalcoholic fatty liver disease model in human induced pluripotent stem cell-derived hepatocytes, created by endoplasmic reticulum stress-induced steatosis. *Dis Model Mech* **11**, (2018); published online EpubSep 25 (10.1242/dmm.033530).
37. A. P. Coll, M. Chen, P. Taskar, D. Rimmington, S. Patel, J. A. Tadross, I. Cimino, M. Yang, P. Welsh, S. Virtue, D. A. Goldspink, E. L. Miedzybrodzka, A. R. Konopka, R. R. Esponda, J. T. Huang, Y. C. L. Tung, S. Rodriguez-Cuenca, R. A. Tomaz, H. P. Harding, A. Melvin, G. S. H. Yeo, D. Preiss, A. Vidal-Puig, L. Vallier, K. S. Nair, N. J. Wareham, D. Ron, F. M. Gribble, F. Reimann, N. Sattar, D. B. Savage, B. B. Allan, S. O'Rahilly, GDF15 mediates the effects of metformin on body weight and energy balance. *Nature* **578**, 444-448 (2020); published online EpubFeb (10.1038/s41586-019-1911-y).
38. C. Lebeaupin, D. Vallee, Y. Hazari, C. Hetz, E. Chevet, B. Bailly-Maitre, Endoplasmic reticulum stress signalling and the pathogenesis of non-alcoholic fatty liver disease. *J Hepatol* **69**, 927-947 (2018); published online EpubOct (10.1016/j.jhep.2018.06.008).
39. K. Hellemans, I. Grinko, K. Rombouts, D. Schuppan, A. Geerts, All-trans and 9-cis retinoic acid alter rat hepatic stellate cell phenotype differentially. *Gut* **45**, 134-142 (1999); published online EpubJul (

40. M. Boyle, D. Tiniakos, J. M. Schattenberg, V. Ratziu, E. Bugianessi, S. Petta, C. P. Oliveira, O. Govaere, R. Younes, S. McPherson, P. Bedossa, M. J. Nielsen, M. Karsdal, D. Leeming, S. Kendrick, Q. M. Anstee, Performance of the PRO-C3 collagen neo-epitope biomarker in non-alcoholic fatty liver disease. *JHEP Rep* **1**, 188-198 (2019); published online EpubSep (10.1016/j.jhepr.2019.06.004).
41. M. Kanno, K. Kawaguchi, M. Honda, R. Horii, H. Takatori, T. Shimakami, K. Kitamura, K. Arai, T. Yamashita, Y. Sakai, T. Yamashita, E. Mizukoshi, S. Kaneko, Serum aldo-keto reductase family 1 member B10 predicts advanced liver fibrosis and fatal complications of nonalcoholic steatohepatitis. *J Gastroenterol*, (2019); published online EpubFeb 1 (10.1007/s00535-019-01551-3).
42. B. K. Koo, S. H. Um, D. S. Seo, S. K. Joo, J. M. Bae, J. H. Park, M. S. Chang, J. H. Kim, J. Lee, W. I. Jeong, W. Kim, Growth differentiation factor 15 predicts advanced fibrosis in biopsy-proven non-alcoholic fatty liver disease. *Liver Int* **38**, 695-705 (2018); published online EpubApr (10.1111/liv.13587).
43. O. Govaere, S. Cockell, M. Van Haele, J. Wouters, W. Van Delm, K. Van den Eynde, A. Bianchi, R. van Eijsden, W. Van Steenbergen, D. Monbaliu, F. Nevens, T. Roskams, High-throughput sequencing identifies aetiology-dependent differences in ductular reaction in human chronic liver disease. *J Pathol*, (2018); published online EpubDec 25 (10.1002/path.5228).
44. S. L. Friedman, V. Ratziu, S. A. Harrison, M. F. Abdelmalek, G. P. Aithal, J. Caballeria, S. Francque, G. Farrell, K. V. Kowdley, A. Craxi, K. Simon, L. Fischer, L. Melchor-Khan, J. Vest, B. L. Wiens, P. Vig, S. Seyedkazemi, Z. Goodman, V. W. Wong, R. Loomba, F. Tacke, A. Sanyal, E. Lefebvre, A Randomized, Placebo-Controlled Trial of Cenicriviroc for Treatment of Nonalcoholic Steatohepatitis with Fibrosis. *Hepatology*, (2017); published online EpubAug 17 (10.1002/hep.29477).
45. M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, G. K. Smyth, limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015); published online EpubApr 20 (10.1093/nar/gkv007).
46. J. D. Storey, R. Tibshirani, Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-9445 (2003); published online EpubAug 5 (10.1073/pnas.1530509100).
47. W. Huang da, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009)10.1038/nprot.2008.211).
48. W. Huang da, B. T. Sherman, R. A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1-13 (2009); published online EpubJan (10.1093/nar/gkn923).
49. W. Walter, F. Sanchez-Cabo, M. Ricote, GOplot: an R package for visually combining expression data with functional analysis. *Bioinformatics* **31**, 2912-2914 (2015); published online EpubSep 1 (10.1093/bioinformatics/btv300).
50. G. Yu, L. G. Wang, Y. Han, Q. Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-287 (2012); published online EpubMay (10.1089/omi.2011.0118).
51. A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, A. A. Alizadeh, Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12**, 453-457 (2015); published online EpubMay (10.1038/nmeth.3337).
52. S. A. Williams, M. Kivimaki, C. Langenberg, A. D. Hingorani, J. P. Casas, C. Bouchard, C. Jonasson, M. A. Sarzynski, M. J. Shipley, L. Alexander, J. Ash, T. Bauer, J. Chadwick, G. Datta, R. K. DeLisle, Y. Hagar, M. Hinterberg, R. Ostroff, S. Weiss, P. Ganz, N. J. Wareham, Plasma protein patterns as comprehensive indicators of health. *Nat Med* **25**, 1851-1857 (2019); published online EpubDec (10.1038/s41591-019-0665-2).

Acknowledgments

The authors would like to thank the Newcastle University Genomics Core Facility, the Newcastle NanoString Core Facility and Newcastle Molecular Pathology Node Proximity Laboratory for their technical support.

Funding

This study has been supported by the EPoS (Elucidating Pathways of Steatohepatitis) consortium funded by the Horizon 2020 Framework Program of the European Union under Grant Agreement 634413, the LITMUS (Liver Investigation: Testing Marker Utility in Steatohepatitis) consortium funded by the Innovative Medicines Initiative (IMI2) Program of the European Union under Grant Agreement 777377, the Newcastle NIHR Biomedical Research Centre, the Cambridge NIHR Biomedical Research Centre and the European NAFLD Registry.

Author contributions

QMA/AKD: study concept; QMA/AKD/OG study design and manuscript drafting; OG/SC/RQ bioinformatics; OG/SC/FR statistical analysis, DT/PB histopathology, all authors contributed to data acquisition, analysis and interpretation, and critically revised the manuscript for important intellectual content.

Competing interests

The authors disclose no conflicts.

Data and materials availability

GEO accession. GSE135251: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE135251

Figures legends

Fig. 1. Experimental study design. A total of 381 NAFLD biopsies and 305 NAFLD serum samples were included in this study. All samples have been centrally read by two expert liver pathologists.

Fig. 2. Unsupervised clustering using the RNAseq discovery cohort from 206 NAFLD patients. Distance map based on the RNAseq data indicating the distribution of clinicopathological features. Two distinct clusters, A and B, were observed.

Fig. 3. Gene signatures associated with progressive NAFLD based on RNAseq data from the discovery cohort (n=206). (A) Venn diagram illustrating the number of differentially expressed genes identified by pairwise analyses using NAFL or NASH F0/1 as a baseline to identify modifiers of steatohepatitis and fibrosis. (B) Heatmap of the 25-gene signature identified by using NAFL or NASH F0/1 as a baseline. Expression fold change is compared to NAFL. (C) 25-gene signature in the comparison of NASH F2-4 to baseline NAFL+NASH F0/1. (D) 25-gene signature in the comparison of unsupervised Cluster A vs baseline Cluster B. Data are presented using a log2 fold change in expression and –log10 of the q-values.

Fig.4. Predictive modelling of histological features using the 25-gene signature. (A) Venn diagram showing overlap of genes linearly associated with the increase in histological grading (adj $p<0.05$). (B) Hierarchical clustering of genes significantly associated (adj $p<0.05$) with at least three histological criteria. Each histological feature was treated as a continuous linear co-variate. (C-F) Association of the 25-gene signature with disease activity and stage in the RNAseq discovery cohort (n=206) using predictive multi-variate models. ROC curves showing the combined predictive model with the single significant co-variates for (C) NAS ≥ 4, (D) SAF activity ≥ 2 and (E) advanced fibrosis F3-4. (F) Comparison of the combined predictive model for advanced fibrosis with the FIB4 score (n=153 from the discovery cohort).

Fig.5. Integrated single cell RNAsequencing. (A) Identification of different cell clusters within the epithelial cells using publicly available scRNAseq data from healthy and end-stage cirrhotic liver (*20*). Expression of selected genes from our 25-gene signature in these different epithelial cell clusters (* adj $p<0.001$ indicating significantly DEG identifying the cell cluster). (B) CIBERSORT analysis to project cluster signatures onto the discovery RNAseq data set from 206 NAFLD patients. (C) Identification of different cell clusters within the macrophage cells using scRNAseq data from healthy and end-stage cirrhotic liver and visualization of selected genes in those clusters (* adj $p<0.001$ indicating significantly DEG identifying the cell cluster). (D) Projection of selected macrophage population signatures on our discovery RNAseq data.

Fig.6. Proteomics analysis on 305 serum samples from patients with histologically proven NAFLD using SomaScan. (A) Overview of detectable proteins from our 25-gene signature and their correlation with histopathological features. Correlation of AKR1B10 and GDF15 with steatosis grade (B-C), SAF activity (D-E) and Brunt fibrosis stage (F-G) for 305 serum samples. (H-I) 59 serum samples had a matching biopsy included in the discovery RNAseq cohort and circulating AKR1B10 and GDF15 were stratified based on the unsupervised clustering. (Kruskal-

Wallis test with post-hoc Bonferroni correction or Mann-Whitney-U test; RFU= relative fluorescent units).

Fig. 7. *In vitro* functional assessment of the 25-gene signature. (A) qPCR analysis for the 25-gene signature on Hep G2 cells treated with lipids (oleic, palmitic acid or combined) or with ER stress inducers tunicamycin or thapsigargin (biological replicates n=3/group). Data are presented as fold change relative to the control. (Unpaired Student's t-test, * $p<0.05$, ** $p<0.01$, *** $p<0.001$) (B) Western blot analysis for AKR1B10, GDF15 and CHOP on treated Hep G2 cells. (C-E) ELISA readout for IL6, TNFA and CCL2 on differentiated monocyte THP-1 cells with or without GDF15 pretreatment, challenged with palmitic acid, combined oleic/palmitic acid or lipopolysaccharide (LPS) (biological replicates n=3/group). Data are presented as mean +/- SD. (Unpaired Student's t-test, * $p<0.05$, ** $p<0.01$, *** $p<0.001$, **** $p<0.0001$).

Table 1. Clinicopathological characteristics of the unsupervised clusters from the RNAseq discovery cohort

Clinical features	n-value	Total (n=206)	Cluster A (n=65)	Cluster B (n=141)	p-value A vs B
Age (mean +/- SD)	206	54 (+/- 11.87)	57.08 (+/- 10.11)	52.57 (+/- 12.37)	**1.58E-02**
Sex					3.90E-01
male	206	123	36	87	
female		83	29	54	
BMI (mean +/- SD)	204	31.34 (+/- 5.04)	32.51 (+/- 4.87)	30.8 (+/- 5.04)	**1.99E-02**
T2DM	206				**6.90E-04**
no		96	19	77	
yes		110	46	64	
HBA1C (mmol/mol +/-SD)	135	48.06 (+/- 14.54)	52.16 (+/- 17.46)	45.8 (+/- 12.19)	**4.17E-02**
ALT (mean +/- SD)	204	67.12 (+/- 41.61)	71.05 (+/- 41.39)	65.33 (+/- 41.73)	1.71E-01
AST (mean +/- SD)	201	44.67 (+/- 23.08)	53.23 (+/- 26.17)	40.67 (+/- 20.39)	**4.62E-04**
Platelet ($x10^9$)	169	229.56 (+/- 65.97)	212.53 (+/- 63.15)	237.78 (+/- 65.99)	**2.05E-02**
Triglycerides (mmol/L)	180	1.95 (+/- 1.41)	2.02 (+/- 1.43)	1.91 (+/- 1.4)	7.40E-01
Total Cholesterol (mmol/L)	178	5.45 (+/- 10.14)	4.65 (+/- 1.25)	5.85 (+/- 12.37)	6.70E-01
Steatosis grade	206				4.78E-01
0		0	0	0	
1		60	16	44	
2		73	25	48	
3		73	24	49	
Ballooning	206				**1.67E-05**
0		52	10	42	
1		98	23	75	
2		56	32	24	
Kleiner Lobular Inflammation	206				**1.70E-03**
0		16	4	12	
1		95	22	73	
2		80	28	52	
3					
SAF lobular inflammation	206				**1.36E-04**
0		16	4	12	
1		140	33	107	
2		50	28	22	
Brunt Fibrosis stage	206				**5.15E-10**
0		38	5	33	
1		47	4	43	
2		53	16	37	
3		54	29	25	
4		14	11	3	
NASH	206				**2.11E-02**
no		53	10	43	
yes		153	55	98	
NAS score ≥ 4	206				**1.49E-02**
no		58	11	47	
yes		148	54	94	
SAF activity score ≥ 2	206				**8.08E-03**
no		53	9	44	
yes		153	56	97	
GCKR rs1260326 (CC/CT/TT)	206	49/105/52	12/33/20	37/72/32	3.20E-01
HSD17B13 rs72613567 (--/-T/TT) [unknown]	188	120/61/7 [18]	33/25/3 [4]	87/36/4 [14]	1.56E-01
PNPLA3 rs738409 (CC/GC/GG)	206	75/89/42	23/22/20	52/67/22	**3.06E-02**
TM6SF2 rs58542926 (CC/CT/TT)	206	156/48/2	46/19/0	110/29/2	2.64E-01

Figure 1

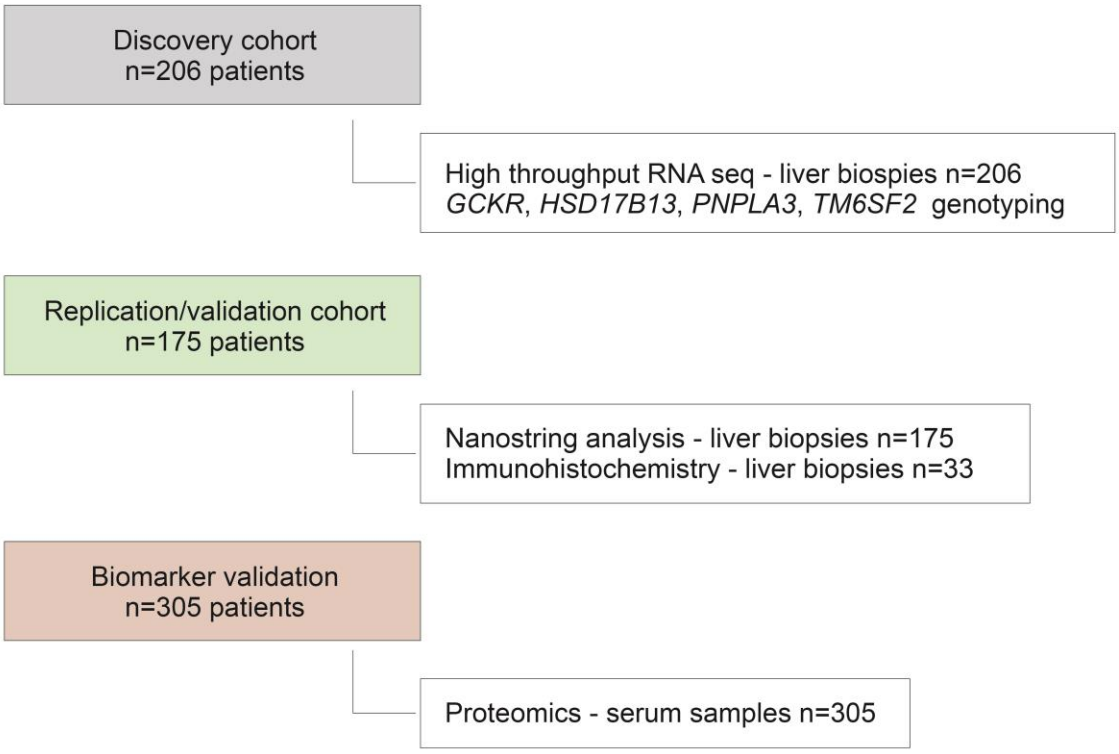

Discovery cohort
n=206 patients
High throughput RNA seq - liver biospies n=206
GCKR, HSD17B13, PNPLA3, TM6SF2 genotyping
Replication/validation cohort
n=175 patients
Nanostring analysis - liver biopsies n=175
Immunohistochemistry - liver biopsies n=33
Biomarker validation
n=305 patients
Proteomics - serum samples n=305

Figure 2

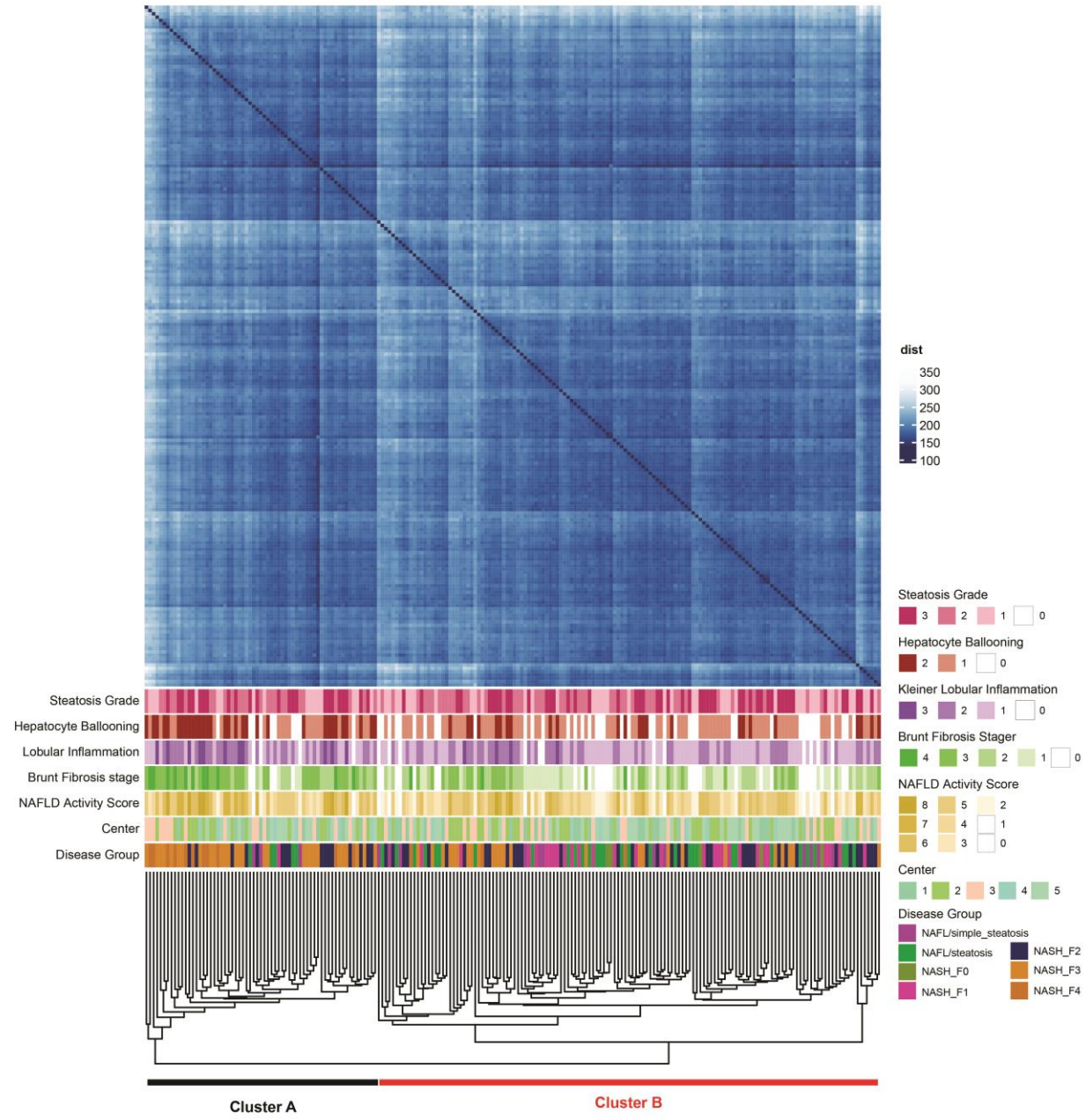

dist
350
300
250
200
150
100
Steatosis Grade
3 2 1 0
Hepatocyte Ballooning
2 1 0
Kleiner Lobular Inflammation
3 2 1 0
Brunt Fibrosis Stager
4 3 2 1 0
NAFLD Activity Score
8 5 2
7 4 1
6 3 0
Center
1 2 3 4 5
Disease Group
NAFL/simple_steatosis
NAFL/steatosis
NASH_F0
NASH_F1
NASH_F2
NASH_F3
NASH_F4
Steatosis Grade
Hepatocyte Ballooning
Lobular Inflammation
Brunt Fibrosis stage
NAFLD Activity Score
Center
Disease Group
Cluster A
Cluster B

Figure 3

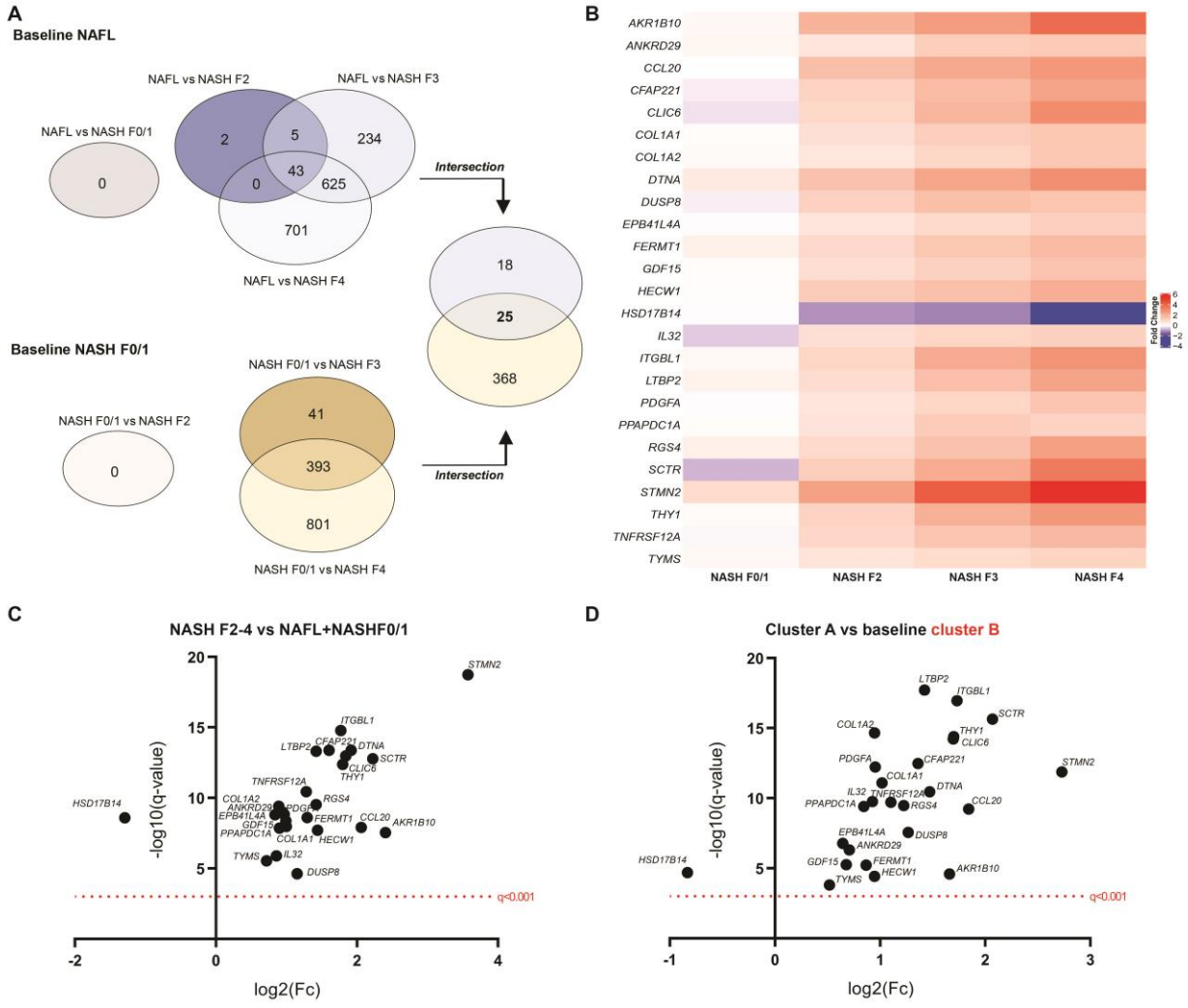

A
Baseline NAFL
NAFL vs NASH F2
NAFL vs NASH F3
NAFL vs NASH F0/1
2
5
234
0
43
0
625
701
NAFL vs NASH F4
Intersection
18
25
368
Baseline NASH F0/1
NASH F0/1 vs NASH F3
NASH F0/1 vs NASH F2
41
0
393
801
NASH F0/1 vs NASH F4
Intersection
B
AKR1B10
ANKRD29
CCL20
CFAP221
CLIC6
COL1A1
COL1A2
DTNA
DUSP8
EPB41L4A
FERMT1
GDF15
HECW1
HSD17B14
IL32
ITGBL1
LTBP2
PDGFA
PPAPDC1A
RGS4
SCTR
STMN2
THY1
TNFRSF12A
TYMS
NASH F0/1
NASH F2
NASH F3
NASH F4
Fold Change
C
NASH F2-4 vs NAFL+NASHF0/1
-log10(q-value)
log2(Fc)
q<0.001
D
Cluster A vs baseline cluster B
-log10(q-value)
log2(Fc)
q<0.001

Figure 4

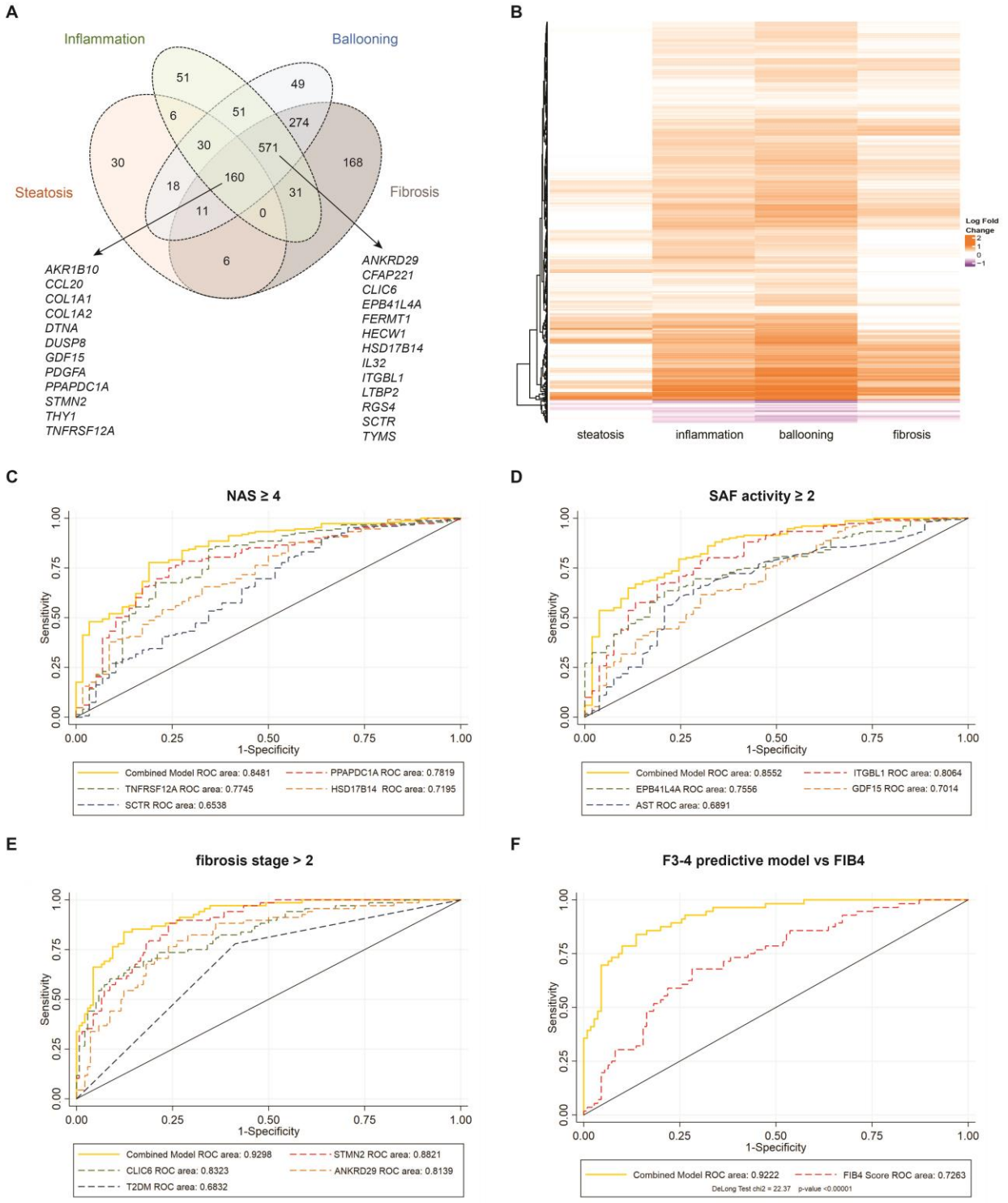

A
Inflammation
Ballooning
Steatosis
Fibrosis
51
49
6
51
274
30
571
30
168
18
160
31
11
0
6
AKR1B10
CCL20
COL1A1
COL1A2
DTNA
DUSP8
GDF15
PDGFA
PPAPDC1A
STMN2
THY1
TNFRSF12A
ANKRD29
CFAP221
CLIC6
EPB41L4A
FERMT1
HECW1
HSD17B14
IL32
ITGBL1
LTBP2
RGS4
SCTR
TYMS
B
Log Fold Change
steatosis
inflammation
ballooning
fibrosis
C
NAS ≥ 4
Sensitivity
1-Specificity
Combined Model ROC area: 0.8481
PPAPDC1A ROC area: 0.7819
TNFRSF12A ROC area: 0.7745
HSD17B14 ROC area: 0.7195
SCTR ROC area: 0.6538
D
SAF activity ≥ 2
Sensitivity
1-Specificity
Combined Model ROC area: 0.8552
ITGBL1 ROC area: 0.8064
EPB41L4A ROC area: 0.7556
GDF15 ROC area: 0.7014
AST ROC area: 0.6891
E
fibrosis stage > 2
Sensitivity
1-Specificity
Combined Model ROC area: 0.9298
STMN2 ROC area: 0.8821
CLIC6 ROC area: 0.8323
ANKRD29 ROC area: 0.8139
T2DM ROC area: 0.6832
F
F3-4 predictive model vs FIB4
Sensitivity
1-Specificity
Combined Model ROC area: 0.9222
FIB4 Score ROC area: 0.7263
DeLong Test chi2 = 22.37 p-value <0.00001

Figure 5

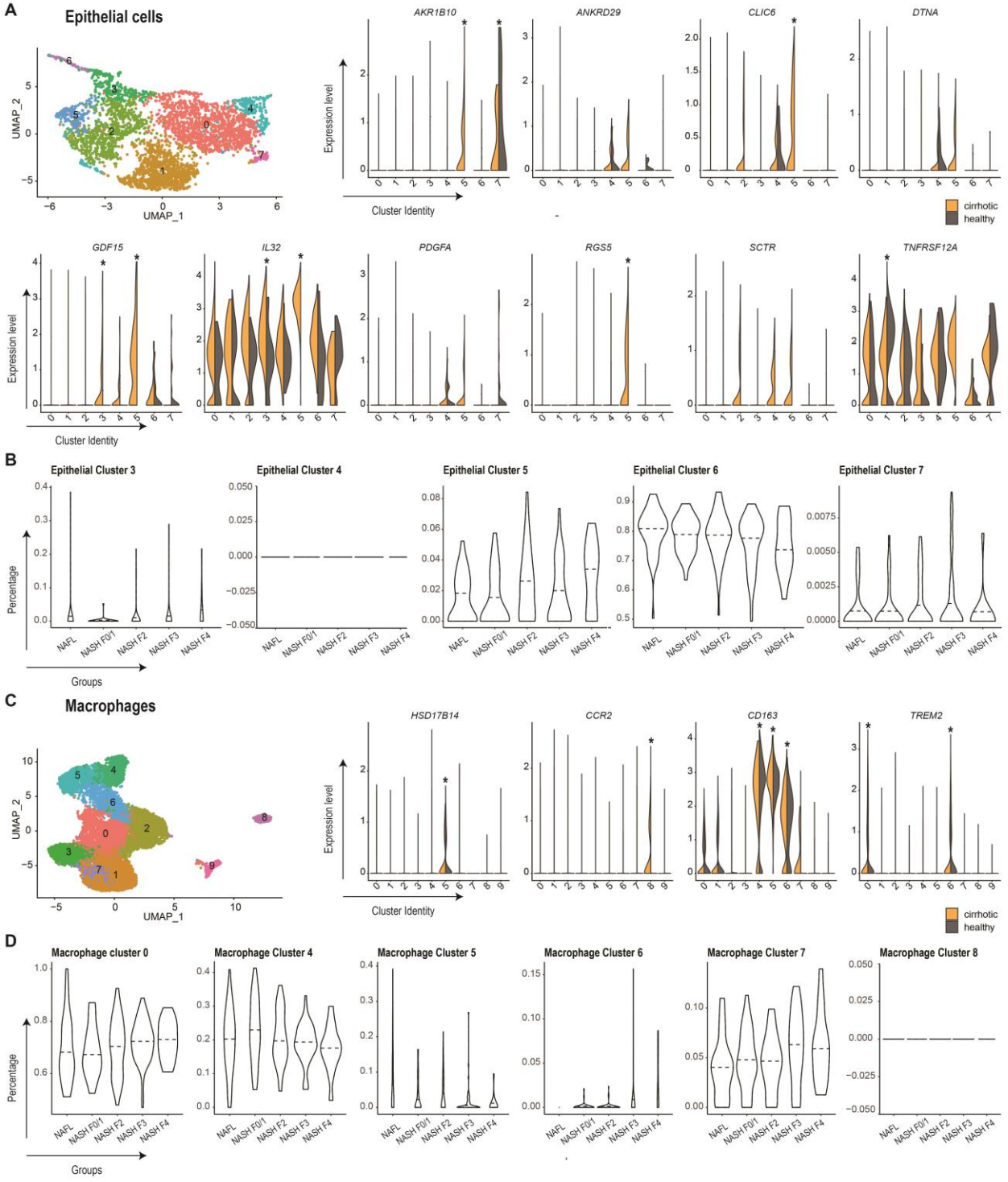

A
Epithelial cells
UMAP_2
UMAP_1
AKR1B10
ANKRD29
CLIC6
DTNA
Expression level
Cluster Identity
cirrhotic
healthy
GDF15
IL32
PDGFA
RGS5
SCTR
TNFRSF12A
B
Epithelial Cluster 3
Epithelial Cluster 4
Epithelial Cluster 5
Epithelial Cluster 6
Epithelial Cluster 7
Percentage
Groups
NAFL
NASH F0/1
NASH F2
NASH F3
NASH F4
C
Macrophages
HSD17B14
CCR2
CD163
TREM2
D
Macrophage cluster 0
Macrophage Cluster 4
Macrophage Cluster 5
Macrophage Cluster 6
Macrophage Cluster 7
Macrophage Cluster 8

Figure 6

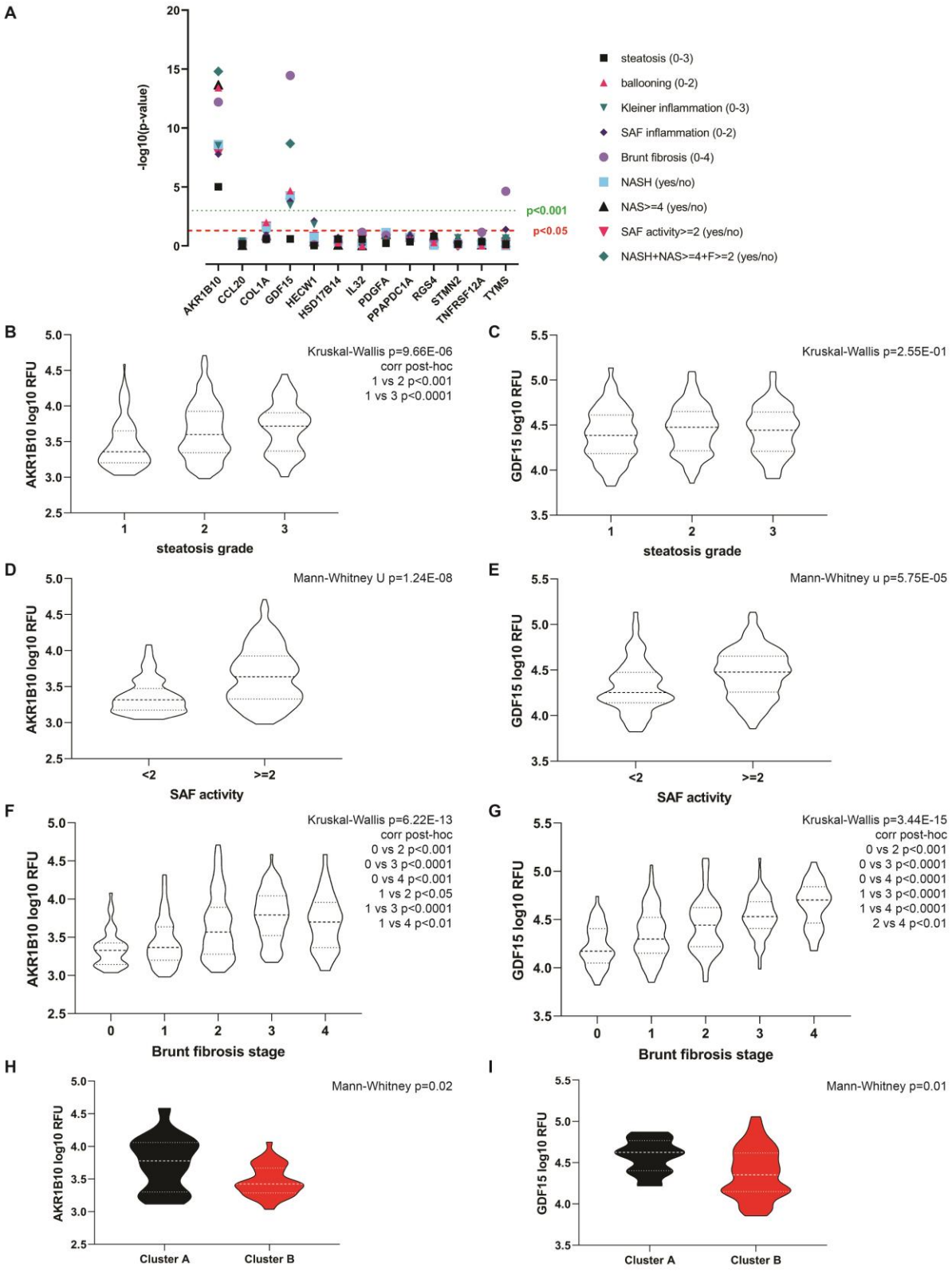
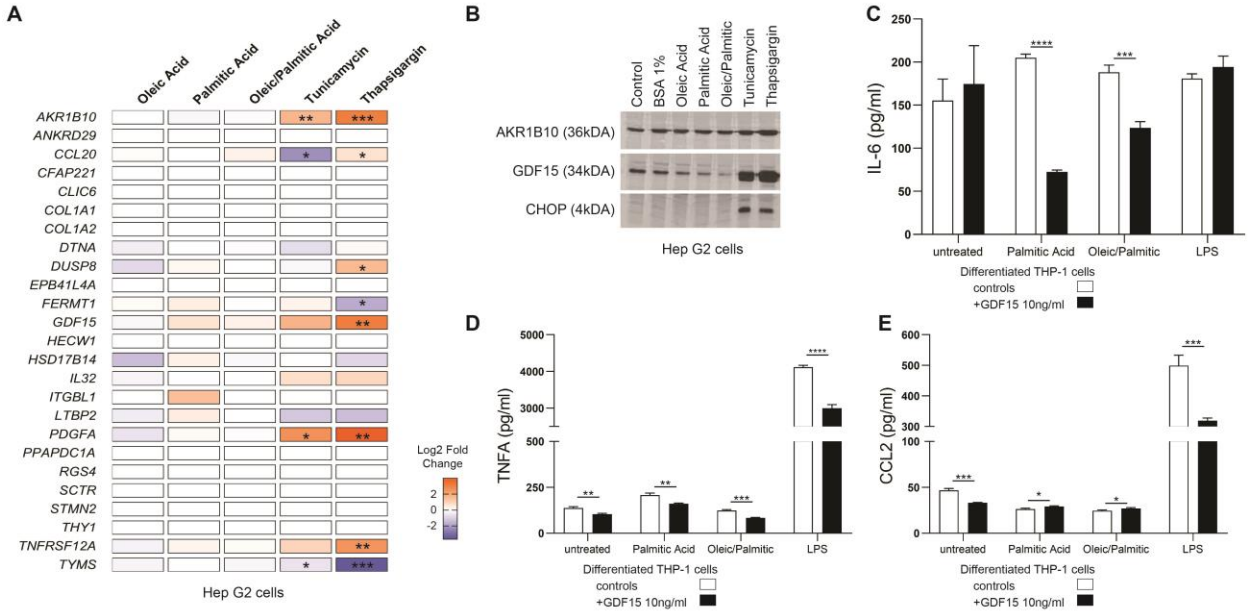

A
-log10(p-value)
p<0.001
p<0.05
steatosis (0-3)
ballooning (0-2)
Kleiner inflammation (0-3)
SAF inflammation (0-2)
Brunt fibrosis (0-4)
NASH (yes/no)
NAS>=4 (yes/no)
SAF activity>=2 (yes/no)
NASH+NAS>=4+F>=2 (yes/no)
AKR1B10
CCL20
COL1A
GDF15
HECW1
HSD17B14
IL32
PDGFA
PPAPDC1A
RGS4
STMN2
TNFRSF12A
TYMS
B
AKR1B10 log10 RFU
Kruskal-Wallis p=9.66E-06
corr post-hoc
1 vs 2 p<0.001
1 vs 3 p<0.0001
steatosis grade
C
GDF15 log10 RFU
Kruskal-Wallis p=2.55E-01
steatosis grade
D
AKR1B10 log10 RFU
Mann-Whitney U p=1.24E-08
SAF activity
E
GDF15 log10 RFU
Mann-Whitney u p=5.75E-05
SAF activity
F
AKR1B10 log10 RFU
Kruskal-Wallis p=6.22E-13
corr post-hoc
0 vs 2 p<0.001
0 vs 3 p<0.0001
0 vs 4 p<0.001
1 vs 2 p<0.05
1 vs 3 p<0.0001
1 vs 4 p<0.01
Brunt fibrosis stage
G
GDF15 log10 RFU
Kruskal-Wallis p=3.44E-15
corr post-hoc
0 vs 2 p<0.001
0 vs 3 p<0.0001
0 vs 4 p<0.0001
1 vs 3 p<0.0001
1 vs 4 p<0.0001
2 vs 4 p<0.01
Brunt fibrosis stage
H
AKR1B10 log10 RFU
Mann-Whitney p=0.02
Cluster A
Cluster B
I
GDF15 log10 RFU
Mann-Whitney p=0.01
Cluster A
Cluster B

Figure 7


A
Oleic Acid
Palmitic Acid
Oleic/Palmitic Acid
Tunicamycin
Thapsigargin
AKR1B10
ANKRD29
CCL20
CFAP221
CLIC6
COL1A1
COL1A2
DTNA
DUSP8
EPB41L4A
FERMT1
GDF15
HECW1
HSD17B14
IL32
ITGBL1
LTBP2
PDGFA
PPAPDC1A
RGS4
SCTR
STMN2
THY1
TNFRSF12A
TYMS
Log2 Fold Change
2
0
-2
Hep G2 cells
B
Control
BSA 1%
Oleic Acid
Palmitic Acid
Oleic/Palmitic
Tunicamycin
Thapsigargin
AKR1B10 (36kDA)
GDF15 (34kDA)
CHOP (4kDA)
Hep G2 cells
C
IL-6 (pg/ml)
untreated
Palmitic Acid
Oleic/Palmitic
LPS
Differentiated THP-1 cells
controls
+GDF15 10ng/ml
D
TNFA (pg/ml)
untreated
Palmitic Acid
Oleic/Palmitic
LPS
Differentiated THP-1 cells
controls
+GDF15 10ng/ml
E
CCL2 (pg/ml)
untreated
Palmitic Acid
Oleic/Palmitic
LPS
Differentiated THP-1 cells
controls
+GDF15 10ng/ml