# A novel gene-screening approach reveals QSOX1/IL1RAP as promising biomarkers for the severity of non-alcoholic fatty liver disease

Wenfeng Ma [1,2,3,4], Jinrong Huang [2], Benqiang Cai [1,4], Mumin Shao [6,7], Xuewen Yu [6,7], Mikkel Breinholt Kjær [5,8], Minling Lv [1,4], Xin Zhong [1,4], Shaomin Xu [1,4], Bolin Zhan [1,4], Qun Li [1,4], Qi Huang [1,4], Mengqing Ma [1,4], Lei Cheng [2], Yonglun Luo [2,3]*, Henning Grønbæk [5]*, Xiaozhou Zhou [1,4]*, Lin Lin [2,3]*

1 Department of Liver Disease, Shenzhen Traditional Chinese Medicine Hospital, Shenzhen, Guangdong 518033, China.

2 Department of Biomedicine, Aarhus University, Aarhus, Denmark.

3 Steno Diabetes Center Aarhus, Aarhus University Hospital, Aarhus, Denmark.

4 Department of Liver Disease, The Fourth Clinical Medical College of Guangzhou University of Chinese Medicine, Shenzhen, 518033, China.

5 Department of Hepatology and Gastroenterology, Aarhus University Hospital, Aarhus, Denmark.

6 Department of Pathology, Shenzhen Traditional Chinese Medicine Hospital, Shenzhen, Guangdong 518033, China.

7 Department of Pathology, The Fourth Clinical Medical College of Guangzhou University of Chinese Medicine, Shenzhen, 518033, China.

8 Department of Clinical Medicine, Aarhus University, Aarhus, Denmark.

* = corresponding author

## Highlights

RNA-seq data from 625 liver specimens comprising healthy controls and NAFLD

patients with increasing severity were utilized for screening NAFLD biomarkers.

- A novel unsupervised method for clustering genes based on the similarity of gene expression across each sample enhanced the discovery of novel effective non-invasive NAFLD biomarkers.
- QSOX1, IL1RAP, and especially the QSOX1/IL1RAP ratio, were found to be associated with NAFLD severity.
- The high sensitivity of the QSOX1/IL1RAP ratio in predicting NAFLD severity was validated with plasma proteomics quantification (AUROC = 0.95) and ELISA (AUROC = 0.82) in two independent patient cohorts.


## Abstract

**Background and Aims:** Non-alcoholic fatty liver disease (NAFLD) is a progressive liver disease that ranges from simple steatosis to inflammation, fibrosis, and cirrhosis. To address the unmet need for new NAFLD biomarkers, we aimed to identify candidate biomarkers using publicly available RNA sequencing (RNA-seq) and proteomics data.

**Methods:** A novel approach involving unsupervised gene clustering was performed using homogeneously processed and integrated RNA-seq data of 625 liver specimens to screen for NAFLD biomarkers, in combination with public proteomics data from healthy controls and NAFLD patients. Additionally, we validated the results in the NAFLD and healthy cohorts using enzyme-linked immunosorbent assay (ELISA) of plasma and immunohistochemical staining (IHC) of liver samples.

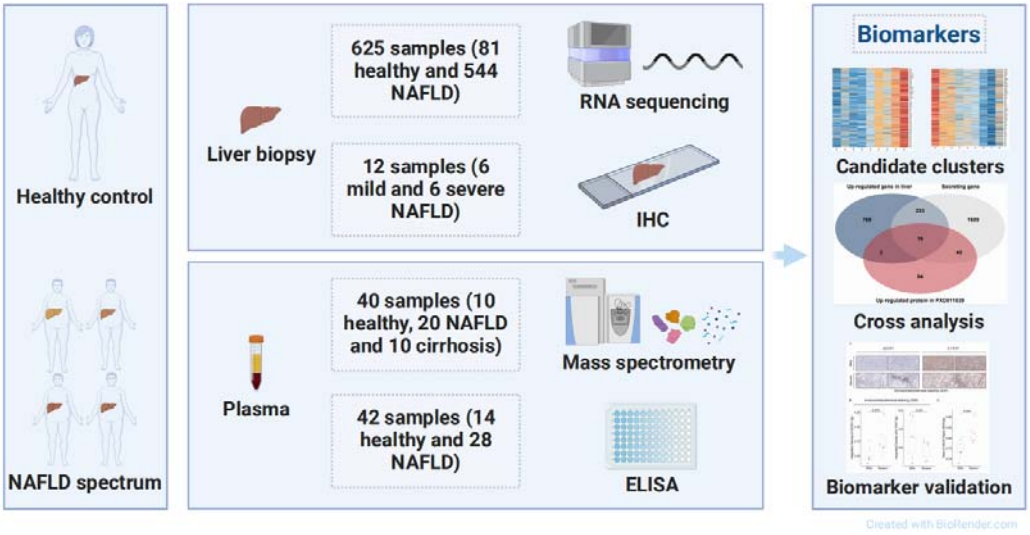
**Results:** We generated a database (https://dreamapp.biomed.au.dk/NAFLD/) for exploring gene expression changes along NAFLD progression to facilitate the identification of genes and pathways involved in the disease's progression. Through cross-analysis of the gene and protein clusters, we identified 38 genes as potential biomarkers for NAFLD severity. Up-regulation of Quiescin sulfhydryl oxidase 1

(*QSOX1*) and down-regulation of Interleukin-1 receptor accessory protein (*IL1RAP*) were associated with increasing NAFLD severity in RNA-seq and proteomics data. Particularly, the QSOX1/IL1RAP ratio in plasma demonstrated effectiveness in diagnosing NAFLD, with an area under the receiver operating characteristic (AUROC) of up to 0.95 as quantified by proteomics profiling, and an AUROC of 0.82 with ELISA. **Conclusions:** We discovered a significant association between the levels of QSOX1 and IL1RAP and NAFLD severity. Furthermore, the QSOX1/IL1RAP ratio shows promise as a non-invasive biomarker for diagnosing NAFLD and assessing its severity. (ChiCTR2300073940).

## Lay Summary

This study aimed to find non-invasive biomarkers for non-alcoholic fatty liver disease (NAFLD). Researchers utilized a new gene clustering method to analyze RNA-seq data from 625 liver samples. The identified biomarkers were further validated using plasma proteomics profiling, enzyme-linked immunosorbent assay (ELISA), and liver immunohistochemical staining (IHC) in three separate groups of healthy controls and NAFLD patients. The study revealed that the levels of QSOX1 were elevated while IL1RAP levels were reduced with increasing severity of NAFLD. Importantly, the ratio of QSOX1 to IL1RAP expression in plasma showed promise as a non-invasive diagnostic tool for assessing the severity of NAFLD, eliminating the reliance on liver biopsy.

## Graphical abstract





## Keywords

Non-alcoholic fatty liver disease, RNA sequencing data integration, non-invasive biomarker, quiescin sulfhydryl oxidase 1, interleukin-1 receptor accessory protein

## Introduction

Non-alcoholic fatty liver disease (NAFLD) is recognized as the hepatic manifestation of the metabolic syndrome, with an estimated global prevalence of around 25-32% (1, 2). The severity of this liver disease ranges from Non-alcoholic Fatty Liver (NAFL) with simple steatosis to Non-alcoholic Steatohepatitis (NASH) with inflammation and fibrosis, which can progress to NASH-induced cirrhosis and increase the risk of hepatocellular carcinoma (HCC).

Liver biopsy is currently the gold standard for histological diagnosis of NAFLD despite its associated side effects such as pain, bleeding, and rare mortality. To address these drawbacks and reduce costs, there is still an unmet need for novel, precise, and cost-effective imaging tools and non-invasive biomarkers (3). Moreover, non-invasive biomarkers are highly needed for replacing repeated liver biopsies when assessing liver histology during pharmacological interventions. Existing NAFLD biomarkers primarily focus on steatosis (e.g., SteatoTest™ or the lipid accumulation product), inflammation (e.g., circulating keratin 18 fragments [CK18), soluble CD163) or fibrosis (e.g., ELF, FibroTest or Pro-C3 tests) (4-8). Despite advancements in biomarker technology, development, and evaluation, an ideal biomarker for the diagnosis, prognosis, and assessment of treatment effects in NAFLD has yet to be identified.

The traditional RNA-seq analysis approach, which relies on established tools such as edgeR (9), DESeq2 (10), and Cufflinks (11), primarily focuses on identifying differentially expressed genes (DEGs) through pairwise comparisons (12). However, for conditions like NAFLD, which involve a complex scoring system and a continuous range of histological variations, this approach has its limitations. NAFLD doesn't involve transitioning between distinct states but represents a dynamic progression through constant histopathological changes. Pairwise comparisons oversimplify the intricate genetic alterations that occur throughout NAFLD's development. What's required is a more advanced analytical method capable of capturing the gradual and

116 overlapping gene expression changes across the entire spectrum of NAFLD. Such an
117 approach would offer a comprehensive representation of NAFLD's complexity and
118 enhance our understanding of its progression. In recent years, advancements in
119 technologies such as RNA sequencing (RNA-seq), single-cell analysis, and spatial
120 transcriptomics have provided deeper insights into the molecular and cellular
121 processes involved in NAFLD progression (13-15). Large-scale profiling efforts,
122 combined with targeted validation approaches, have led to the discovery of potential
123 biomarkers (16, 17). However, the majority of available RNA-seq data are derived
124 from smaller cohorts of NAFLD patients, which limits the comprehensive
125 understanding of NAFLD severity.
126
127 In this study, modularity optimization methods were utilized to cluster genes by
128 employing a graph-based strategy, taking into account the gene expression patterns
129 throughout the progression of NAFLD. We propose that integrating and analyzing
130 these datasets with the unbiased gene-based profiling strategy will provide further
131 insights into the molecular progression of NAFLD and the identification of biomarkers
132 associated with NAFLD severity. In the present study, we identified over 300 NAFLD
133 biomarkers by integrating and analyzing RNA-seq data from 625 liver
134 samples, including their NAFLD activity scores (NAS) and fibrosis scores,
135 along with public proteomics data. We further validated these findings in two
136 independent NAFLD cohorts, demonstrating the potential of the QSOX1/IL1RAP ratio
137 as a non-invasive biomarker for diagnosing NAFLD and assessing its severity.
138

## 139 Materials and methods

### 140 RNA-seq Data Analysis

141 1. Data Collection

142 Genome-wide RNA-seq data of human NAFLD and associated healthy controls were
143 collected from the NCBI GEO (https://www.ncbi.nlm.nih.gov/gds, access date until
144 May 2022). Only datasets that provided detailed NAS and fibrosis scores were

145 included for further investigation.

146

147 2. Data Normalization

148 The SRA-formatted data were converted into FASTQ format using 'SraToolkit'
149 (sratoolkit.2.8.2-1-centos_linux64) (https://github.com/ncbi/sra-tools). Sequencing
150 reads were aligned to the hg19 UCSC RNA sequences Genome Reference Consortium
151 Human Build 37 (GRCh37) using 'bowtie2' (bowtie2-2.2.5)
152 (https://rnnh.github.io/bioinfo-notebook/docs/bowtie2.html). Only protein-coding
153 transcripts were considered, and Transcript Per Million (TPM) values were obtained
154 by transforming the mapped transcript reads using 'RSEM' (rsem-1.2.12)
155 (https://github.com/deweylab/RSEM). Then, TPM values were then subjected to
156 Trimmed Mean of M-values (TMM) normalization across all samples using
157 'metaseqR ' (metaseqR 1.12.2) (18). The data from various sources involved in this
158 study were integrated and log1p-transformed, followed by batch correction using the
159 'removeBatchEffect' function in the R package 'limma' (limma 3.54.2)
160 (https://kasperdanielhansen.github.io/genbioconductor/html/limma.html) (19).
161 Subsequently, the data were expanded (10^x) for further analysis (**Figure 1A**).

162

163 3. RNA-seq Data Analysis and Database Construction

164 After normalization and batch correction, the RNA-seq data were subjected to
165 Principal Components Analysis (PCA) and unsupervised clustering using the R
166 package 'Seurat' (Seurat-4.3.0)
167 (https://satijalab.org/seurat/articles/get_started.html). We utilized the
168 "LogNormalize" method for global-scaling normalization, which normalized the
169 feature expression measurements across different samples for each gene by the total
170 expression. The normalized values were multiplied by a scale factor (default:
171 10,000) and log1p-transformed. Subsequently, scaling was applied to the identified
172 variable features (default: 2,000). PCA was then performed on the scaled data, with a
173 default setting of computing and storing 50 Principal Components (PCs). To cluster
174 the genes, we employed modularity optimization techniques using a graph-based

175 clustering approach. The dimensions of reduction were set to 1:20, and
176 the resolution parameter was set to 2.3 (20).
177
178 To show the gene expression variation during the development of NAFLD associated
179 with both NAS and fibrosis scores, we generated an RNA-seq database using
180 'ShinyCell' (https://github.com/SGDDNB/ShinyCell). This database was deployed at
181 https://dreamapp.biomed.au.dk/NAFLD/.
182
183 **Proteomics Data Collection and Analysis**
184 The proteomics cohort dataset (PXD011839) includes 10 healthy controls, 10 NAFLD
185 patients with normal glucose tolerance (NAFLD_ngt), 10 NAFLD patients with type 2
186 diabetes (NAFLD_T2D), and 10 NAFLD patients with cirrhosis (21). We performed the
187 statistical analysis using R-4.3.0 on the dataset (EV1, tab4)
188 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6396370/bin/MSB-15-e8793-s003.
189 xlsx). The NAFLD_ngt and NAFLD_T2D groups were merged into a single NAFLD group,
190 resulting in three groups: healthy controls, NAFLD, and cirrhosis.
191
192 Similar to RNA-seq analysis, we used 'Seurat' (Seurat-4.3.0) and employed the
193 "LogNormalize" method for global-scaling normalization. This method normalized
194 the feature expression measurements across different samples for each protein by
195 the total expression. The normalized values were multiplied by a scale factor (default:
196 10,000) and log1p-transformed. Subsequently, scaling was applied to the identified
197 variable features (default: 2,000). PCA was performed on the scaled data, with a total
198 of 39 Principal Components (PCs) computed and stored. Additionally, we calculated
199 the log fold-change of the average expression between two groups (avg_log2FC) by
200 comparing the health and NAFLD groups, as well as the NAFLD and cirrhosis groups.
201 By setting avg_log2FC > 0, we selected up- and down-regulated proteins associated
202 with increasing severity of NAFLD.
203
204 **Validation of QSOX1 and IL1RAP as Biomarkers in NAFLD**

205 1. Study Design

206 Our objective was to investigate whether *QSOX1* and *IL1RAP* gene expression levels,
207 as well as their encoded proteins, could predict the histological severity of NAFLD. To
208 address this question, we examined the plasma concentrations of QSOX1 and IL1RAP
209 in the proteomics data of healthy and NAFLD cohorts. Additionally, we conducted
210 enzyme-linked immunosorbent assay (ELISA) tests for plasma QSOX1/IL1RAP in a
211 cohort comprising healthy subjects and NAFLD patients recruited at Shenzhen
212 Traditional Chinese Medicine Hospital, China (SZTCMH).

213

214 2. Human Samples

215 A total of 28 ultrasound-proven adult NAFLD patients, including NAFLD_ngt and
216 NAFLD_T2D, and 14 healthy controls were enrolled from SZTCMH. Other diagnoses
217 and etiologies, such as excessive alcohol consumption, viral hepatitis, autoimmune
218 liver disease, and the use of steatogenic compounds, were excluded. Archived
219 plasma samples were collected between October and December 2022. Informed
220 consent was obtained from the healthy subjects and NAFLD patients, following the
221 approved clinical protocols of the Ethical Committee of SZTCMH. Clinical information,
222 including body mass index (BMI) and standard biochemistry (liver, kidney,
223 hematology) with metabolic profiling (glucose, insulin, lipids), was collected.
224 Fibroscan with controlled attenuation parameter (CAP) values was performed to
225 assess fibrosis and steatosis. Clinical information for the healthy controls and NAFLD
226 patients can be found in **Supplementary Table 4**.

227

228 For immunohistochemical staining (IHC), 12 fixed liver tissues were collected from
229 archived histological samples at SZTCMH between 2014 and 2023. These samples
230 were scored based on the NAS score (N0 to N8) and fibrosis score (F0 to F4) by two
231 pathologists (MMS and XWY). Six samples were from mild NAFLD patients (N0-4,
232 F0-2), and six were from severe NAFLD patients (N5-8, F3-4). The clinical study was
233 approved by the Ethical Committee of SZTCMH, and the approved clinical protocols
234 adhere to the Helsinki Declaration (No. K2022-174-01).

235

236 3. ELISA

237 Blind ELISA tests were conducted on the collected plasma samples. Randomly
238 assigned sample identifiers and positions were used to ensure blindness to the
239 clinical information and NAFLD stages. The levels of QSOX1 and IL1RAP were
240 measured using QSOX1 ELISA Kits (Catalog No. YJ145587, Lot No. 12/2022 from
241 Enzyme-linked Biotechnology, Shanghai, China) and IL1RAP ELISA Kits (Catalog No.
242 YJ130558, Lot No. 12/2022 from Enzyme-linked Biotechnology, Shanghai, China),
243 respectively. The measurements followed the manufacturer's instructions and the
244 absorbance values were measured at 450nm. To ensure the reliability of the ELISA
245 Kits, a pre-experiment was conducted three times before the formal experiment.

246

247 4. IHC

248 We examined the association of QSOX1 and IL1RAP with human NAFLD severity by
249 performing IHC on formalin-fixed and paraffin-embedded liver sections from 6 mild
250 NAFLD patients and 6 severe NAFLD patients. The 3 μm-thick paraffin sections were
251 deparaffinized and rehydrated with distilled water. Antigen retrieval was carried out
252 using pH 9.0 EDTA buffer, followed by 20 minutes of boiling and washing with 1X PBS.
253 Subsequently, the slides were blocked with 1% bovine serum albumin in 1X PBS for
254 15 minutes and then incubated overnight at 4℃ with QSOX1 (Rabbit anti-human,
255 Catalog No. Ab235444, Lot No. GR3386311-2 from Abcam) or IL1RAP (Rabbit
256 anti-human, Catalog No.35605, Lot No. 4926 from Sabbiotech) antibodies at a
257 concentration of 20 μg/ml. The following day, the slides were washed with 1X PBS
258 and incubated with Goat anti-rabbit IgG H&L (Catalog No. Ab205718, Lot No.
259 ab205718 from Abcam) for 15 minutes at room temperature, followed by another
260 wash with 1X PBS. The images were captured using a light microscope and
261 3DHISTECH digital scanner (https://www.3dhistech.com/).

262

263 The IHC results were analyzed using the software tool '**ImagineJ (Fiji)'.** To prevent

264 potential bias, we randomly selected five locations of the same size from each
265 sample at 20x magnification using 3DHISTECH CaseViewer_2.4
266 (https://www.3dhistech.com/solutions/caseviewer/). Using 'ImagineJ', we applied
267 the "Colour Deconvolution" tool with vectors=[H DAB]; followed by selecting the
268 Colour_2 pictures and running "8-bit". Standard thresholds were used (QSOX1:
269 setThreshold (60, 230), IL1RAP: setThreshold (94, 214)) (22, 23). The average
270 integrated density from the five sites was calculated and used as the integrated
271 density value for each sample, which was then subjected to statistical analysis.
272
273 5. Statistical analysis
274 The significance for all statistical tests was two-sided, with $P < 0.05$. All data analysis
275 was presented in the plots using R-4.3.0, and MedCalc was used to calculate the
276 AUROC, sensitivity, specificity, optimal cutoff value, and sample size.

## Results

### Overview of RNA-seq data and NAFLD patient cohorts

After applying stringent filtering criteria based on the availability of histological NAS and fibrosis scores, five datasets (including GSE115193, GSE134422, GSE135448, GSE160016, and GSE164441) were excluded from the analysis, while seven datasets (GSE105127, GSE107650, GSE126848, GSE130970, GSE135251, GSE162694, and GSE167523) were included. These datasets collectively comprise 81 healthy controls (N0F0) and 544 NAFLD patients. The severity of NAFLD patients was classified based on the NAS score (ranging from N1 to N8) and the fibrosis score (ranging from F0 to F4) using the scoring systems proposed by Brunt (24) and Kleiner (25), respectively (**Table 1**). A positive correlation was observed between NAS and fibrosis scores (Pearson R = 0.64, P = 4.94E-74, **Table 1**), indicating an association with NAFLD severity.

### Normalization and Integration of RNA-seq Data

To address the issue of batch effects resulting from differences in sequencing technology and studies, we processed the integrated data as depicted in **Figure 1A**. Genes with low expression were filtered out, resulting in a total of 17,946 protein-coding genes. Principal component analysis (PCA) demonstrated that normalization effectively eliminated noticeable batch effects (**Figure 1B**). Moreover, neither the NAS nor fibrosis scores appeared to be the main factors contributing to sample separation (**Figure 1D, E**). Instead, the normalized RNA abundance (nCount) in each sample emerged as the key component influencing transcriptome profiles (**Figure 1C**).

### Unsupervised Gene Clustering Identifies Clusters of Genes Associated with NAFLD Severity.

To identify genes associated with NAFLD severity, we utilized a previously developed unsupervised gene clustering method based on the similarity of gene expression

306 patterns across each sample (26). We employ gene clustering, grouping genes
307 according to their expression patterns during the progression of NAFLD. By setting a
308 resolution of 2.3, we identified a total of 37 gene clusters (**Supplementary Figure S1**).
309 Notably, cluster 4, consisting of 1021 genes, consistently exhibited increased
310 expression with higher NAS and fibrosis scores (**Figure 2A, Supplementary Figure**
311 **S2A**). Conversely, cluster 14, comprising 643 genes, showed decreased expression
312 with increasing NAS and fibrosis scores (**Figure 2B**, **Supplementary Figure S2B**). As
313 illustrated in figures (**Figure 2A, B**, **Supplementary Figure S2A, B, C, D**), this
314 innovative approach efficiently sorts genes into distinct groups based on their
315 expression patterns across NAFLD severity stages. It offers a more structured
316 depiction of gene expression variations, enabling a deeper understanding of NAFLD's
317 molecular pathogenesis. Through visualizing and categorizing these gene expression
318 changes, we can acquire a more comprehensive insight into the underlying
319 mechanisms and factors that propel NAFLD progression.
320
321 To explore the biological functions of these gene clusters, we performed Gene
322 Ontology (GO) analysis using the R package 'ClusterProfiler' (ClusterProfiler-4.8.0).
323 Specifically, we focused on cluster 4, which consisted of up-regulated genes. The GO
324 analysis revealed significant enrichment of genes involved in the fibrosis-related
325 process, such as extracellular matrix (ECM) organization (p.adjust = 9.77E-34),
326 extracellular structure organization (p.adjust = 9.77E-34), external encapsulating
327 structure organization (p.adjust = 1.09E-33), and cell-substrate adhesion (p.adjust =
328 3.35E-17). Notably, the expression of multiple genes involved in the ECM processes,
329 such as *COL5A3, FBLN5, SPINT2, COL1A1, COL1A2, COL3A1, COL4A1, COL4A4,*
330 *COL12A1, COL15A1*, and *COL16A1*, showed a gradual up-regulation during the
331 progression of NAFLD (**Figure 2A, Supplementary Figure S3**).
332
333 In contrast, cluster 14, which displayed a reverse correlation with NAS and fibrosis
334 scores, was significantly enriched in metabolic processes, indicating an association
335 between NAFLD progression and attenuated liver metabolism. The down-regulated

genes in this cluster were particularly enriched in processes such as organic acid catabolic process (p.adjust = 2.57E-22), carboxylic acid catabolic process (p.adjust = 2.57E-22), small molecule catabolic process (p.adjust = 5.28E-22), alpha-amino acid metabolic process (p.adjust = 6.03E-20), fatty acid metabolic process (p.adjust = 2.60E-10), and alcohol metabolic process (p.adjust = 1.35E-09). Notably, genes encoding enzymes of the Cytochrome P450 superfamily, including *CYP1A2, CYP2C19, CYP2J2, CYP2E1, CYP4A11, CYP4A22, CYP4F11, CYP2C8,* and *CYP3A4*, were down-regulated with increasing NAFLD severity (**Figure 2B**, **Supplementary Figure S4**). For a comprehensive list of all enriched GO terms for genes in cluster 4 and 14, please refer to **Supplementary Table 1**.

In addition, we developed a NAFLD gene expression database (NAFLD-DB) to facilitate the exploration and comparison of all identified protein-coding genes based on NAFLD severity. The NAFLD-DB (https://dreamapp.biomed.au.dk/NAFLD/) was constructed using the ShinyCell framework (27), which was specifically designed for convenient exploration and sharing of single-cell transcriptome data.

**Identification of Candidate Diagnostic Biomarkers.**

We employed an additional complementary strategy to further refine our list of candidate genes. In this approach, we first analyzed the plasma protein levels from a NAFLD cohort in a proteomics dataset (PXD011839) (21). We selected proteins that showed positive correlations with increasing NAFLD severity (avg_log2FC > 0) and proteins that showed negative correlations. As a result, we identified 148 up-regulated proteins and 114 down-regulated proteins associated with increasing NAFLD severity (**Figure 2C, D**).

The secretome, which consists of secreted proteins, has emerged as a valuable resource for disease diagnostics (28-30). In our study, we aim to identify potential diagnostic markers among the candidate genes, by comparing our gene clusters with the secretome database from the Human Protein Atlas (31). This cross-analysis

366 revealed a total of 349 genes encoding secreted proteins, with 249 genes showing
367 up-regulation and 100 genes showing down-regulation (**Figure 2E**). Notably, our
368 approach successfully identified a comprehensive list of previously known NAFLD
369 diagnostic and prognostic markers, including *ADAMTSL2* (32), *AEBP1* (33), and *BGN*
370 (34) (**Supplementary Table 2**), further validating the effectiveness of our approach.
371
372 Next, we intersected the protein-encoding genes of these proteins with the
373 secretome genes and the candidate genes generated from our RNA-seq analysis.
374 Through this cross-comparison, we identified 16 up-regulated secreting genes (*A2M,*
375 *C7, COL6A3, COLEC11, ENPP2, FBLN1, FBN1, FCGBP, IGFBP6, LCN2, LUM, MMP2,*
376 *PAPLN, PTGDS,* ***QSOX1****, VWF*) and 22 down-regulated secreting genes (*AZGP1, C1RL,*
377 *C4BPA, C6, C8B, CFHR3, CNDP1, F2, GC, HP, HPR,* ***IL1RAP****, ITIH1, ITIH2, ITIH4, KLKB1,*
378 *PON3, SERPINA10, SERPINC1, SERPING1, SMPDL3A, TTR*) associated with increasing
379 NAFLD severity in both the RNA-seq and proteomics data (**Supplementary Figure S5**).
380
381 **QSOX1 and IL1RAP are promising biomarkers for NAFLD severity**
382 To demonstrate the applicability of our NAFLD-DB and validate the association of
383 differential gene expression with increasing NAFLD severity, we selected two
384 representative genes, *QSOX1* and *IL1RAP,* which showed positive and negative
385 correlations with increasing NAFLD severity**.** We examined their expression levels in
386 comparison to patients with a NAS or fibrosis score of 0 (N0 or F0). The expression of
387 *QSOX1* was significantly correlated with the severity of NAFLD compared to N0 or F0
388 patients: N1-4 (p = 0.003), N5-8 (p = 1.9E-10), F1-2 (p = 0.001), F3-4 (p = 6.5E-8)
389 (**Figure 3A, B**). On the other hand, *IL1RAP* expression was significantly lower in
390 patients with increased NAFLD severity compared to N0 or F0: N1-4 (p = 1E-5), N5-8
391 (p = 4.7E-10), F1-2 (p = 0.00012), F3-4 (p = 0.00013) (**Figure 3C, D**).
392
393 Since *QSOX1* and *IL1RAP* exhibited opposite correlations with NAFLD severity, we
394 further explored whether the ratio of QSOX1/IL1RAP could better distinguish
395 between patient groups. Our results showed that compared to N0 or F0 patients, the

396 ratio of QSOX1 to IL1RAP mRNA levels showed even greater separation: N1-4 (p =
397 7.6E-8), N5-8 (5.9E-16), F1-2 (4.6E-6), F3-4 (6.8E-8) (**Figure 3E, F**). These findings
398 suggest that the QSOX1/IL1RAP ratio has the potential as a biomarker for diagnosing
399 NAFLD severity.
400
401 **Validation of Plasma QSOX1/IL1RAP Levels as Biomarkers for NAFLD Severity with**
402 **NAFLD Proteomics Cohort**
403 To further validate the potential of QSOX1 and IL1RAP as biomarkers for NAFLD
404 severity, we analyzed the plasma levels of QSOX1 and IL1RAP in a NAFLD proteomics
405 cohort (PXD011839) previously conducted by Niu L. and colleagues. This cohort
406 consisted of 10 healthy controls and 30 patients with NAFLD or NAFLD-induced
407 cirrhosis (21).
408
409 Consistent with our liver RNA profiling results in livers, the analysis of plasma
410 proteomics data from this independent NAFLD cohort showed a significant increase
411 in plasma QSOX1 levels in patients with NAFLD (Wilcoxon rank sum test, p = 0.021)
412 and cirrhosis (p = 0.049) compared to healthy controls (**Figure 4A**). Conversely,
413 IL1RAP levels were significantly reduced in patients with NAFLD (p = 5.8E-5) and
414 cirrhosis (p = 0.0011) (**Figure 4B**). Moreover, when considering the combined marker
415 of plasma QSOX1/IL1RAP ratio, it demonstrated even greater significance in
416 distinguishing NAFLD (p = 9.3E-6) and cirrhosis (p = 0.00013) patients from the
417 control group, compared to using QSOX1 or IL1RAP alone (**Figure 4C**).
418
419 To access the diagnostic sensitivity and specificity of QSOX1, IL1RAP, and their ratio
420 for NAFLD severity, we conducted ROC curve analysis using the 'MedCalc' tool (30).
421 The sample sizes for each comparison were evaluated and are listed in
422 **Supplementary Table 3**. The AUROC of the QSOX1/IL1RAP ratio for distinguishing
423 NAFLD patients from healthy controls was 0.95, with a cutoff value of 1.12. The
424 sensitivity was determined to be 90%, and the specificity was 100%. Notably, the
425 efficacy of the QSOX1/IL1RAP ratio was superior to that of IL1RAP alone

426 (AUROC=0.92) or QSOX1 alone (not significant). Similarly, when assessing the
427 differentiation between cirrhosis patients and healthy controls, the AUROC of the
428 QSOX1/IL1RAP ratio was 0.96, with a cutoff value of 1.12. The sensitivity was 90%,
429 and the specificity was 100%.
430
431 These results indicate that the QSOX1/IL1RAP ratio holds promise as a highly
432 effective biomarker for diagnosing NAFLD severity, surpassing the individual
433 biomarkers alone, and maintaining better sensitivity and specificity in distinguishing
434 NAFLD patients and cirrhosis patients from healthy individuals.
435
436 **Validation of QSOX1 and IL1RAP as biomarkers for NAFLD in another patient cohort**
437 To further validate the utility of QSOX1 and IL1RAP as biomarkers for NAFLD, we
438 conducted a validation study in healthy controls and NAFLD patients recruited from
439 the Department of Liver Disease of Shenzhen Traditional Chinese Medicine Hospital.
440 Plasma samples were collected from 14 healthy subjects and 28 newly diagnosed
441 NAFLD patients. Clinical information for the healthy controls and NAFLD patients can
442 be found in **Supplementary Table 4**.
443
444 We measured plasma levels of QSOX1 and IL1RAP using an enzyme-linked
445 immunosorbent assay (ELISA). Consistent with our previous findings, plasma levels of
446 QSOX1 (Wilcoxon rank sum test, $p = 0.043$), IL1RAP ($p = 0.035$), and the
447 QSOX1/IL1RAP ratio ($p = 0.00061$) were significantly different between NAFLD
448 patients and controls (**Figure 4D, E, F**).
449
450 To assess the diagnostic value of QSOX1 and IL1RAP as non-invasive biomarkers for
451 NAFLD by ELISA, we calculated the AUROC of the QSOX1/IL1RAP ratio in the ELISA
452 test to distinguish NAFLD patients from healthy controls. The QSOX1/IL1RAP ratio
453 exhibited an AUROC of 0.82. Using a cutoff of 0.05, the sensitivity was 93% and the
454 specificity was 57%. In comparison, the AUROC of QSOX1/IL1RAP ratio quantified by
455 ELISA showed less efficacy in distinguishing NAFLD patients from healthy controls

456 (**Supplementary Table 3**), which may be attributed to the sensitivity of protein
457 quantification methods and small sample size.
458
459 To further validate the association between QSOX1 and IL1RAP protein levels and
460 NAFLD severity, we assessed their levels in liver biopsies from mild and severe NAFLD
461 patients using IHC. Our results consistently demonstrated a significant correlation
462 between QSOX1 and IL1RAP levels and NAFLD severity (**Figure 5A**). Quantification of
463 QSOX1 and IL1RAP levels based on IHC confirmed that the QSOX1/IL1RAP ratio ($p =$
464 $0.027$) could distinguish the severe NAFLD group (n=6; NAS 5-8, fibrosis score 3-4)
465 from the mild NAFLD group (n=6; NAS 0-4, Fibrosis score 0-2) (**Figure 5B, C**).
466 Collectively, these findings suggest that the QSOX1/IL1RAP ratio holds promise as an
467 effective biomarker for the early diagnosis and prediction of NAFLD severity.

468 ## Discussion

469 This study is the first to integrate publicly available RNA-seq datasets from over 600
470 NAFLD patients with varying stages of disease severity, combined with proteomics
471 data analysis of publicly available datasets. The key findings suggest that the
472 QSOX1/IL1RAP, and particularly the QSOX1/IL1RAP ratio hold promise as potential
473 biomarkers for NAFLD severity assessment. These results align with recent research
474 highlighting the importance of different transcriptional profiles specific to NAS and
475 fibrosis scores, offering valuable insights into the molecular mechanisms driving
476 disease progression from simple steatosis to inflammation and fibrosis (35, 36).

477

478 **The Advantages of Utilizing Integrated RNA-seq Data for Investigating NAFLD**
479 **Biomarkers**

480 Despite the growing availability of RNA-seq data in this field, many original studies
481 have been limited by small sample sizes and biased sample distribution, making it
482 challenging to accurately decipher transcriptional differences across various stages of
483 NAFLD(35-39). Several studies have attempted to identify diagnostic biomarkers and
484 potential drug targets. For instance, Brosch et al. conducted a positional analysis of
485 transcriptomes across three micro-dissected liver zones from 19 NAFLD patients (37).
486 Suppli et al. demonstrated that immunohistochemical markers offer greater
487 objectivity in distinguishing hepatocyte injury between NASH and NAFL (38). In the
488 pursuit of diagnostic genes and novel drug targets, Hoang et al. studied 6
489 histologically normal and 72 NAFLD patients, while Pantano et al. studied 31
490 histologically normal and 112 NAFLD patients. These studies revealed that specific
491 cells proportion and candidate gene signatures can accurately predict fibrosis stage
492 and disease progression (35, 36). Likewise, Govaere et al. observed the correlation
493 between gene expression and histology in a cohort of 10 controls and 206 NAFLD
494 patients (39). In contrast to the studies above that identified sets of potential
495 biomarker genes, Kozumi et al. validated thrombospondin 2 (THBS2) as a noninvasive
496 biomarker for NAFLD. They confirmed its potential in identifying the disease stages

497 among 98 NAFLD patients, and the serum levels of its encoded protein TSP-2,
498 measured by ELISA, showed an AUROC of 0.78 in the diagnosing of NASH among 213
499 patients with biopsy-proven NAFLD (40). The major challenge of combining and
500 analyzing these diverse datasets lies in achieving homogeneous processing, which
501 requires substantial time and computational resources (41, 42). To generate more
502 robust and compelling results, we employed unbiased integration of comprehensive
503 NAFLD data to profile the liver transcriptome across a broad spectrum of NAFLD
504 severity in our study, incorporating all the aforementioned samples.
505
506 **The Superiority of QSOX1 and IL1RAP as Potential Biomarkers of NAFLD**
507 The high prevalence and associated risks of NAFLD have driven global efforts to
508 identify improved diagnostic biomarkers. However, most existing biomarkers are
509 primarily suited for evaluating fibrosis (3, 43-45). The Fibrosis-4 (FIB-4) test
510 commonly used in clinical practice, is sub-optimal for screening purposes, as it carries
511 the risks of both overdiagnosis and false negatives, particularly in patients at risk of
512 chronic liver disease (8). Although the patented ELF™ test was highly recommended
513 for ruling out advanced fibrosis, it comes with higher costs. Several steatosis scores,
514 such as the SteatoTest™ and the fatty liver index (FLI), have been proposed for
515 steatosis detection, but they do not provide substantial additional information
516 beyond routine clinical, laboratory, and imaging examinations conducted in
517 patients suspected of having NAFLD(8). Non-coding RNAs (ncRNAs), which exhibit
518 aberrant expression associated with NAFLD, have emerged as potential biomarkers
519 for NAFLD pathology, and circulating ncRNAs including miR-122 and lncRNAs are
520 proposed as potential biomarkers for NAFLD severity and progression (46-52).
521 Despite the development of new biomarkers, there is still uncertainty surrounding
522 their predictive value, underscoring the urgent need to develop novel, cost-effective,
523 and efficient biomarkers with high sensitivity and specificity for NAFLD prediction
524 and monitoring (4, 53).
525
526 The approach by Hoang et al. (35) centered on identifying genes with diverse

527 expressions associated with NAFLD severity, inspired us to develop our gene
528 clustering method. Our approach surpasses the constraints of conventional RNA-seq
529 data analysis, which predominantly relies on pairwise comparisons. Instead, it
530 classifies genes according to their dynamic expression patterns, enabling a more
531 comprehensive and dynamic perspective of molecular alterations as NAFLD
532 progresses. This method has the potential to map NAFLD severity and progression
533 solely through gene expressions, thus avoiding invasive procedures like liver biopsies.
534 Moreover, the gene-based scoring system can forecast NAFLD progression,
535 facilitating early interventions for patients at risk of advancing to severe disease
536 stages.
537
538 Previous studies have explored the relationship between QSOX1, IL1RAP, and NAFLD
539 or steatosis(16, 54). QSOX1 has been suggested as a potential diagnostic biomarker
540 for NAFLD, playing a significant role in lipid metabolism as an enzyme expressed in
541 various tissues, particularly in quiescent fibroblasts (37, 55, 56). IL1RAP is localized in
542 vesicles and cytosol, and it is secreted into the bloodstream. Notably, IL1RAP
543 expression at the RNA level was specifically detected in the liver and hepatocytes
544 (31). Hence, the combination of QSOX1 and IL1RAP as secretome genes and proteins
545 was selected as a potential biomarker combination.
546
547 The potential of QSOX1, IL1RAP, and their ratio as biomarkers for NAFLD was
548 demonstrated through the analysis of public RNA-seq and proteomics data, ELISA
549 tests conducted on patients' plasma, and IHC performed on fixed liver slides. These
550 findings suggest that QSOX1, IL1RAP, and their ratio hold promise as effective
551 biomarkers for NAFLD. Notably, the higher AUROC values for NAFLD diagnosis
552 achieved by QSOX1, IL1RAP, and their ratio highlight their efficacy as NAFLD
553 biomarkers.
554
555 **Limitation and Future Prospects.**
556 The current study possesses several strengths, including the integration and

557 processing of RNA-seq data from over 600 NAFLD patients with varying degrees of
558 NAFLD severity, as well as validation using proteomics data and samples from NAFLD
559 patients and controls. Furthermore, the well-established database with a
560 user-friendly interface could benefit the research community in exploring
561 differentially expressed genes in NAFLD at various stages. However, there are also
562 limitations to consider. For instance, some samples in the GSE126848 and GSE167523
563 datasets lacked individual NAS and fibrosis scores. To address this issue,
564 we standardized scores based on their categories in the original articles, and the
565 impact on the results was deemed negligible due to the provision of general stages
566 and unsupervised gene clustering. Machine learning, an essential tool for biomarker
567 validation and sample classification validation, should be employed to train large
568 cohorts of biopsy-proven NAFLD patients and healthy controls. However, this would
569 require an extended recruiting period(57) to determine the sensitivity and specificity
570 of the QSOX1/IL1RAP ratio for NAFLD diagnosis and staging.
571
572 Although newer technologies such as single-cell RNA sequencing (scRNA-seq) and
573 spatial sequencing have gained popularity, RNA-seq still serves as a valuable tool in
574 uncovering the pathogenesis of NAFLD (17). Computational analysis limitations make
575 it impractical for large cohort research, and single-cell suspension processing may
576 affect cell abundance and cell type representation, particularly in hepatic ballooning
577 cells in NAFLD (58). Single-nuclei RNA sequencing (snRNA-seq) captures cell
578 frequency more accurately than scRNA-seq but captures lower gene expression.
579 Spatial transcriptomics and proteomics have limitations for discovering invasive
580 biomarkers of NAFLD as they focus on small sampling areas (15). The combination of
581 all these biological tools holds potential for future research.
582
583 In conclusion, through a novel approach of unsupervised gene clustering performed
584 on integrated RNA-seq data, we have discovered a significant association between
585 QSOX1 and IL1RAP levels and NAFLD severity, with their ratio showing potential as a
586 non-invasive biomarker for diagnosing and assessing the severity of NAFLD.

587 Validation of our plasma-level findings in larger cohorts of liver biopsies is required,
588 but it holds promise as a new tool to diagnose NAFLD severity and reduce the need
589 for liver biopsies. Our approach may lead to the discovery of more NAFLD biomarkers,
590 and the ratios of other up-regulated and down-regulated genes associated with
591 increasing NAFLD severity also have the potential to be verified as potential
592 biomarkers.
593

## 594 Abbreviations

595 AUROC, area under the receiver operating characteristic; avg_log2FC, log fold-change
596 of the average expression between the two groups; BMI, body mass index; CK18,
597 circulating keratin 18 fragments; ECM, extracellular matrix; ELISA, enzyme-linked
598 immunosorbent assay; F, Fibrosis score; FIB-4, Fibrosis-4; GEO, Gene Expression
599 Omnibus; GO, Gene Ontology; GRCh37, Genome Reference Consortium Human Build
600 37; HCC, hepatocellular carcinoma; IHC, immunohistochemistry staining; IL1RAP,
601 Interleukin-1 receptor accessory protein; lg, log10; PCA, Principal components
602 analysis; QSOX1, Quiescin sulfhydryl oxidase 1; RNA-seq, RNA sequencing; N, NAS
603 score; NAFL, Non-alcoholic Fatty Liver; NAFLD, Non-alcoholic fatty liver disease;
604 NAFLD-DB, NAFLD gene expression database; NAFLD_ngt, NAFLD with normal
605 glucose tolerance; NAFLD_T2D, NAFLD with type 2 diabetes; NAS, NAFLD activity
606 scores; NASH, Non-alcoholic Steatohepatitis; ncRNAs, non-coding RNAs; scRNA-seq,
607 single-cell RNA sequencing; snRNA-seq, Single-nuclei RNA sequencing; SZTCMH,
608 Shenzhen Traditional Chinese Medicine Hospital, China; THBS2, thrombospondin 2;
609 TMM, Trimmed Mean of M-values; TPM, Transcript Per Million.
610

## 611 Financial support

612 This research was funded by the Shenzhen Science and Technology Project and
613 Sanming Project of Medicine in Shenzhen, China (grant nos. SZSM201612074,
614 JCYJ20210324120405015).
615

## Authors' contributions

YLL, HG, WFM and LL designed the study and interpreted the data. The analysis strategy has been developed by LL and WFM. WFM, JRH, BQC, MLL, XZ, SMX and MBK collected and assembled the data. WFM drafted the manuscript. WFM, JRH and LC performed data analysis and/or interpretation. Technical support: YLL, LL, JRH, ZXZ, MMS and XWY. Study participant inclusion: ZXZ, WFM, BQC, MLL, XZ, SMX, BLZ, QL, QH, MQM. All authors reviewed and approved the final version of the manuscript.

## Data and code availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## Conflict of interest

Henning Grønbæk has received research grants from Abbvie, Intercept, ARLA Food for Health, ADS AIPHIA Development Services AG. Consulting Fees from Ipsen, NOVO, Pfizer. Lecturer for AstraZeneca and EISAI; and on Data Monitoring Committee at CAMURUS AB. All other authors have no conflicts of interest to declare.

## Reference


1. Lazarus JV, Mark HE, Anstee QM, Arab JP, Batterham RL, Castera L, Cortez-Pinto H, et al. Advancing the global public health agenda for NAFLD: a consensus statement. Nat Rev Gastroenterol Hepatol 2022;19:60-78.
2. Riazi K, Azhari H, Charette JH, Underwood FE, King JA, Afshar EE, Swain MG, et al. The prevalence and incidence of NAFLD worldwide: a systematic review and meta-analysis. Lancet Gastroenterol Hepatol 2022;7:851-861.
3. Nassir F. NAFLD: Mechanisms, Treatments, and Biomarkers. Biomolecules 2022;12.
4. Vilar-Gomez E, Chalasani N. Non-invasive assessment of non-alcoholic fatty liver disease: Clinical prediction rules and blood-based biomarkers. J Hepatol 2018;68:305-315.
5. Kazankov K, Barrera F, Moller HJ, Rosso C, Bugianesi E, David E, Younes R, et al. The macrophage activation marker sCD163 is associated with morphological disease stages in patients with non-alcoholic fatty liver disease. Liver Int 2016;36:1549-1557.
6. Anstee QM, Castera L, Loomba R. Impact of non-invasive biomarkers on hepatology practice: Past, present and future. J Hepatol 2022;76:1362-1378.

649 7. Piazzolla VA, Mangia A. Noninvasive Diagnosis of NAFLD and NASH. Cells 2020;9.
650 8. European Association for the Study of the Liver. Electronic address eee, Clinical Practice
651 Guideline P, Chair, representative EGB, Panel m. EASL Clinical Practice Guidelines on non-invasive tests
652 for evaluation of liver disease severity and prognosis - 2021 update. J Hepatol 2021;75:659-689.
653 9. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression
654 analysis of digital gene expression data. Bioinformatics 2010;26:139-140.
655 10. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq
656 data with DESeq2. Genome Biol 2014;15:550.
657 11. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, et al. Differential gene and
658 transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc
659 2012;7:562-578.
660 12. Costa-Silva J, Domingues DS, Menotti D, Hungria M, Lopes FM. Temporal progress of gene
661 expression analysis with RNA-Seq data: A review on the relationship between computational methods.
662 Comput Struct Biotechnol J 2023;21:86-98.
663 13. Barreby E, Chen P, Aouadi M. Macrophage functional diversity in NAFLD - more than
664 inflammation. Nat Rev Endocrinol 2022;18:461-472.
665 14. Wang ZY, Keogh A, Waldt A, Cuttat R, Neri M, Zhu S, Schuierer S, et al. Single-cell and bulk
666 transcriptomics of the liver reveals potential targets of NASH with fibrosis. Sci Rep 2021;11:19396.
667 15. Guilliams M, Bonnardel J, Haest B, Vanderborght B, Wagner C, Remmerie A, Bujko A, et al. Spatial
668 proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. Cell
669 2022;185:379-396 e338.
670 16. Green CD, Dozmorov MG, Spiegel S. Analysis of Liver Responses to Non-alcoholic Steatohepatitis
671 by mRNA-Sequencing. Methods Mol Biol 2022;2455:163-179.
672 17. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet
673 2019;20:631-656.
674 18. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling,
675 and transformations: improving the biological information content of metabolomics data. BMC
676 Genomics 2006;7:142.
677 19. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential
678 expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47.
679 20. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A, Lee MJ, et al. Integrated
680 analysis of multimodal single-cell data. Cell 2021;184:3573-3587 e3529.
681 21. Niu L, Geyer PE, Wewer Albrechtsen NJ, Gluud LL, Santos A, Doll S, Treit PV, et al. Plasma
682 proteome profiling discovers novel proteins associated with non-alcoholic fatty liver disease. Mol Syst
683 Biol 2019;15:e8793.
684 22. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat
685 Methods 2012;9:671-675.
686 23. Shu J, Dolman GE, Duan J, Qiu G, Ilyas M. Statistical colour models: an automated digital image
687 analysis method for quantification of histological biomarkers. Biomed Eng Online 2016;15:46.
688 24. Brunt EM, Janney CG, Di Bisceglie AM, Neuschwander-Tetri BA, Bacon BR. Nonalcoholic
689 steatohepatitis: a proposal for grading and staging the histological lesions. Am J Gastroenterol
690 1999;94:2467-2474.

691 25. Kleiner DE, Brunt EM, Van Natta M, Behling C, Contos MJ, Cummings OW, Ferrell LD, et al. Design
692 and validation of a histological scoring system for nonalcoholic fatty liver disease. Hepatology
693 2005;41:1313-1321.
694 26. Karlsson M, Sjostedt E, Oksvold P, Sivertsson A, Huang J, Alvez MB, Arif M, et al. Genome-wide
695 annotation of protein-coding genes in pig. BMC Biol 2022;20:25.
696 27. Ouyang JF, Kamaraj US, Cao EY, Rackham OJL. ShinyCell: Simple and sharable visualisation of
697 single-cell gene expression data. Bioinformatics 2021.
698 28. Liu Y, Ciotti GE, Eisinger-Mathason TSK. Hypoxia and the Tumor Secretome. Adv Exp Med Biol
699 2019;1136:57-69.
700 29. Crescenzi E, Leonardi A, Pacifico F. NGAL as a Potential Target in Tumor Microenvironment. Int J
701 Mol Sci 2021;22.
702 30. Zhou X, Zhang J, Lv W, Zhao C, Xia Y, Wu Y, Zhang Q. The pleiotropic roles of adipocyte secretome
703 in remodeling breast cancer. J Exp Clin Cancer Res 2022;41:203.
704 31. Uhlen M, Karlsson MJ, Hober A, Svensson AS, Scheffel J, Kotol D, Zhong W, et al. The human
705 secretome. Sci Signal 2019;12.
706 32. Corey KE, Pitts R, Lai M, Loureiro J, Masia R, Osganian SA, Gustafson JL, et al. ADAMTSL2 protein
707 and a soluble biomarker signature identify at-risk non-alcoholic steatohepatitis and fibrosis in adults
708 with NAFLD. J Hepatol 2022;76:25-33.
709 33. Gerhard GS, Hanson A, Wilhelmsen D, Piras IS, Still CD, Chu X, Petrick AT, et al. AEBP1 expression
710 increases with severity of fibrosis in NASH and is regulated by glucose, palmitate, and miR-372-3p.
711 PLoS One 2019;14:e0219764.
712 34. Cengiz M, Yilmaz G, Ozenirler S. Serum Biglycan as a Diagnostic Marker for Non-Alcoholic
713 Steatohepatitis and Liver Fibrosis. Clin Lab 2021;67.
714 35. Hoang SA, Oseini A, Feaver RE, Cole BK, Asgharpour A, Vincent R, Siddiqui M, et al. Gene
715 Expression Predicts Histological Severity and Reveals Distinct Molecular Profiles of Nonalcoholic Fatty
716 Liver Disease. Sci Rep 2019;9:12541.
717 36. Pantano L, Agyapong G, Shen Y, Zhuo Z, Fernandez-Albert F, Rust W, Knebel D, et al. Molecular
718 characterization and cell type composition deconvolution of fibrosis in NAFLD. Sci Rep 2021;11:18045.
719 37. Brosch M, Kattler K, Herrmann A, von Schonfels W, Nordstrom K, Seehofer D, Damm G, et al.
720 Epigenomic map of human liver reveals principles of zonated morphogenic and metabolic control. Nat
721 Commun 2018;9:4150.
722 38. Suppli MP, Rigbolt KTG, Veidal SS, Heeboll S, Eriksen PL, Demant M, Bagger JI, et al. Hepatic
723 transcriptome signatures in patients with varying degrees of nonalcoholic fatty liver disease compared
724 with healthy normal-weight individuals. Am J Physiol Gastrointest Liver Physiol 2019;316:G462-G472.
725 39. Govaere O, Cockell S, Tiniakos D, Queen R, Younes R, Vacca M, Alexander L, et al. Transcriptomic
726 profiling across the nonalcoholic fatty liver disease spectrum reveals gene signatures for
727 steatohepatitis and fibrosis. Sci Transl Med 2020;12.
728 40. Kozumi K, Kodama T, Murai H, Sakane S, Govaere O, Cockell S, Motooka D, et al. Transcriptomics
729 Identify Thrombospondin-2 as a Biomarker for NASH and Advanced Liver Fibrosis. Hepatology
730 2021;74:2452-2466.
731 41. Stupnikov A, McInerney CE, Savage KI, McIntosh SA, Emmert-Streib F, Kennedy R, Salto-Tellez M,
732 et al. Robustness of differential gene expression analysis of RNA-seq. Comput Struct Biotechnol J
733 2021;19:3470-3481.

42. Shakola F, Palejev D, Ivanov I. A Framework for Comparison and Assessment of Synthetic RNA-Seq Data. Genes (Basel) 2022;13.
43. Tapper EB, Loomba R. Noninvasive imaging biomarker assessment of liver fibrosis by elastography in NAFLD. Nat Rev Gastroenterol Hepatol 2018;15:274-282.
44. Masoodi M, Gastaldelli A, Hyotylainen T, Arretxe E, Alonso C, Gaggini M, Brosnan J, et al. Metabolomics and lipidomics in NAFLD: biomarkers and non-invasive diagnostic tests. Nat Rev Gastroenterol Hepatol 2021;18:835-856.
45. Wong VW, Adams LA, de Ledinghen V, Wong GL, Sookoian S. Noninvasive biomarkers in NAFLD and NASH - current progress and future promise. Nat Rev Gastroenterol Hepatol 2018;15:461-478.
46. Panera N, Gnani D, Crudele A, Ceccarelli S, Nobili V, Alisi A. MicroRNAs as controlled systems and controllers in non-alcoholic fatty liver disease. World J Gastroenterol 2014;20:15079-15086.
47. Fang Z, Dou G, Wang L. MicroRNAs in the Pathogenesis of Nonalcoholic Fatty Liver Disease. Int J Biol Sci 2021;17:1851-1863.
48. Qian X, Zong W, Ma L, Yang Z, Chen W, Yan J, Xu J. MM-associated circular RNA downregulates microRNA-19a through methylation to suppress proliferation of pancreatic adenocarcinoma cells. Bioengineered 2022;13:9294-9300.
49. Jampoka K, Muangpaisarn P, Khongnomnan K, Treeprasertsuk S, Tangkijvanich P, Payungporn S. Serum miR-29a and miR-122 as Potential Biomarkers for Non-Alcoholic Fatty Liver Disease (NAFLD). Microrna 2018;7:215-222.
50. Akuta N, Kawamura Y, Suzuki F, Saitoh S, Arase Y, Fujiyama S, Sezaki H, et al. Analysis of association between circulating miR-122 and histopathological features of nonalcoholic fatty liver disease in patients free of hepatocellular carcinoma. BMC Gastroenterol 2016;16:141.
51. Wang W, Min L, Qiu X, Wu X, Liu C, Ma J, Zhang D, et al. Biological Function of Long Non-coding RNA (LncRNA) Xist. Front Cell Dev Biol 2021;9:645647.
52. Atanasovska B, Rensen SS, Marsman G, Shiri-Sverdlov R, Withoff S, Kuipers F, Wijmenga C, et al. Long Non-Coding RNAs Involved in Progression of Non-Alcoholic Fatty Liver Disease to Steatohepatitis. Cells 2021;10.
53. Chee D, Ng CH, Chan KE, Huang DQ, Teng M, Muthiah M. The Past, Present, and Future of Noninvasive Test in Chronic Liver Diseases. Med Clin North Am 2023;107:397-421.
54. Bozaoglu K, Attard C, Kulkarni H, Cummings N, Diego VP, Carless MA, Shields KA, et al. Plasma levels of soluble interleukin 1 receptor accessory protein are reduced in obesity. J Clin Endocrinol Metab 2014;99:3435-3443.
55. Gao R, Wang J, He X, Wang T, Zhou L, Ren Z, Yang J, et al. Comprehensive analysis of endoplasmic reticulum-related and secretome gene expression profiles in the progression of non-alcoholic fatty liver disease. Front Endocrinol (Lausanne) 2022;13:967016.
56. Reznik N, Fass D. Disulfide bond formation and redox regulation in the Golgi apparatus. FEBS Lett 2022;596:2859-2872.
57. Ledesma D, Symes S, Richards S. Advancements within Modern Machine Learning Methodology: Impacts and Prospects in Biomarker Discovery. Curr Med Chem 2021;28:6512-6531.
58. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, Khodadoust MS, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol 2019;37:773-782.

## Figure Legend

**Figure 1. RNA sequencing data processing, integration, and analysis**

A. Illustration of data processing.
B. Principal component analysis (PCA) based on the origin of datasets.
C. PCA based on normalized RNA abundance (nCount).
D. PCA based on NAS score.
E. PCA based on fibrosis scores.

**Figure 2. Integrative transcriptome and proteomics analysis to identify NAFLD biomarkers**

A. Heatmap presentation of 1021 up-regulated genes in cluster 4 associated with increasing NAS scores.
B. Heatmap presentation of 643 down-regulated genes in cluster 14 associated with increasing NAS scores.
C. Protein cluster of 148 up-regulated proteins associated with increasing NAFLD severity in PXD011839.
D. Protein cluster of 114 down-regulated proteins associated with increasing NAFLD severity in PXD011839.
E. (UP) Venn diagram showing 16 overlapping genes between up-regulated genes identified by RNA-seq, up-regulated proteins in the plasma, and secreting proteins. (DOWN) Venn diagram showing 22 overlapping genes between down-regulated genes identified by RNA-seq, down-regulated proteins in the plasma, and secreting proteins.

**Figure 3. Relationship between QSOX1/IL1RAP and NAS/fibrosis scores in integrative RNA-seq data of the human liver.**

A. Box plot of QSOX1 gene expressions grouped by NAS scores (N0, N1-4, N5-8).
B. Box plot of QSOX1 gene expressions grouped by fibrosis stages (F0, F1-2, F3-4).
C. Box plot of IL1RAP gene expressions grouped by NAS scores (N0, N1-4, N5-8).
D. Box plot of IL1RAP gene expressions grouped by fibrosis stages (F0, F1-2, F3-4).
E. Box plot of QSOX1/IL1RAP gene expression ratio grouped by NAS scores (N0, N1-4, N5-8).
F. Box plot of QSOX1/ IL1RAP gene expression ratio grouped by fibrosis stages (F0, F1-2, F3-4).
Statistical testing was performed using the Wilcoxon rank sum test, with p-values shown in the plot.

**Figure 4. Comparison of plasma protein QSOX1/IL1RAP between healthy individuals and NAFLD at various stages.**

A-B. Box plots of plasma QSOX1 (A) and IL1RAP (B) protein levels in healthy individuals, NAFLD patients, and cirrhosis patients quantified by mass spectrometry in the NAFLD proteomics cohort.

C. Box plots of plasma QSOX1 and IL1RAP protein ratios in healthy controls, NAFLD patients, and cirrhosis patients quantified by mass spectrometry in the NAFLD proteomics cohort.

D-E. Box plots of plasma QSOX1 (D) and IL1RAP (E) protein levels in healthy controls and NAFLD groups (pg/ml) measured by ELISA.

F. Box plot of plasma QSOX1 and IL1RAP protein ratio.

Statistical testing was performed using the Wilcoxon rank sum test, with p-values shown in the plot.

**Figure 5. Quantification of liver QSOX1 and IL1RAP levels in NAFLD patients by IHC.**

A. Representative IHC images of QSOX1 and IL1RAP in liver biopsies from mild NAFLD patients (N0-4, F0-2) and severe NAFLD patients (N5-8, F3-4).

B. The integrated density of QSOX1 IHC. Box plots showing the log10 value.

C. The integrated density of IL1RAP IHC. Box plots showing the log10 value.

D. Box plot of QSOX1 and IL1RAP ratio of the integrated density quantified with IHC.

Statistical testing was performed using t-test, with p-values shown in the plot.

**Table 1**

**Table 1: Characteristics of the populations studied**

A total of 625 human liver samples of the full histological range from normal, NAFL, NASH to cirrhosis with the NAS (N) and fibrosis (F) scores provided in the database or original articles.

| Sample distribution of NAS and FIB | F0 | F1 | F2 | F3 | F4 | Age (yrs) | BMI (kg/m2) |
|---|---|---|---|---|---|---|---|
| | | | | | | mean ± S.D. | mean ± S.D. |
| N0 (n) | 81 | 1 | 1 | - | 1 | 46.4 ± 10.0 | 35.4 ± 9.1 |
| N1 (n) | 29 | 9 | 1 | - | 1 | 42.8 ± 12.7 | 36.1 ± 6.6 |
| N2 (n) | 27 | 10 | 1 | - | 1 | 51.9 ± 5.8 | 35.1 ± 6.0 |
| N3 (n) | 46 | 72 | 9 | 5 | 3 | 49.8 ± 9.4 | 34.6 ± 8.9 |
| N4 (n) | 30 | 19 | 20 | 15 | 3 | 49.6 ± 9.5 | 33.9 ± 4.6 |
| N5 (n) | 7 | 41 | 75 | 20 | 5 | 52.4 ± 11.7 | 32.5 ± 5.5 |
| N6 (n) | 1 | 18 | 17 | 18 | 3 | 53.2 ± 8.5 | 34.5 ± 5.6 |
| N7 (n) | - | 1 | 14 | 11 | 1 | 49.4 ± 9.8 | 36.3 ± 7.2 |
| N8 (n) | - | - | 2 | 4 | 2 | 54 ± 0.0 | 31.3± 0.0 |
| Age (yrs) mean ± S.D. | 46.6 ± 9.6 | 49.0 ± 10.6 | 53.2 ± 11.3 | 54.7 ± 6.0 | 54.8 ± 5.4 | | |
| BMI(kg/m2) mean ± S.D. | 37.1± 8.3 | 32.4 ± 5.6 | 33.1 ± 7.0 | 32.5 ± 2.9 | 32.6 ± 2.2 | | |
| Gender n(male/female) | 221(105/116)* | 171 (108/63)* | 140 (64/76)* | 73 (35/38)* | 20 (11/9)* | | |

**Table 1.** *The ratio of gender was estimated according to the gender ratio in the original articles. BMI, body mass index; N, NAS score; F, Fibrosis score.

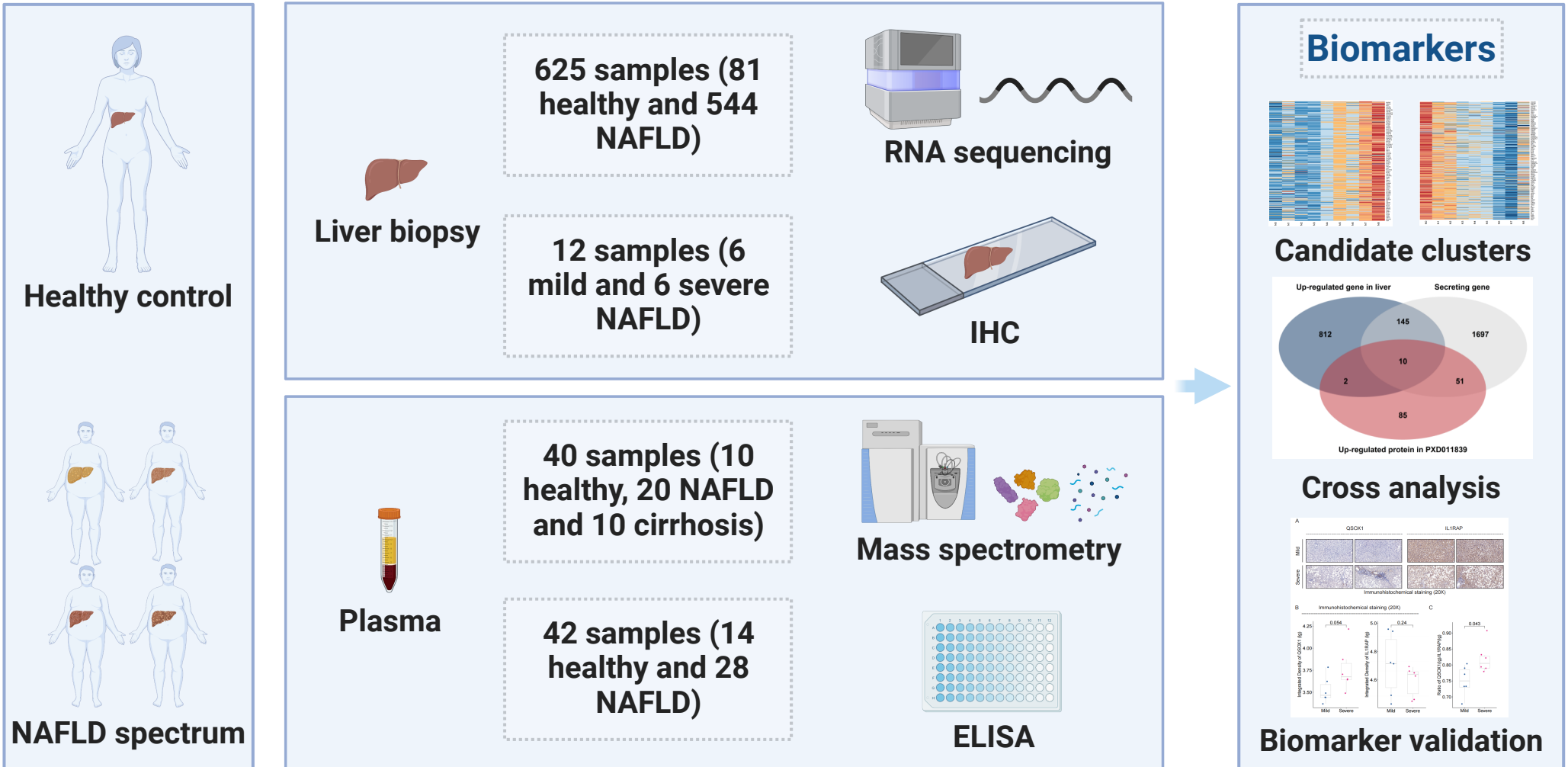

Healthy control
NAFLD spectrum
Liver biopsy
625 samples (81 healthy and 544 NAFLD)
12 samples (6 mild and 6 severe NAFLD)
RNA sequencing
IHC
Plasma
40 samples (10 healthy, 20 NAFLD and 10 cirrhosis)
42 samples (14 healthy and 28 NAFLD)
Mass spectrometry
ELISA
Biomarkers
Candidate clusters
Up-regulated gene in liver
Secreting gene
812
145
1697
10
2
51
85
Up-regulated protein in PXD011839
Cross analysis
Biomarker validation
Created with BioRender.com

**Figure 1**

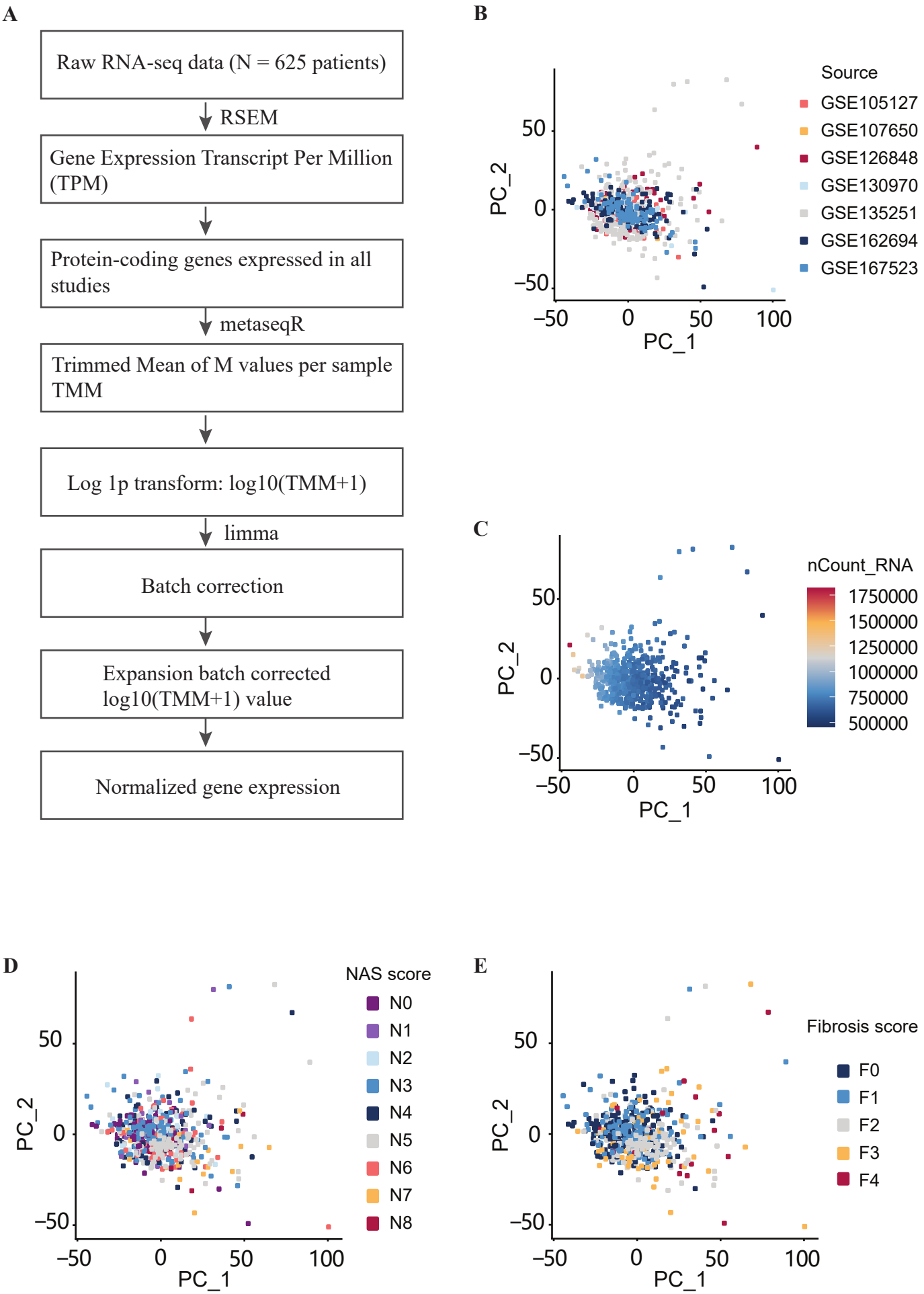

A
Raw RNA-seq data (N = 625 patients)
RSEM
Gene Expression Transcript Per Million (TPM)
Protein-coding genes expressed in all studies
metaseqR
Trimmed Mean of M values per sample TMM
Log 1p transform: log10(TMM+1)
limma
Batch correction
Expansion batch corrected log10(TMM+1) value
Normalized gene expression
B
Source
GSE105127
GSE107650
GSE126848
GSE130970
GSE135251
GSE162694
GSE167523
PC_2
PC_1
C
nCount_RNA
1750000
1500000
1250000
1000000
750000
500000
PC_2
PC_1
D
NAS score
N0
N1
N2
N3
N4
N5
N6
N7
N8
PC_2
PC_1
E
Fibrosis score
F0
F1
F2
F3
F4
PC_2
PC_1

**Figure 2**

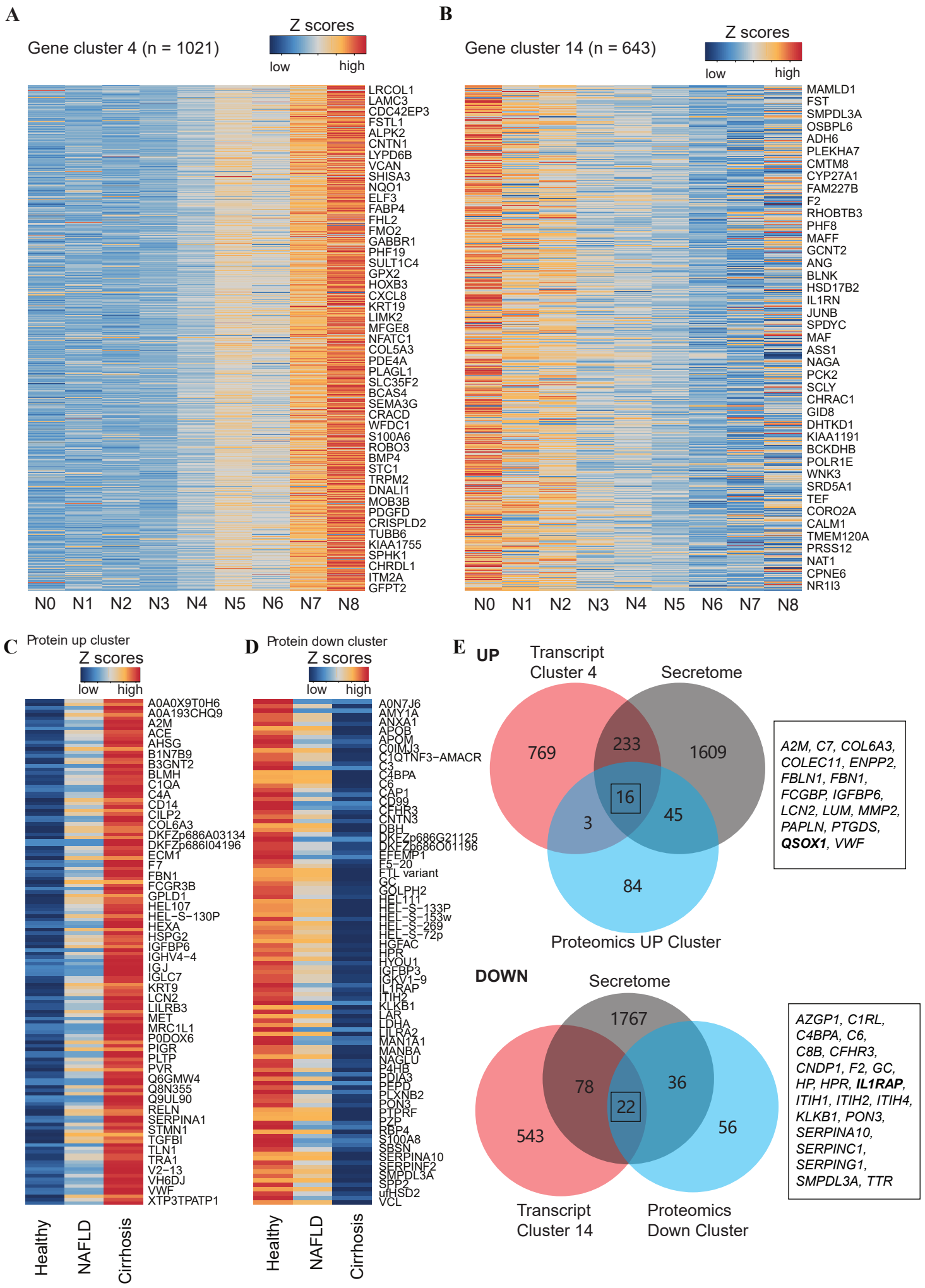

A
Gene cluster 4 (n = 1021)
Z scores
low
high
LRCOL1
LAMC3
CDC42EP3
FSTL1
ALPK2
CNTN1
LYPD6B
VCAN
SHISA3
NQO1
ELF3
FABP4
FHL2
FMO2
GABBR1
PHF19
SULT1C4
GPX2
HOXB3
CXCL8
KRT19
LIMK2
MFGE8
NFATC1
COL5A3
PDE4A
PLAGL1
SLC35F2
BCAS4
SEMA3G
CRACD
WFDC1
S100A6
ROBO3
BMP4
STC1
TRPM2
DNALI1
MOB3B
PDGFD
CRISPLD2
TUBB6
KIAA1755
SPHK1
CHRDL1
ITM2A
GFPT2
N0 N1 N2 N3 N4 N5 N6 N7 N8
B
Gene cluster 14 (n = 643)
Z scores
low
high
MAMLD1
FST
SMPDL3A
OSBPL6
ADH6
PLEKHA7
CMTM8
CYP27A1
FAM227B
F2
RHOBTB3
PHF8
MAFF
GCNT2
ANG
BLNK
HSD17B2
IL1RN
JUNB
SPDYC
MAF
ASS1
NAGA
PCK2
SCLY
CHRAC1
GID8
DHTKD1
KIAA1191
BCKDHB
POLR1E
WNK3
SRD5A1
TEF
CORO2A
CALM1
TMEM120A
PRSS12
NAT1
CPNE6
NR1I3
N0 N1 N2 N3 N4 N5 N6 N7 N8
C
Protein up cluster
Z scores
low
high
A0A0X9T0H6
A0A193CHQ9
A2M
ACE
AHSG
B1N7B9
B3GNT2
BLMH
C1QA
C4A
CD14
CILP2
COL6A3
DKFZp686A03134
DKFZp686I04196
ECM1
F7
FBN1
FCGR3B
GPLD1
HEL107
HEL-S-130P
HEXA
HSPG2
IGFBP6
IGHV4-4
IGJ
IGLC7
KRT9
LCN2
LILRB3
MET
MRC1L1
P0DOX6
PIGR
PLTP
PVR
Q6GMW4
Q8N355
Q9UL90
RELN
SERPINA1
STMN1
TGFBI
TLN1
TRA1
V2-13
VH6DJ
VWF
XTP3TPATP1
Healthy
NAFLD
Cirrhosis
D
Protein down cluster
Z scores
low
high
A0N7J6
AMY1A
ANXA1
APOB
APOM
C0IMJ3
C1QTNF3-AMACR
C3
C4BPA
C6
CAP1
CD99
CFHR3
CNTN3
DBH
DKFZp686G21125
DKFZp686O01196
EFEMP1
F5-20
FTL variant
GC
GOLPH2
HEL111
HEL-S-133P
HEL-S-153w
HEL-S-269
HEL-S-72p
HGFAC
HPR
HYOU1
IGFBP3
IGKV1-9
IL1RAP
ITIH2
KLKB1
LAR
LDHA
LILRA2
MAN1A1
MANBA
NAGLU
P4HB
PDIA3
PEPD
PLXNB2
PON3
PTPRF
PZP
RBP4
S100A8
SBSN
SERPINA10
SERPINF2
SMPDL3A
SPP2
ufHSD2
VCL
Healthy
NAFLD
Cirrhosis
E
UP
Transcript Cluster 4
Secretome
769
233
1609
16
3
45
84
Proteomics UP Cluster
A2M, C7, COL6A3, COLEC11, ENPP2, FBLN1, FBN1, FCGBP, IGFBP6, LCN2, LUM, MMP2, PAPLN, PTGDS, QSOX1, VWF
DOWN
Secretome
1767
78
36
22
543
56
Transcript Cluster 14
Proteomics Down Cluster
AZGP1, C1RL, C4BPA, C6, C8B, CFHR3, CNDP1, F2, GC, HP, HPR, IL1RAP, ITIH1, ITIH2, ITIH4, KLKB1, PON3, SERPINA10, SERPINC1, SERPING1, SMPDL3A, TTR

**Figure 3**

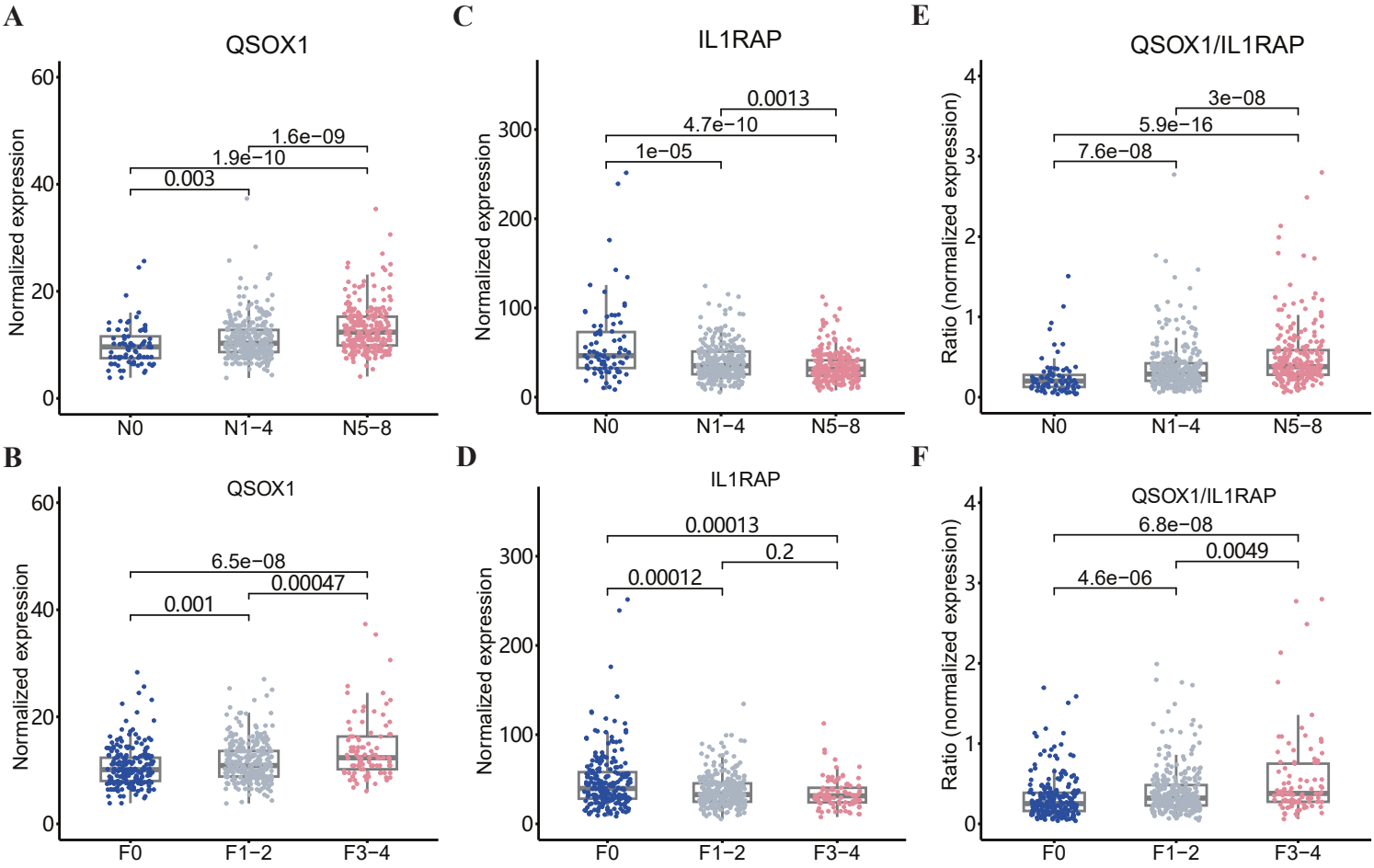

A
QSOX1
Normalized expression
0.003
1.9e−10
1.6e−09
N0
N1−4
N5−8
B
QSOX1
Normalized expression
0.001
6.5e−08
0.00047
F0
F1−2
F3−4
C
IL1RAP
Normalized expression
1e−05
4.7e−10
0.0013
N0
N1−4
N5−8
D
IL1RAP
Normalized expression
0.00012
0.00013
0.2
F0
F1−2
F3−4
E
QSOX1/IL1RAP
Ratio (normalized expression)
7.6e−08
5.9e−16
3e−08
N0
N1−4
N5−8
F
QSOX1/IL1RAP
Ratio (normalized expression)
4.6e−06
6.8e−08
0.0049
F0
F1−2
F3−4

**Figure 4**

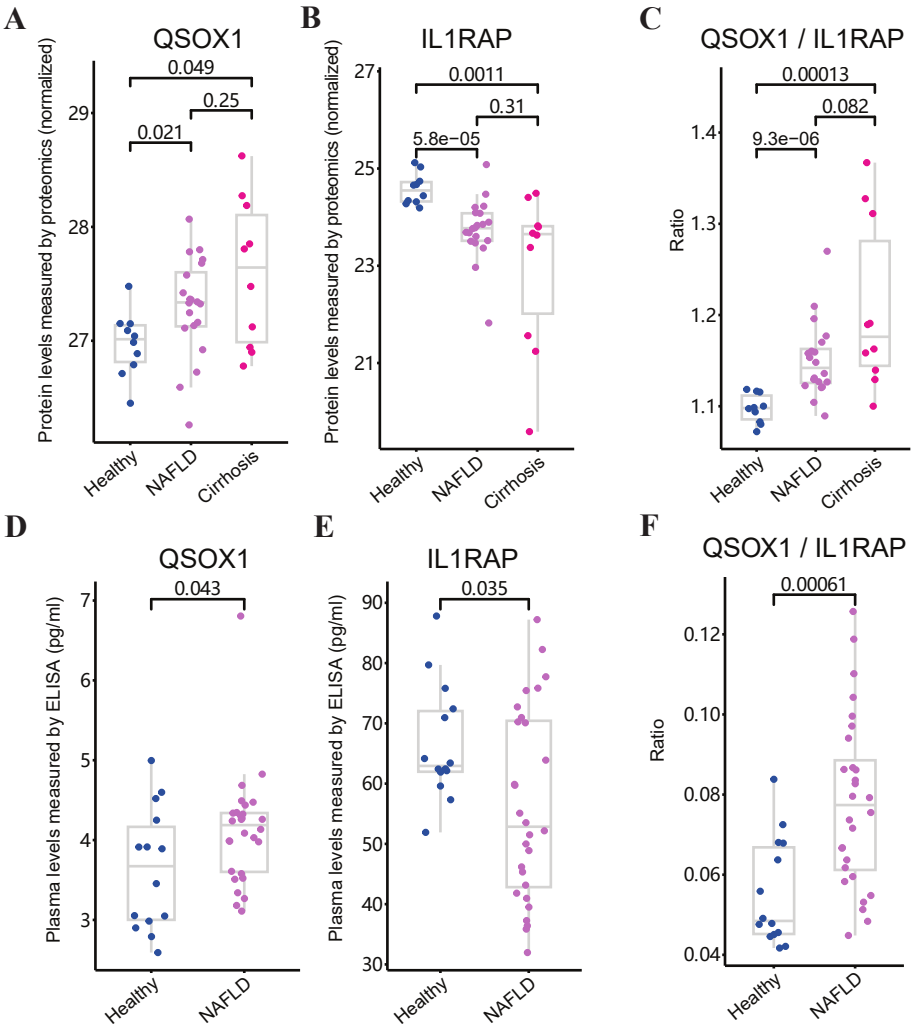
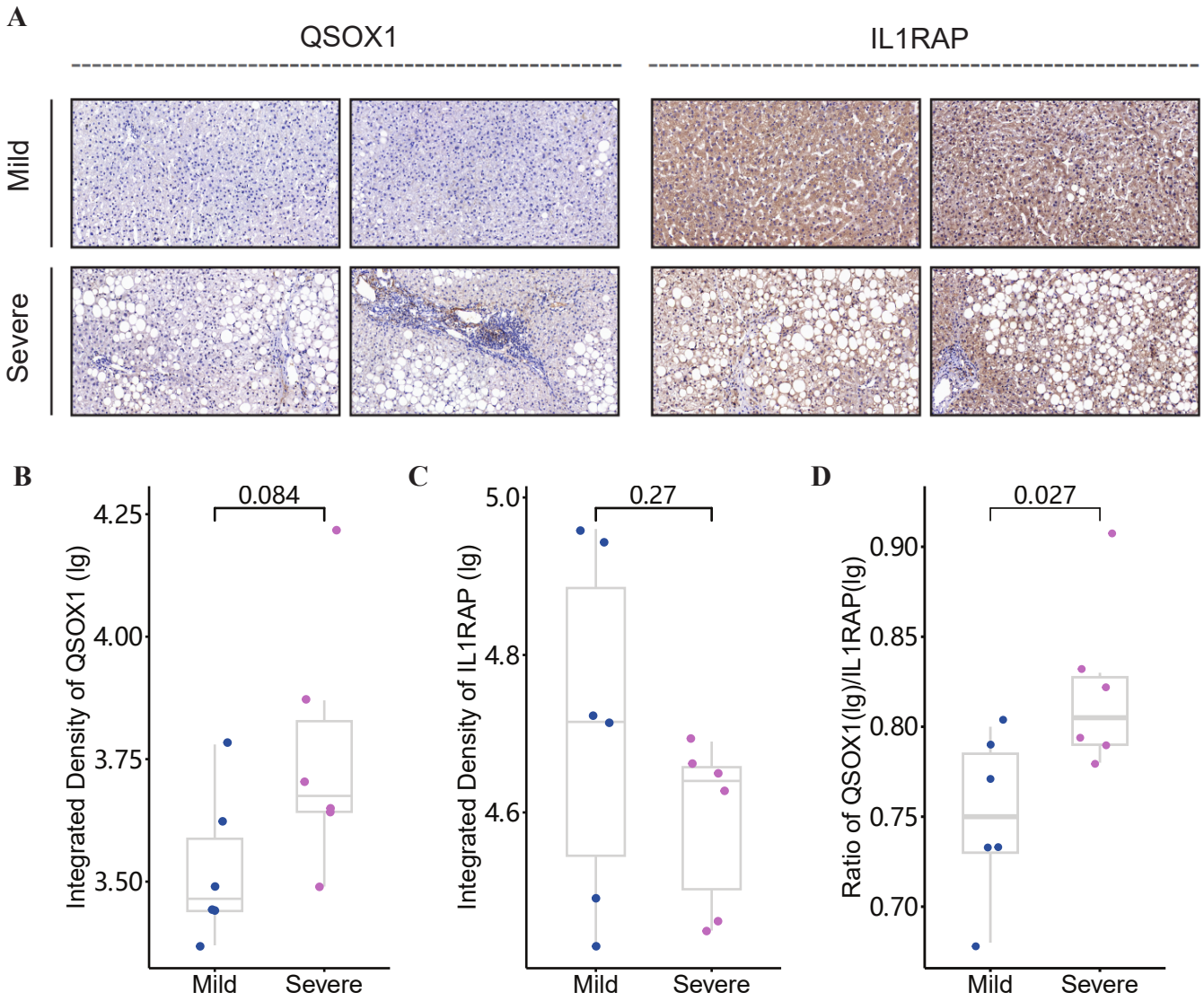

A
QSOX1
Protein levels measured by proteomics (normalized)
0.049
0.25
0.021
29
28
27
Healthy
NAFLD
Cirrhosis
B
IL1RAP
Protein levels measured by proteomics (normalized)
0.0011
0.31
5.8e−05
27
25
23
21
Healthy
NAFLD
Cirrhosis
C
QSOX1 / IL1RAP
Ratio
0.00013
0.082
9.3e−06
1.4
1.3
1.2
1.1
Healthy
NAFLD
Cirrhosis
D
QSOX1
Plasma levels measured by ELISA (pg/ml)
0.043
7
6
5
4
3
Healthy
NAFLD
E
IL1RAP
Plasma levels measured by ELISA (pg/ml)
0.035
90
80
70
60
50
40
30
Healthy
NAFLD
F
QSOX1 / IL1RAP
Ratio
0.00061
0.12
0.10
0.08
0.06
0.04
Healthy
NAFLD

**Figure 5**


A
QSOX1
IL1RAP
Mild
Severe
B
0.084
Integrated Density of QSOX1 (Ig)
4.25
4.00
3.75
3.50
Mild
Severe
C
0.27
Integrated Density of IL1RAP (Ig)
5.0
4.8
4.6
Mild
Severe
D
0.027
Ratio of QSOX1(Ig)/IL1RAP(Ig)
0.90
0.85
0.80
0.75
0.70
Mild
Severe