A gene-based clustering 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.
- An unsupervised method for clustering genes based on the similarity of gene expression trajectory across all samples 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: An 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.

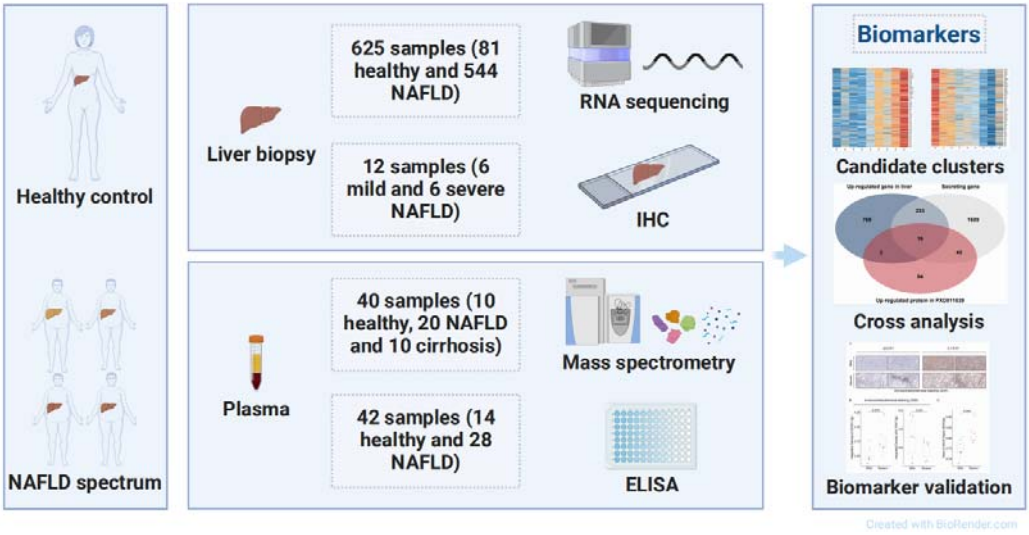
Lay Summary

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.

Keywords

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

Graphical abstract



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

119 overlapping gene expression changes across the entire spectrum of NAFLD. Such an
120 approach would offer a comprehensive representation of NAFLD's complexity and
121 enhance our understanding of its progression. In recent years, advancements in
122 technologies such as RNA sequencing (RNA-seq), single-cell analysis, and spatial
123 transcriptomics have provided deeper insights into the molecular and cellular
124 processes involved in NAFLD progression (13-15). Large-scale profiling efforts,
125 combined with targeted validation approaches, have led to the discovery of potential
126 biomarkers (16, 17). However, the majority of available RNA-seq data are derived
127 from smaller cohorts of NAFLD patients, which limits the comprehensive
128 understanding of NAFLD severity.
129
130 In this study, modularity optimization methods were utilized to cluster genes by
131 employing a graph-based strategy, taking into account the gene expression patterns
132 throughout the progression of NAFLD. We propose that integrating and analyzing
133 these datasets with the unbiased gene-based profiling strategy will provide further
134 insights into the molecular progression of NAFLD and the identification of biomarkers
135 associated with NAFLD severity. In the present study, we identified over 300 NAFLD
136 biomarkers by integrating and analyzing RNA-seq data from 625 liver
137 samples, including their NAFLD activity scores (NAS) and fibrosis scores,
138 along with public proteomics data. We further validated these findings in two
139 independent NAFLD cohorts, demonstrating the potential of the QSOX1/IL1RAP ratio
140 as a non-invasive biomarker for diagnosing NAFLD and assessing its severity.
141

142 ## Materials and methods

143

144 ***Data Collection***

145 Genome-wide RNA-seq data of human NAFLD and associated healthy controls were
146 collected from the NCBI GEO (https://www.ncbi.nlm.nih.gov/gds, access date until
147 May 2022). Only datasets that provided detailed NAS and fibrosis scores were
148 included for further investigation, including seven datasets (GSE105127(18),

149 GSE107650(19), GSE126848(20), GSE130970(21), GSE135251(22, 23), GSE162694(24),
150 and GSE167523(25). (**Supplementary Table 1**)
151
152 ***Data Normalization***
153 The SRA-formatted data were converted into FASTQ format using 'SraToolkit'
154 (sratoolkit.2.8.2-1-centos_linux64) (https://github.com/ncbi/sra-tools). Sequencing
155 reads were aligned to the hg19 UCSC RNA sequences Genome Reference Consortium
156 Human Build 37 (GRCh37) using 'bowtie2' (bowtie2-2.2.5)
157 (https://rnnh.github.io/bioinfo-notebook/docs/bowtie2.html). Only protein-coding
158 transcripts were considered, and Transcript Per Million (TPM) values were obtained
159 by transforming the mapped transcript reads using 'RSEM' (rsem-1.2.12)
160 (https://github.com/deweylab/RSEM). Then, TPM values were then subjected to
161 Trimmed Mean of M-values (TMM) normalization across all samples using
162 'metaseqR ' (metaseqR 1.12.2) (26). The data from various sources involved in this
163 study were integrated and log1p-transformed, followed by batch correction using the
164 'removeBatchEffect' function in the R package 'limma' (limma 3.54.2)
165 (https://kasperdanielhansen.github.io/genbioconductor/html/limma.html) (27).
166 Subsequently, the data were expanded (10^x) for further analysis (**Figure 1A**).
167
168 ***RNA-seq Data Analysis and Database Construction***
169 After normalization and batch correction, the RNA-seq data were subjected to
170 Principal Components Analysis (PCA) and unsupervised clustering using the R
171 package 'Seurat' (Seurat-4.3.0)
172 (https://satijalab.org/seurat/articles/get_started.html). We utilized the
173 "LogNormalize" method for global-scaling normalization, which normalized the
174 feature expression measurements across different samples for each gene by the total
175 expression. The normalized values were multiplied by a scale factor (default:
176 10,000) and log1p-transformed. Subsequently, scaling was applied to the identified
177 variable features (default: 2,000). PCA was then performed on the scaled data, with a
178 default setting of computing and storing 50 Principal Components (PCs). To cluster

179 the genes, we employed modularity optimization techniques using a graph-based
180 clustering approach. The dimensions of reduction were set to 1:20, and
181 the resolution parameter was set to 2.3 (28).
182
183 To show the gene expression variation during the development of NAFLD associated
184 with both NAS and fibrosis scores, we generated an RNA-seq database using
185 'ShinyCell' (https://github.com/SGDDNB/ShinyCell). This database was deployed at
186 https://dreamapp.biomed.au.dk/NAFLD/.
187
188 ***Proteomics Data Collection and Analysis***
189 The proteomics cohort dataset (PXD011839) includes 10 healthy controls, 10 NAFLD
190 patients with normal glucose tolerance (NAFLD_ngt), 10 NAFLD patients with type 2
191 diabetes (NAFLD_T2D), and 10 NAFLD patients with cirrhosis (29). We performed the
192 statistical analysis using R-4.3.0 on the dataset (EV1, tab4)
193 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6396370/bin/MSB-15-e8793-s003.
194 xlsx). The NAFLD_ngt and NAFLD_T2D groups were merged into a single NAFLD group,
195 resulting in three groups: healthy controls, NAFLD, and cirrhosis.
196
197 Similar to RNA-seq analysis, we used 'Seurat' (Seurat-4.3.0) and employed the
198 "LogNormalize" method for global-scaling normalization. This method normalized
199 the feature expression measurements across different samples for each protein by
200 the total expression. The normalized values were multiplied by a scale factor (default:
201 10,000) and log1p-transformed. Subsequently, scaling was applied to the identified
202 variable features (default: 2,000). PCA was performed on the scaled data, with a total
203 of 39 Principal Components (PCs) computed and stored. Additionally, we calculated
204 the log fold-change of the average expression between two groups (avg_log2FC) by
205 comparing the health and NAFLD groups, as well as the NAFLD and cirrhosis groups.
206 By setting avg_log2FC > 0, we selected up- and down-regulated proteins associated
207 with increasing severity of NAFLD.
208

209 ***Validation of QSOX1 and IL1RAP as Biomarkers in NAFLD***

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

217

218 ***Human Samples***

219 A total of 28 ultrasound-proven adult NAFLD patients, including NAFLD_ngt and
220 NAFLD_T2D, and 14 healthy controls were enrolled from SZTCMH. Other diagnoses
221 and etiologies, such as excessive alcohol consumption, viral hepatitis, autoimmune
222 liver disease, and the use of steatogenic compounds, were excluded. Archived
223 plasma samples were collected between October and December 2022. Informed
224 consent was obtained from the healthy subjects and NAFLD patients, following the
225 approved clinical protocols of the Ethical Committee of SZTCMH. Clinical information,
226 including body mass index (BMI) and standard biochemistry (liver, kidney,
227 hematology) with metabolic profiling (glucose, insulin, lipids), was collected.
228 Fibroscan with controlled attenuation parameter (CAP) values was performed to
229 assess fibrosis and steatosis. Clinical information for the healthy controls and NAFLD
230 patients can be found in **Supplementary Table 5**.

231

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

239

240 ***ELISA***

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

250

251 ***Immunohistochemisgtry Assay***

252 We examined the association of QSOX1 and IL1RAP with human NAFLD severity by
253 performing IHC on formalin-fixed and paraffin-embedded liver sections from 6 mild
254 NAFLD patients and 6 severe NAFLD patients. The 3 μm-thick paraffin sections were
255 deparaffinized and rehydrated with distilled water. Antigen retrieval was carried out
256 using pH 9.0 EDTA buffer, followed by 20 minutes of boiling and washing with 1X PBS.
257 Subsequently, the slides were blocked with 1% bovine serum albumin in 1X PBS for
258 15 minutes and then incubated overnight at 4℃ with QSOX1 (Rabbit anti-human,
259 Catalog No. Ab235444, Lot No. GR3386311-2 from Abcam) or IL1RAP (Rabbit
260 anti-human, Catalog No.35605, Lot No. 4926 from Sabbiotech) antibodies at a
261 concentration of 20 μg/ml. The following day, the slides were washed with 1X PBS
262 and incubated with Goat anti-rabbit IgG H&L (Catalog No. Ab205718, Lot No.
263 ab205718 from Abcam) for 15 minutes at room temperature, followed by another
264 wash with 1X PBS. The images were captured using a light microscope and
265 3DHISTECH digital scanner (https://www.3dhistech.com/).

266

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

268 potential bias, we randomly selected five locations of the same size from each
269 sample at 20x magnification using 3DHISTECH CaseViewer_2.4
270 (https://www.3dhistech.com/solutions/caseviewer/). Using 'ImagineJ', we applied
271 the "Colour Deconvolution" tool with vectors=[H DAB]; followed by selecting the
272 Colour_2 pictures and running "8-bit". Standard thresholds were used (QSOX1:
273 setThreshold (60, 230), IL1RAP: setThreshold (94, 214)) (30, 31). The average
274 integrated density from the five sites was calculated and used as the integrated
275 density value for each sample, which was then subjected to statistical analysis.
276
277 ***Statistical analysis***
278 The significance for all statistical tests was two-sided, with $P < 0.05$. All data analysis
279 was presented in the plots using R-4.3.0, and MedCalc was used to calculate the
280 AUROC, sensitivity, specificity, optimal cutoff value, and sample size.

281 ## Results

282 **Overview of RNA-seq data and NAFLD patient cohorts**

283 After applying stringent filtering criteria based on the availability of histological NAS
284 and fibrosis scores, five datasets including GSE115193 (32), GSE134422 (33),
285 GSE135448 (34), GSE160016 (35), and GSE164441 (36) were excluded from the
286 analysis, while seven datasets (GSE105127 (18), GSE107650 (19), GSE126848 (20),
287 GSE130970 (21), GSE135251 (22, 23), GSE162694 (24), and GSE167523 (25)) were
288 included. These datasets collectively comprise 81 healthy controls (including healthy
289 obsese individuals) (N0F0) and 544 NAFLD patients. The severity of NAFLD patients
290 was classified based on the NAS score (ranging from N1 to N8) and the fibrosis score
291 (ranging from F0 to F4) using the scoring systems proposed by Brunt (37) and Kleiner
292 (38), respectively (**Table 1, Supplementary Table 1**). A positive correlation was
293 observed between NAS and fibrosis scores (Pearson R = 0.64, P = 4.94E-74, **Table 1**),
294 indicating an association with NAFLD severity.

295

296 **Normalization and Integration of RNA-seq Data**

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

306

307 **Unsupervised Gene Clustering Identifies Clusters of Genes Associated with NAFLD**
308 **Severity.**

309 To identify genes associated with NAFLD severity, we utilized a previously developed

310 unsupervised gene clustering method based on the similarity of gene expression
311 patterns across each sample (39). We employed gene clustering, grouping genes
312 according to their expression patterns during the progression of NAFLD. By setting a
313 resolution of 2.3, we identified a total of 37 gene clusters (**Supplementary Figure S1**).
314 Notably, cluster 4, consisting of 1021 genes, consistently exhibited increased
315 expression with higher NAS and fibrosis scores (**Figure 2A, Supplementary Figure**
316 **S2A**). Conversely, cluster 14, comprising 643 genes, showed decreased expression
317 with increasing NAS and fibrosis scores (**Figure 2B**, **Supplementary Figure S2B**). As
318 illustrated in figures (**Figure 2A, B**, **Supplementary Figure S2A, B, C, D**), this approach
319 efficiently clustered genes into distinct groups based on their expression patterns
320 across NAFLD severity stages. It offers a more structured depiction of gene
321 expression variations, enabling a deeper understanding of NAFLD's molecular
322 pathogenesis. Through visualizing and categorizing these gene expression changes,
323 we can acquire a more comprehensive insight into the underlying mechanisms and
324 factors that propel NAFLD progression.
325
326 To explore the biological functions of these gene clusters, we performed Gene
327 Ontology (GO) analysis using the R package 'ClusterProfiler' (ClusterProfiler-4.8.0).
328 Specifically, we focused on cluster 4, which consisted of up-regulated genes. The GO
329 analysis revealed significant enrichment of genes involved in the fibrosis-related
330 process, such as extracellular matrix (ECM) organization (p.adjust = 9.77E-34),
331 extracellular structure organization (p.adjust = 9.77E-34), external encapsulating
332 structure organization (p.adjust = 1.09E-33), and cell-substrate adhesion (p.adjust =
333 3.35E-17). Notably, the expression of multiple genes involved in the ECM processes,
334 such as *COL5A3, FBLN5, SPINT2, COL1A1, COL1A2, COL3A1, COL4A1, COL4A4,*
335 *COL12A1, COL15A1*, and *COL16A1*, showed a gradual up-regulation during the
336 progression of NAFLD (**Figure 2A, Supplementary Figure S3**).
337
338 In contrast, cluster 14, which displayed a reverse correlation with NAS and fibrosis
339 scores, was significantly enriched in metabolic processes, indicating an association

340 between NAFLD progression and attenuated liver metabolism. The down-regulated
341 genes in this cluster were particularly enriched in processes such as organic acid
342 catabolic process (p.adjust = 2.57E-22), carboxylic acid catabolic process (p.adjust =
343 2.57E-22), small molecule catabolic process (p.adjust = 5.28E-22), alpha-amino acid
344 metabolic process (p.adjust = 6.03E-20), fatty acid metabolic process (p.adjust =
345 2.60E-10), and alcohol metabolic process (p.adjust = 1.35E-09). Notably, genes
346 encoding enzymes of the Cytochrome P450 superfamily, including *CYP1A2, CYP2C19,*
347 *CYP2J2, CYP2E1, CYP4A11, CYP4A22, CYP4F11, CYP2C8,* and *CYP3A4*, were
348 down-regulated with increasing NAFLD severity (**Figure 2B**, **Supplementary Figure**
349 **S4**). For a comprehensive list of all enriched GO terms for genes in cluster 4 and 14,
350 please refer to **Supplementary Table 2**.
351
352 In addition, we developed a NAFLD gene expression database (NAFLD-DB) to
353 facilitate the exploration and comparison of all identified protein-coding genes based
354 on NAFLD severity. The NAFLD-DB (https://dreamapp.biomed.au.dk/NAFLD/) was
355 constructed using the ShinyCell framework (40), which was specifically designed for
356 convenient exploration and sharing of single-cell transcriptome data.
357
358 **Identification of Candidate Diagnostic Biomarkers**
359 We employed an additional complementary strategy to further refine our list of
360 candidate genes. In this approach, we first analyzed the plasma protein levels from a
361 NAFLD cohort in a proteomics dataset (PXD011839) (29). We selected proteins that
362 showed positive correlations with increasing NAFLD severity (avg_log2FC > 0) and
363 proteins that showed negative correlations. As a result, we identified 148
364 up-regulated proteins and 114 down-regulated proteins associated with increasing
365 NAFLD severity (**Figure 2C, D**).
366
367 The secretome, which consists of secreted proteins, has emerged as a valuable
368 resource for disease diagnostics (41-43). In our study, we aim to identify potential
369 diagnostic markers among the candidate genes, by comparing our gene clusters with

370 the secretome database from the Human Protein Atlas (44). This cross-analysis
371 revealed a total of 349 genes encoding secreted proteins, with 249 genes showing
372 up-regulation and 100 genes showing down-regulation (**Figure 2E**). Notably, our
373 approach successfully identified a comprehensive list of previously known NAFLD
374 diagnostic and prognostic markers, including *ADAMTSL2* (45), *AEBP1* (46), and *BGN*
375 (47) (**Supplementary Table 3**), further validating the effectiveness of our approach.
376
377 Next, we intersected the protein-encoding genes of these proteins with the
378 secretome genes and the candidate genes generated from our RNA-seq analysis.
379 Through this cross-comparison, we identified 16 up-regulated secreting genes (*A2M,*
380 *C7, COL6A3, COLEC11, ENPP2, FBLN1, FBN1, FCGBP, IGFBP6, LCN2, LUM, MMP2,*
381 *PAPLN, PTGDS, QSOX1, VWF*) and 22 down-regulated secreting genes (*AZGP1, C1RL,*
382 *C4BPA, C6, C8B, CFHR3, CNDP1, F2, GC, HP, HPR, IL1RAP, ITIH1, ITIH2, ITIH4, KLKB1,*
383 *PON3, SERPINA10, SERPINC1, SERPING1, SMPDL3A, TTR*) associated with increasing
384 NAFLD severity in both the RNA-seq and proteomics data (**Supplementary Figure S5**).
385
386 **QSOX1 and IL1RAP are promising biomarkers for NAFLD severity**
387 To demonstrate the applicability of our NAFLD-DB and validate the association of
388 differential gene expression with increasing NAFLD severity, we selected two
389 representative genes, *QSOX1* and *IL1RAP,* which showed positive and negative
390 correlations with increasing NAFLD severity, and their roles as biomarkers were
391 under explored as compared to other NAFLD biomarkers (**Supplementary Figure S5**).
392 We examined their expression levels in comparison to patients with a NAS or fibrosis
393 score of 0 (N0 or F0). The expression of *QSOX1* was significantly correlated with the
394 severity of NAFLD compared to N0 or F0 patients: N1-4 ($p = 0.003$), N5-8 ($p =$
395 $1.9E-10$), F1-2 ($p = 0.001$), F3-4 ($p = 6.5E-8$) (**Figure 3A, B**). On the other hand, *IL1RAP*
396 expression was significantly lower in patients with increased NAFLD severity
397 compared to N0 or F0: N1-4 ($p = 1E-5$), N5-8 ($p = 4.7E-10$), F1-2 ($p = 0.00012$), F3-4 (p
398 $= 0.00013$) (**Figure 3C, D**).
399

400 Since *QSOX1* and *IL1RAP* exhibited opposite correlations with NAFLD severity, we
401 further explored whether the ratio of QSOX1/IL1RAP could better distinguish
402 between patient groups. Our results showed that compared to N0 or F0 patients, the
403 ratio of QSOX1 to IL1RAP mRNA levels showed even greater separation: N1-4 (p =
404 7.6E-8), N5-8 (5.9E-16), F1-2 (4.6E-6), F3-4 (6.8E-8) (**Figure 3E, F**). These findings
405 suggest that the QSOX1/IL1RAP ratio has the potential as a biomarker for diagnosing
406 NAFLD severity.
407

408 **Validation of Plasma QSOX1/IL1RAP Levels as Biomarkers for NAFLD Severity with**
409 **NAFLD Proteomics Cohort**

410 To further validate the potential of QSOX1 and IL1RAP as biomarkers for NAFLD
411 severity, we analyzed the plasma levels of QSOX1 and IL1RAP in a NAFLD proteomics
412 cohort (PXD011839) previously conducted by Niu L and colleagues (29). Consistent
413 with our liver RNA profiling results in livers, the analysis of plasma proteomics data
414 from this independent NAFLD cohort showed a significant increase in plasma QSOX1
415 levels in patients with NAFLD (Wilcoxon rank sum test, p = 0.021) and cirrhosis (p =
416 0.049) compared to healthy controls (**Figure 4A**). Conversely, IL1RAP levels were
417 significantly reduced in patients with NAFLD (p = 5.8E-5) and cirrhosis (p = 0.0011)
418 (**Figure 4B**). Moreover, when considering the combined marker of plasma
419 QSOX1/IL1RAP ratio, it demonstrated even greater significance in distinguishing
420 NAFLD (p = 9.3E-6) and cirrhosis (p = 0.00013) patients from the control group,
421 compared to using QSOX1 or IL1RAP alone (**Figure 4C**).
422

423 To access the diagnostic sensitivity and specificity of QSOX1, IL1RAP, and their ratio
424 for NAFLD severity, we conducted ROC curve analysis using the 'MedCalc' tool (30).
425 The sample sizes for each comparison were evaluated and are listed in
426 **Supplementary Table 4**. The AUROC of the QSOX1/IL1RAP ratio for distinguishing
427 NAFLD patients from healthy controls was 0.95, with a cutoff value of 1.12. The
428 sensitivity was determined to be 90%, and the specificity was 100%. Notably, the
429 efficacy of the QSOX1/IL1RAP ratio was superior to that of IL1RAP alone

430 (AUROC=0.92) or QSOX1 alone (not significant). Similarly, when assessing the
431 differentiation between cirrhosis patients and healthy controls, the AUROC of the
432 QSOX1/IL1RAP ratio was 0.96, with a cutoff value of 1.12. The sensitivity was 90%,
433 and the specificity was 100%.
434
435 These results indicate that the QSOX1/IL1RAP ratio holds promise as a highly
436 effective biomarker for diagnosing NAFLD severity, surpassing the individual
437 biomarkers alone, and maintaining better sensitivity and specificity in distinguishing
438 NAFLD patients and cirrhosis patients from healthy individuals.
439
440 **Validation of QSOX1 and IL1RAP as biomarkers for NAFLD in another patient cohort**
441 To further validate the utility of QSOX1 and IL1RAP as biomarkers for NAFLD, we
442 conducted a validation study in healthy controls and NAFLD patients recruited from
443 the Department of Liver Disease of Shenzhen Traditional Chinese Medicine Hospital.
444 Plasma samples were collected from 14 healthy subjects and 28 newly diagnosed
445 NAFLD patients. Clinical information for the healthy controls and NAFLD patients can
446 be found in **Supplementary Table 5**.
447
448 We measured plasma levels of QSOX1 and IL1RAP using an enzyme-linked
449 immunosorbent assay (ELISA). Consistent with our previous findings, plasma levels of
450 QSOX1 (Wilcoxon rank sum test, $p = 0.043$), IL1RAP ($p = 0.035$), and the
451 QSOX1/IL1RAP ratio ($p = 0.00061$) were significantly different between NAFLD
452 patients and controls (**Figure 4D, E, F**).
453
454 To assess the diagnostic value of QSOX1 and IL1RAP as non-invasive biomarkers for
455 NAFLD by ELISA, we calculated the AUROC of the QSOX1/IL1RAP ratio in the ELISA
456 test to distinguish NAFLD patients from healthy controls. The QSOX1/IL1RAP ratio
457 exhibited an AUROC of 0.82. Using a cutoff of 0.05, the sensitivity was 93% and the
458 specificity was 57%. In comparison, the AUROC of QSOX1/IL1RAP ratio quantified by
459 ELISA showed less efficacy in distinguishing NAFLD patients from healthy controls

460 (**Supplementary Table 4**), which may be attributed to the sensitivity of protein
461 quantification methods and small sample size.
462
463 To further validate the association between QSOX1 and IL1RAP protein levels and
464 NAFLD severity, we assessed their levels in liver biopsies from mild and severe NAFLD
465 patients using IHC. Our results consistently demonstrated a significant correlation
466 between QSOX1 and IL1RAP levels and NAFLD severity (**Figure 5A**). Quantification of
467 QSOX1 and IL1RAP levels based on IHC confirmed that the QSOX1/IL1RAP ratio ($p =$
468 0.027) could distinguish the severe NAFLD group (n=6; NAS 5-8, fibrosis score 3-4)
469 from the mild NAFLD group (n=6; NAS 0-4, Fibrosis score 0-2) (**Figure 5B, C**).
470 Collectively, these findings suggest that the QSOX1/IL1RAP ratio holds promise as an
471 effective biomarker for the early diagnosis and prediction of NAFLD severity.

472 ## Discussion

473 This study is the first to integrate publicly available RNA-seq datasets from over 600
474 NAFLD patients with varying stages of disease severity, combined with proteomics
475 data analysis of publicly available datasets. The key findings suggest that the
476 QSOX1/IL1RAP, and particularly the QSOX1/IL1RAP ratio hold promise as potential
477 biomarkers for NAFLD severity assessment. These results align with recent research
478 highlighting the importance of different transcriptional profiles specific to NAS and
479 fibrosis scores, offering valuable insights into the molecular mechanisms driving
480 disease progression from simple steatosis to inflammation and fibrosis (21, 24).
481

482 **The Advantages of Utilizing Integrated RNA-seq Data for Investigating NAFLD**
483 **Biomarkers**

484 Despite the growing availability of RNA-seq data in this field, many original studies
485 have been limited by small sample sizes and biased sample distribution, making it
486 challenging to accurately decipher transcriptional differences across various stages of
487 NAFLD(18, 20-22, 24). Several studies have attempted to identify diagnostic
488 biomarkers and potential drug targets. For instance, Brosch et al. conducted a
489 positional analysis of transcriptomes across three micro-dissected liver zones from 19
490 NAFLD patients (18). Suppli et al. demonstrated that immunohistochemical markers
491 offer greater objectivity in distinguishing hepatocyte injury between NASH and NAFL
492 (20). In the pursuit of diagnostic genes and novel drug targets, Hoang et al. studied 6
493 histologically normal and 72 NAFLD patients, while Pantano et al. studied 31
494 histologically normal and 112 NAFLD patients. These studies revealed that specific
495 cells proportion and candidate gene signatures can accurately predict fibrosis stage
496 and disease progression (21, 24). Likewise, Govaere et al. observed the correlation
497 between gene expression and histology in a cohort of 10 controls and 206 NAFLD
498 patients (22). In contrast to the studies above that identified sets of potential
499 biomarker genes, Kozumi et al. validated thrombospondin 2 (THBS2) as a noninvasive
500 biomarker for NAFLD. They confirmed its potential in identifying the disease stages

501 among 98 NAFLD patients, and the serum levels of its encoded protein TSP-2,
502 measured by ELISA, showed an AUROC of 0.78 in the diagnosing of NASH among 213
503 patients with biopsy-proven NAFLD (25). The major challenge of combining and
504 analyzing these diverse datasets lies in achieving homogeneous processing, which
505 requires substantial time and computational resources (48, 49). To generate more
506 robust and compelling results, we employed unbiased integration of comprehensive
507 NAFLD data to profile the liver transcriptome across a broad spectrum of NAFLD
508 severity in our study, incorporating all the aforementioned samples.
509
510 **The Superiority of QSOX1 and IL1RAP as Potential Biomarkers of NAFLD**
511 The high prevalence and associated risks of NAFLD have driven global efforts to
512 identify improved diagnostic biomarkers. However, most existing biomarkers are
513 primarily suited for evaluating fibrosis (3, 50-52). The Fibrosis-4 (FIB-4) test
514 commonly used in clinical practice, is sub-optimal for screening purposes, as it carries
515 the risks of both overdiagnosis and false negatives, particularly in patients at risk of
516 chronic liver disease (8). Although the patented ELF™ test was highly recommended
517 for ruling out advanced fibrosis, it comes with higher costs. Several steatosis scores,
518 such as the SteatoTest™ and the fatty liver index (FLI), have been proposed for
519 steatosis detection, but they do not provide substantial additional information
520 beyond routine clinical, laboratory, and imaging examinations conducted in
521 patients suspected of having NAFLD(8). Non-coding RNAs (ncRNAs), which exhibit
522 aberrant expression associated with NAFLD, have emerged as potential biomarkers
523 for NAFLD pathology, and circulating ncRNAs including miR-122 and lncRNAs are
524 proposed as potential biomarkers for NAFLD severity and progression (53-59).
525 Despite the development of new biomarkers, there is still uncertainty surrounding
526 their predictive value, underscoring the urgent need to develop novel, cost-effective,
527 and efficient biomarkers with high sensitivity and specificity for NAFLD prediction
528 and monitoring (4, 60).
529
530 The approach by Hoang et al. (21) centered on identifying genes with diverse

531 expressions associated with NAFLD severity, inspired us to develop our gene
532 clustering method. Our approach surpasses the constraints of conventional RNA-seq
533 data analysis, which predominantly relies on pairwise comparisons. Instead, it
534 classifies genes according to their dynamic expression patterns, enabling a more
535 comprehensive and dynamic perspective of molecular alterations as NAFLD
536 progresses. This method has the potential to map NAFLD severity and progression
537 solely through gene expressions, thus avoiding invasive procedures like liver biopsies.
538 Moreover, the gene-based scoring system can forecast NAFLD progression,
539 facilitating early interventions for patients at risk of advancing to severe disease
540 stages.
541
542 Previous studies have explored the relationship between QSOX1, IL1RAP, and NAFLD
543 or steatosis(16, 61). QSOX1 has been suggested as a potential diagnostic biomarker
544 for NAFLD, playing a significant role in lipid metabolism as an enzyme expressed in
545 various tissues, particularly in quiescent fibroblasts (18, 62, 63). IL1RAP is localized in
546 vesicles and cytosol, and it is secreted into the bloodstream. Notably, IL1RAP
547 expression at the RNA level was specifically detected in the liver and hepatocytes
548 (44). Hence, the combination of QSOX1 and IL1RAP as secretome genes and proteins
549 was selected as a potential biomarker combination.
550
551 The potential of QSOX1, IL1RAP, and their ratio as biomarkers for NAFLD was
552 demonstrated through the analysis of public RNA-seq and proteomics data, ELISA
553 tests conducted on patients' plasma, and IHC performed on fixed liver slides. These
554 findings suggest that QSOX1, IL1RAP, and their ratio hold promise as effective
555 biomarkers for NAFLD. Notably, the higher AUROC values for NAFLD diagnosis
556 achieved by QSOX1, IL1RAP, and their ratio highlight their efficacy as NAFLD
557 biomarkers.
558
559 **Limitation and Future Prospects.**
560 The current study possesses several strengths, including the integration and

561 processing of RNA-seq data from over 600 NAFLD patients with varying degrees of
562 NAFLD severity, as well as validation using proteomics data and samples from NAFLD
563 patients and controls. Furthermore, the well-established database with a
564 user-friendly interface could benefit the research community in exploring
565 differentially expressed genes in NAFLD at various stages. However, there are also
566 limitations to consider. For instance, some samples in the GSE126848 and GSE167523
567 datasets lacked individual NAS and fibrosis scores. To address this issue,
568 we standardized scores based on their categories in the original articles, and the
569 impact on the results was deemed negligible due to the provision of general stages
570 and unsupervised gene clustering. Machine learning, an essential tool for biomarker
571 validation and sample classification validation, should be employed to train large
572 cohorts of biopsy-proven NAFLD patients and healthy controls. However, this would
573 require an extended recruiting period(64) to determine the sensitivity and specificity
574 of the QSOX1/IL1RAP ratio for NAFLD diagnosis and staging.
575
576 Although newer technologies such as single-cell RNA sequencing (scRNA-seq) and
577 spatial sequencing have gained popularity, RNA-seq still serves as a valuable tool in
578 uncovering the pathogenesis of NAFLD (17). Computational analysis limitations make
579 it impractical for large cohort research, and single-cell suspension processing may
580 affect cell abundance and cell type representation, particularly in hepatic ballooning
581 cells in NAFLD (65). Single-nuclei RNA sequencing (snRNA-seq) captures cell
582 frequency more accurately than scRNA-seq but captures lower gene expression.
583 Spatial transcriptomics and proteomics have limitations for discovering invasive
584 biomarkers of NAFLD as they focus on small sampling areas (15). The combination of
585 all these biological tools holds potential for future research.
586
587 In conclusion, through a novel approach of unsupervised gene clustering performed
588 on integrated RNA-seq data, we have discovered a significant association between
589 QSOX1 and IL1RAP levels and NAFLD severity, with their ratio showing potential as a
590 non-invasive biomarker for diagnosing and assessing the severity of NAFLD.

591 Validation of our plasma-level findings in larger cohorts of liver biopsies is required,
592 but it holds promise as a new tool to diagnose NAFLD severity and reduce the need
593 for liver biopsies. Our approach may lead to the discovery of more NAFLD biomarkers,
594 and the ratios of other up-regulated and down-regulated genes associated with
595 increasing NAFLD severity also have the potential to be verified as potential
596 biomarkers.
597

598 **Abbreviations**

599 AUROC, area under the receiver operating characteristic; avg_log2FC, log fold-change
600 of the average expression between the two groups; BMI, body mass index; CK18,
601 circulating keratin 18 fragments; ECM, extracellular matrix; ELISA, enzyme-linked
602 immunosorbent assay; F, Fibrosis score; FIB-4, Fibrosis-4; GEO, Gene Expression
603 Omnibus; GO, Gene Ontology; GRCh37, Genome Reference Consortium Human Build
604 37; HCC, hepatocellular carcinoma; IHC, immunohistochemistry staining; IL1RAP,
605 Interleukin-1 receptor accessory protein; lg, log10; PCA, Principal components
606 analysis; QSOX1, Quiescin sulfhydryl oxidase 1; RNA-seq, RNA sequencing; N, NAS
607 score; NAFL, Non-alcoholic Fatty Liver; NAFLD, Non-alcoholic fatty liver disease;
608 NAFLD-DB, NAFLD gene expression database; NAFLD_ngt, NAFLD with normal
609 glucose tolerance; NAFLD_T2D, NAFLD with type 2 diabetes; NAS, NAFLD activity
610 scores; NASH, Non-alcoholic Steatohepatitis; ncRNAs, non-coding RNAs; scRNA-seq,
611 single-cell RNA sequencing; snRNA-seq, Single-nuclei RNA sequencing; SZTCMH,
612 Shenzhen Traditional Chinese Medicine Hospital, China; THBS2, thrombospondin 2;
613 TMM, Trimmed Mean of M-values; TPM, Transcript Per Million.
614

615 **Financial support**

616 This research was funded by the Shenzhen Science and Technology Project and
617 Sanming Project of Medicine in Shenzhen, China (grant nos. SZSM201612074,
618 JCYJ20210324120405015).
619

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.

653 7. Piazzolla VA, Mangia A. Noninvasive Diagnosis of NAFLD and NASH. Cells 2020;9.
654 8. European Association for the Study of the Liver. Electronic address eee, Clinical Practice
655 Guideline P, Chair, representative EGB, Panel m. EASL Clinical Practice Guidelines on non-invasive tests
656 for evaluation of liver disease severity and prognosis - 2021 update. J Hepatol 2021;75:659-689.
657 9. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression
658 analysis of digital gene expression data. Bioinformatics 2010;26:139-140.
659 10. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq
660 data with DESeq2. Genome Biol 2014;15:550.
661 11. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, et al. Differential gene and
662 transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc
663 2012;7:562-578.
664 12. Costa-Silva J, Domingues DS, Menotti D, Hungria M, Lopes FM. Temporal progress of gene
665 expression analysis with RNA-Seq data: A review on the relationship between computational methods.
666 Comput Struct Biotechnol J 2023;21:86-98.
667 13. Barreby E, Chen P, Aouadi M. Macrophage functional diversity in NAFLD - more than
668 inflammation. Nat Rev Endocrinol 2022;18:461-472.
669 14. Wang ZY, Keogh A, Waldt A, Cuttat R, Neri M, Zhu S, Schuierer S, et al. Single-cell and bulk
670 transcriptomics of the liver reveals potential targets of NASH with fibrosis. Sci Rep 2021;11:19396.
671 15. Guilliams M, Bonnardel J, Haest B, Vanderborght B, Wagner C, Remmerie A, Bujko A, et al. Spatial
672 proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. Cell
673 2022;185:379-396 e338.
674 16. Green CD, Dozmorov MG, Spiegel S. Analysis of Liver Responses to Non-alcoholic Steatohepatitis
675 by mRNA-Sequencing. Methods Mol Biol 2022;2455:163-179.
676 17. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet
677 2019;20:631-656.
678 18. Brosch M, Kattler K, Herrmann A, von Schonfels W, Nordstrom K, Seehofer D, Damm G, et al.
679 Epigenomic map of human liver reveals principles of zonated morphogenic and metabolic control. Nat
680 Commun 2018;9:4150.
681 19. Mardinoglu A, Wu H, Bjornson E, Zhang C, Hakkarainen A, Rasanen SM, Lee S, et al. An Integrated
682 Understanding of the Rapid Metabolic Benefits of a Carbohydrate-Restricted Diet on Hepatic Steatosis
683 in Humans. Cell Metab 2018;27:559-571 e555.
684 20. Suppli MP, Rigbolt KTG, Veidal SS, Heeboll S, Eriksen PL, Demant M, Bagger JI, et al. Hepatic
685 transcriptome signatures in patients with varying degrees of nonalcoholic fatty liver disease compared
686 with healthy normal-weight individuals. Am J Physiol Gastrointest Liver Physiol 2019;316:G462-G472.
687 21. Hoang SA, Oseini A, Feaver RE, Cole BK, Asgharpour A, Vincent R, Siddiqui M, et al. Gene
688 Expression Predicts Histological Severity and Reveals Distinct Molecular Profiles of Nonalcoholic Fatty
689 Liver Disease. Sci Rep 2019;9:12541.
690 22. Govaere O, Cockell S, Tiniakos D, Queen R, Younes R, Vacca M, Alexander L, et al. Transcriptomic
691 profiling across the nonalcoholic fatty liver disease spectrum reveals gene signatures for
692 steatohepatitis and fibrosis. Sci Transl Med 2020;12.
693 23. Pfister D, Nunez NG, Pinyol R, Govaere O, Pinter M, Szydlowska M, Gupta R, et al. NASH limits
694 anti-tumour surveillance in immunotherapy-treated HCC. Nature 2021;592:450-456.
695 24. Pantano L, Agyapong G, Shen Y, Zhuo Z, Fernandez-Albert F, Rust W, Knebel D, et al. Molecular
696 characterization and cell type composition deconvolution of fibrosis in NAFLD. Sci Rep 2021;11:18045.

697 25. Kozumi K, Kodama T, Murai H, Sakane S, Govaere O, Cockell S, Motooka D, et al. Transcriptomics
698 Identify Thrombospondin-2 as a Biomarker for NASH and Advanced Liver Fibrosis. Hepatology
699 2021;74:2452-2466.
700 26. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling,
701 and transformations: improving the biological information content of metabolomics data. BMC
702 Genomics 2006;7:142.
703 27. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential
704 expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47.
705 28. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A, Lee MJ, et al. Integrated
706 analysis of multimodal single-cell data. Cell 2021;184:3573-3587 e3529.
707 29. Niu L, Geyer PE, Wewer Albrechtsen NJ, Gluud LL, Santos A, Doll S, Treit PV, et al. Plasma
708 proteome profiling discovers novel proteins associated with non-alcoholic fatty liver disease. Mol Syst
709 Biol 2019;15:e8793.
710 30. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat
711 Methods 2012;9:671-675.
712 31. Shu J, Dolman GE, Duan J, Qiu G, Ilyas M. Statistical colour models: an automated digital image
713 analysis method for quantification of histological biomarkers. Biomed Eng Online 2016;15:46.
714 32. Febbraio MA, Reibe S, Shalapour S, Ooi GJ, Watt MJ, Karin M. Preclinical Models for Studying
715 NASH-Driven HCC: How Useful Are They? Cell Metab 2019;29:18-26.
716 33. Florentino RM, Fraunhoffer NA, Morita K, Takeishi K, Ostrowska A, Achreja A, Animasahun O, et
717 al. Cellular Location of HNF4alpha is Linked With Terminal Liver Failure in Humans. Hepatol Commun
718 2020;4:859-875.
719 34. Xu C, Markova M, Seebeck N, Loft A, Hornemann S, Gantert T, Kabisch S, et al. High-protein diet
720 more effectively reduces hepatic fat than low-protein diet despite lower autophagy and FGF21 levels.
721 Liver Int 2020;40:2982-2997.
722 35. Hou J, Zhang J, Cui P, Zhou Y, Liu C, Wu X, Ji Y, et al. TREM2 sustains macrophage-hepatocyte
723 metabolic coordination in nonalcoholic fatty liver disease and sepsis. J Clin Invest 2021;131.
724 36. Yang W, Feng Y, Zhou J, Cheung OK, Cao J, Wang J, Tang W, et al. A selective HDAC8 inhibitor
725 potentiates antitumor immunity and efficacy of immune checkpoint blockade in hepatocellular
726 carcinoma. Sci Transl Med 2021;13.
727 37. Brunt EM, Janney CG, Di Bisceglie AM, Neuschwander-Tetri BA, Bacon BR. Nonalcoholic
728 steatohepatitis: a proposal for grading and staging the histological lesions. Am J Gastroenterol
729 1999;94:2467-2474.
730 38. Kleiner DE, Brunt EM, Van Natta M, Behling C, Contos MJ, Cummings OW, Ferrell LD, et al. Design
731 and validation of a histological scoring system for nonalcoholic fatty liver disease. Hepatology
732 2005;41:1313-1321.
733 39. Karlsson M, Sjostedt E, Oksvold P, Sivertsson A, Huang J, Alvez MB, Arif M, et al. Genome-wide
734 annotation of protein-coding genes in pig. BMC Biol 2022;20:25.
735 40. Ouyang JF, Kamaraj US, Cao EY, Rackham OJL. ShinyCell: Simple and sharable visualisation of
736 single-cell gene expression data. Bioinformatics 2021.
737 41. Liu Y, Ciotti GE, Eisinger-Mathason TSK. Hypoxia and the Tumor Secretome. Adv Exp Med Biol
738 2019;1136:57-69.
739 42. Crescenzi E, Leonardi A, Pacifico F. NGAL as a Potential Target in Tumor Microenvironment. Int J
740 Mol Sci 2021;22.

741 43. Zhou X, Zhang J, Lv W, Zhao C, Xia Y, Wu Y, Zhang Q. The pleiotropic roles of adipocyte secretome
742 in remodeling breast cancer. J Exp Clin Cancer Res 2022;41:203.
743 44. Uhlen M, Karlsson MJ, Hober A, Svensson AS, Scheffel J, Kotol D, Zhong W, et al. The human
744 secretome. Sci Signal 2019;12.
745 45. Corey KE, Pitts R, Lai M, Loureiro J, Masia R, Osganian SA, Gustafson JL, et al. ADAMTSL2 protein
746 and a soluble biomarker signature identify at-risk non-alcoholic steatohepatitis and fibrosis in adults
747 with NAFLD. J Hepatol 2022;76:25-33.
748 46. Gerhard GS, Hanson A, Wilhelmsen D, Piras IS, Still CD, Chu X, Petrick AT, et al. AEBP1 expression
749 increases with severity of fibrosis in NASH and is regulated by glucose, palmitate, and miR-372-3p.
750 PLoS One 2019;14:e0219764.
751 47. Cengiz M, Yilmaz G, Ozenirler S. Serum Biglycan as a Diagnostic Marker for Non-Alcoholic
752 Steatohepatitis and Liver Fibrosis. Clin Lab 2021;67.
753 48. Stupnikov A, McInerney CE, Savage KI, McIntosh SA, Emmert-Streib F, Kennedy R, Salto-Tellez M,
754 et al. Robustness of differential gene expression analysis of RNA-seq. Comput Struct Biotechnol J
755 2021;19:3470-3481.
756 49. Shakola F, Palejev D, Ivanov I. A Framework for Comparison and Assessment of Synthetic
757 RNA-Seq Data. Genes (Basel) 2022;13.
758 50. Tapper EB, Loomba R. Noninvasive imaging biomarker assessment of liver fibrosis by
759 elastography in NAFLD. Nat Rev Gastroenterol Hepatol 2018;15:274-282.
760 51. Masoodi M, Gastaldelli A, Hyotylainen T, Arretxe E, Alonso C, Gaggini M, Brosnan J, et al.
761 Metabolomics and lipidomics in NAFLD: biomarkers and non-invasive diagnostic tests. Nat Rev
762 Gastroenterol Hepatol 2021;18:835-856.
763 52. Wong VW, Adams LA, de Ledinghen V, Wong GL, Sookoian S. Noninvasive biomarkers in NAFLD
764 and NASH - current progress and future promise. Nat Rev Gastroenterol Hepatol 2018;15:461-478.
765 53. Panera N, Gnani D, Crudele A, Ceccarelli S, Nobili V, Alisi A. MicroRNAs as controlled systems and
766 controllers in non-alcoholic fatty liver disease. World J Gastroenterol 2014;20:15079-15086.
767 54. Fang Z, Dou G, Wang L. MicroRNAs in the Pathogenesis of Nonalcoholic Fatty Liver Disease. Int J
768 Biol Sci 2021;17:1851-1863.
769 55. Qian X, Zong W, Ma L, Yang Z, Chen W, Yan J, Xu J. MM-associated circular RNA downregulates
770 microRNA-19a through methylation to suppress proliferation of pancreatic adenocarcinoma cells.
771 Bioengineered 2022;13:9294-9300.
772 56. Jampoka K, Muangpaisarn P, Khongnomnan K, Treeprasertsuk S, Tangkijvanich P, Payungporn S.
773 Serum miR-29a and miR-122 as Potential Biomarkers for Non-Alcoholic Fatty Liver Disease (NAFLD).
774 Microrna 2018;7:215-222.
775 57. Akuta N, Kawamura Y, Suzuki F, Saitoh S, Arase Y, Fujiyama S, Sezaki H, et al. Analysis of
776 association between circulating miR-122 and histopathological features of nonalcoholic fatty liver
777 disease in patients free of hepatocellular carcinoma. BMC Gastroenterol 2016;16:141.
778 58. Wang W, Min L, Qiu X, Wu X, Liu C, Ma J, Zhang D, et al. Biological Function of Long Non-coding
779 RNA (LncRNA) Xist. Front Cell Dev Biol 2021;9:645647.
780 59. Atanasovska B, Rensen SS, Marsman G, Shiri-Sverdlov R, Withoff S, Kuipers F, Wijmenga C, et al.
781 Long Non-Coding RNAs Involved in Progression of Non-Alcoholic Fatty Liver Disease to Steatohepatitis.
782 Cells 2021;10.
783 60. Chee D, Ng CH, Chan KE, Huang DQ, Teng M, Muthiah M. The Past, Present, and Future of
784 Noninvasive Test in Chronic Liver Diseases. Med Clin North Am 2023;107:397-421.

785 61. Bozaoglu K, Attard C, Kulkarni H, Cummings N, Diego VP, Carless MA, Shields KA, et al. Plasma
786 levels of soluble interleukin 1 receptor accessory protein are reduced in obesity. J Clin Endocrinol
787 Metab 2014;99:3435-3443.
788 62. Gao R, Wang J, He X, Wang T, Zhou L, Ren Z, Yang J, et al. Comprehensive analysis of endoplasmic
789 reticulum-related and secretome gene expression profiles in the progression of non-alcoholic fatty
790 liver disease. Front Endocrinol (Lausanne) 2022;13:967016.
791 63. Reznik N, Fass D. Disulfide bond formation and redox regulation in the Golgi apparatus. FEBS Lett
792 2022;596:2859-2872.
793 64. Ledesma D, Symes S, Richards S. Advancements within Modern Machine Learning Methodology:
794 Impacts and Prospects in Biomarker Discovery. Curr Med Chem 2021;28:6512-6531.
795 65. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, Khodadoust MS, et al.
796 Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat
797 Biotechnol 2019;37:773-782.
798

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.

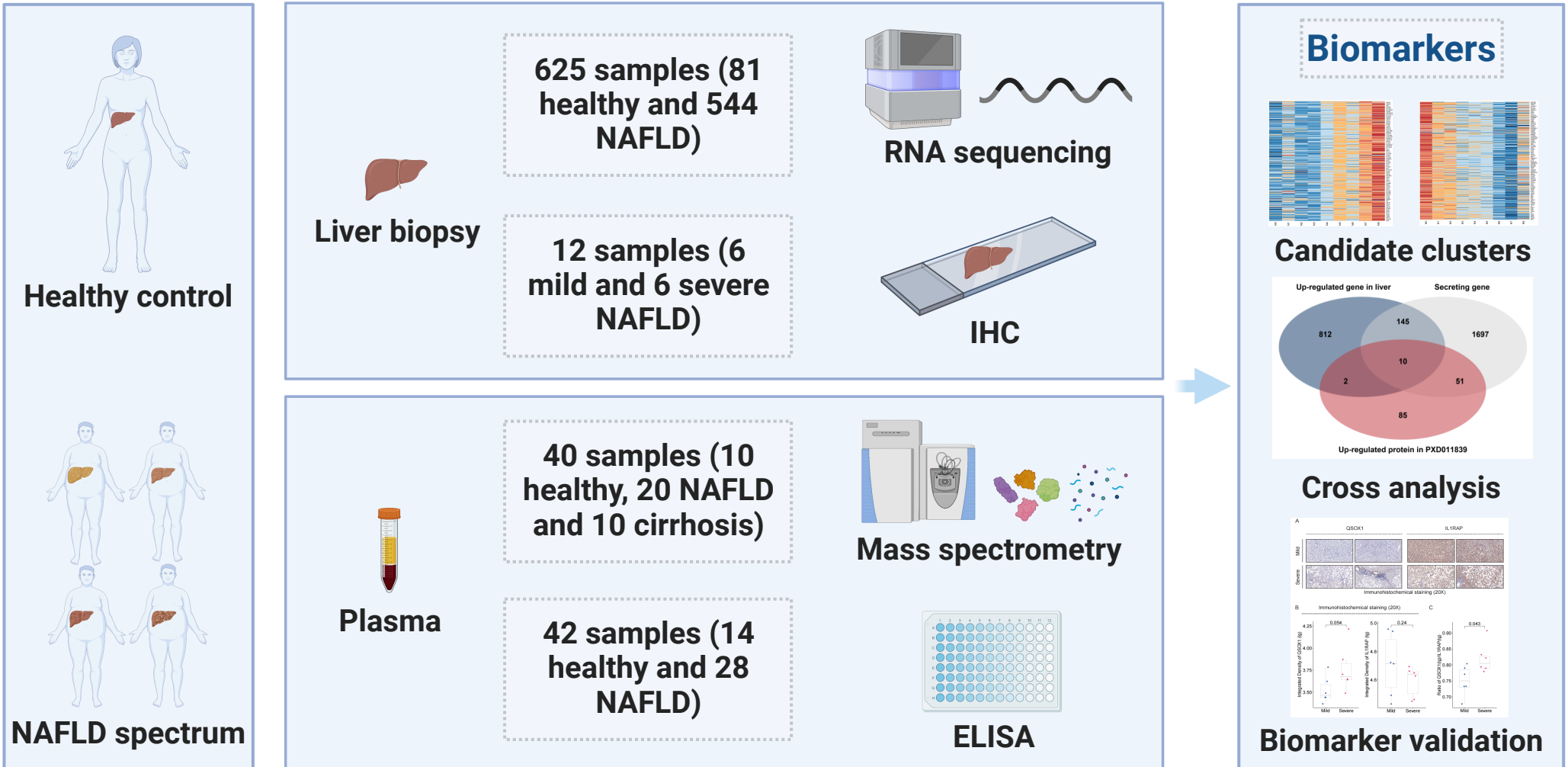

Healthy control
NAFLD spectrum
Liver biopsy
625 samples (81 healthy and 544 NAFLD)
12 samples (6 mild and 6 severe NAFLD)
RNA sequencing
IHC
Plasma
40 samples (10 healthy, 20 NAFLD and 10 cirrhosis)
42 samples (14 healthy and 28 NAFLD)
Mass spectrometry
ELISA
Biomarkers
Candidate clusters
Up-regulated 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

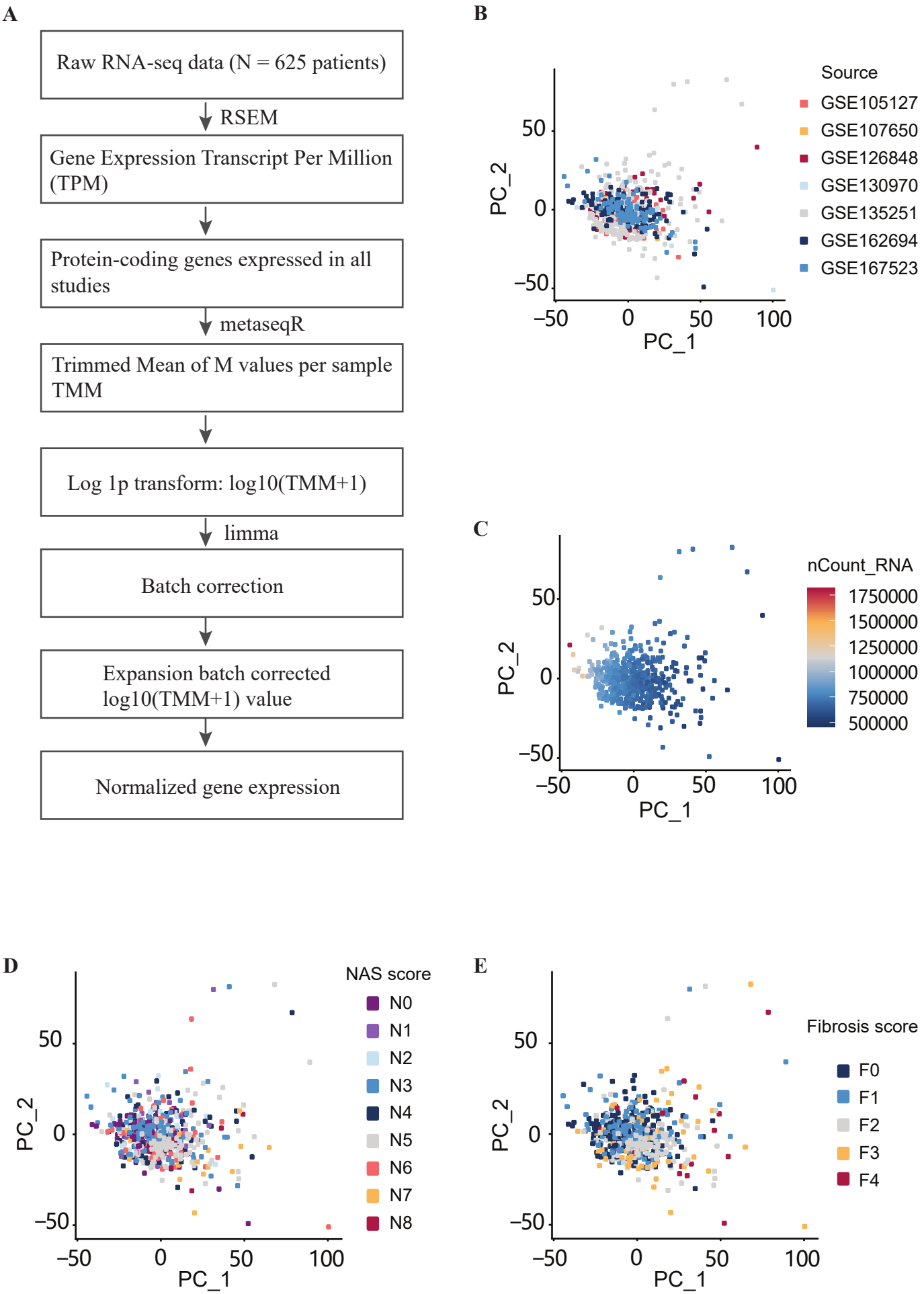

A
Raw RNA-seq data (N = 625 patients)
RSEM
Gene Expression Transcript Per Million (TPM)
Protein-coding genes expressed in all studies
metaseqR
Trimmed Mean of M values per sample TMM
Log 1p transform: log10(TMM+1)
limma
Batch correction
Expansion batch corrected log10(TMM+1) value
Normalized gene expression
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

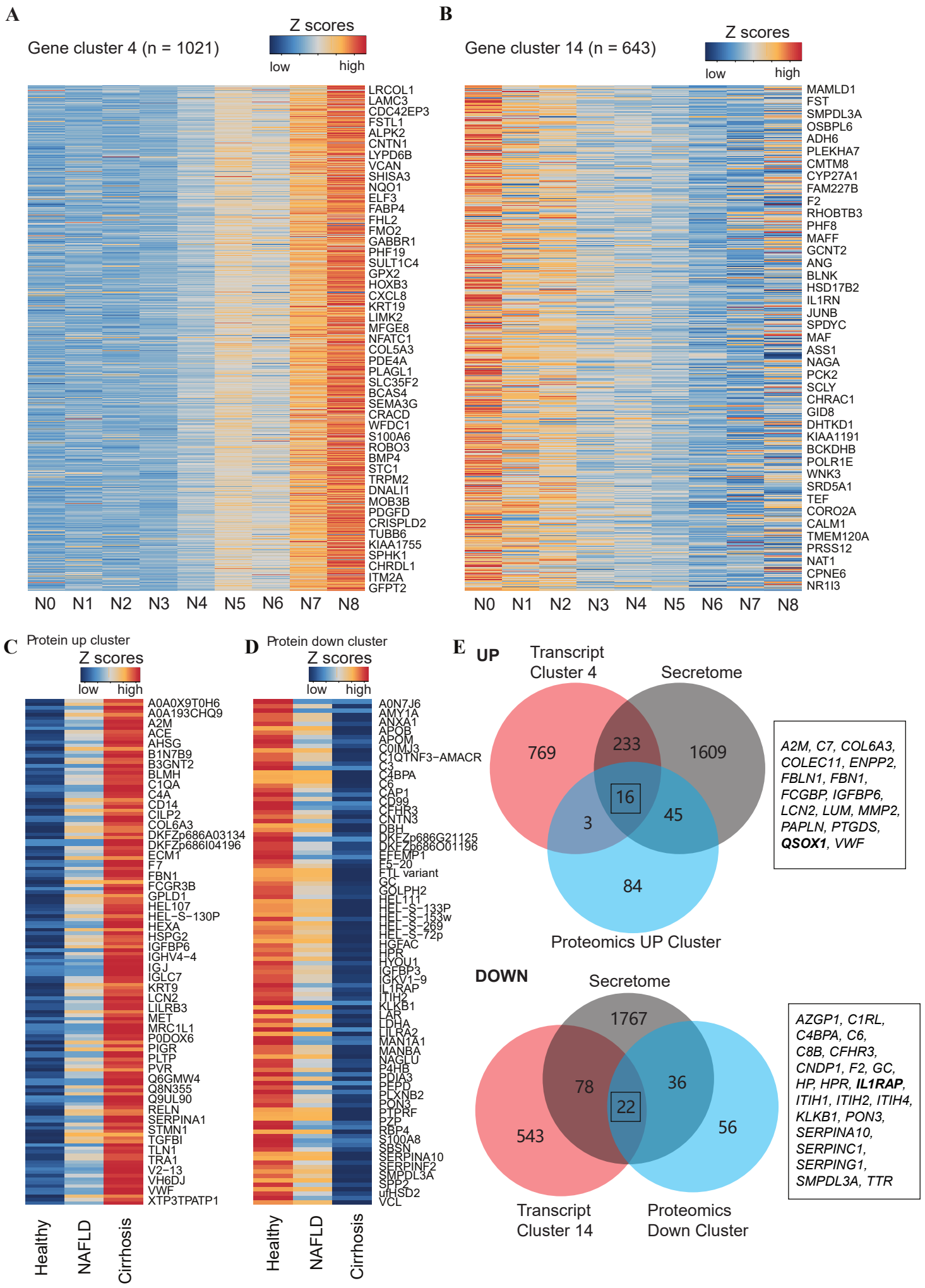

A
Gene cluster 4 (n = 1021)
Z scores
low
high
LRCOL1
LAMC3
CDC42EP3
FSTL1
ALPK2
CNTN1
LYPD6B
VCAN
SHISA3
NQO1
ELF3
FABP4
FHL2
FMO2
GABBR1
PHF19
SULT1C4
GPX2
HOXB3
CXCL8
KRT19
LIMK2
MFGE8
NFATC1
COL5A3
PDE4A
PLAGL1
SLC35F2
BCAS4
SEMA3G
CRACD
WFDC1
S100A6
ROBO3
BMP4
STC1
TRPM2
DNALI1
MOB3B
PDGFD
CRISPLD2
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

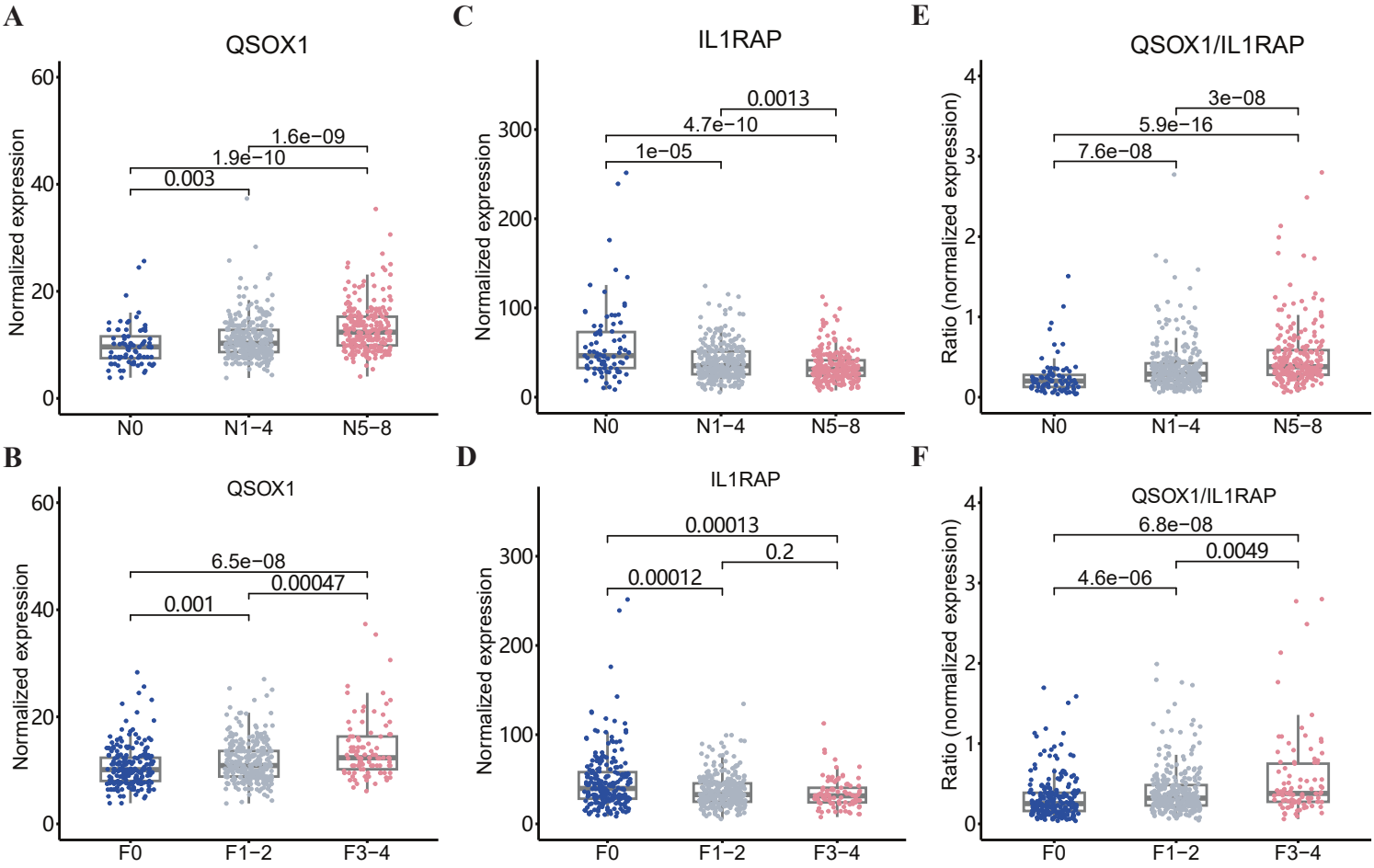

A
QSOX1
Normalized expression
0.003
1.9e−10
1.6e−09
N0
N1−4
N5−8
B
QSOX1
Normalized expression
0.001
6.5e−08
0.00047
F0
F1−2
F3−4
C
IL1RAP
Normalized expression
1e−05
4.7e−10
0.0013
N0
N1−4
N5−8
D
IL1RAP
Normalized expression
0.00012
0.00013
0.2
F0
F1−2
F3−4
E
QSOX1/IL1RAP
Ratio (normalized expression)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

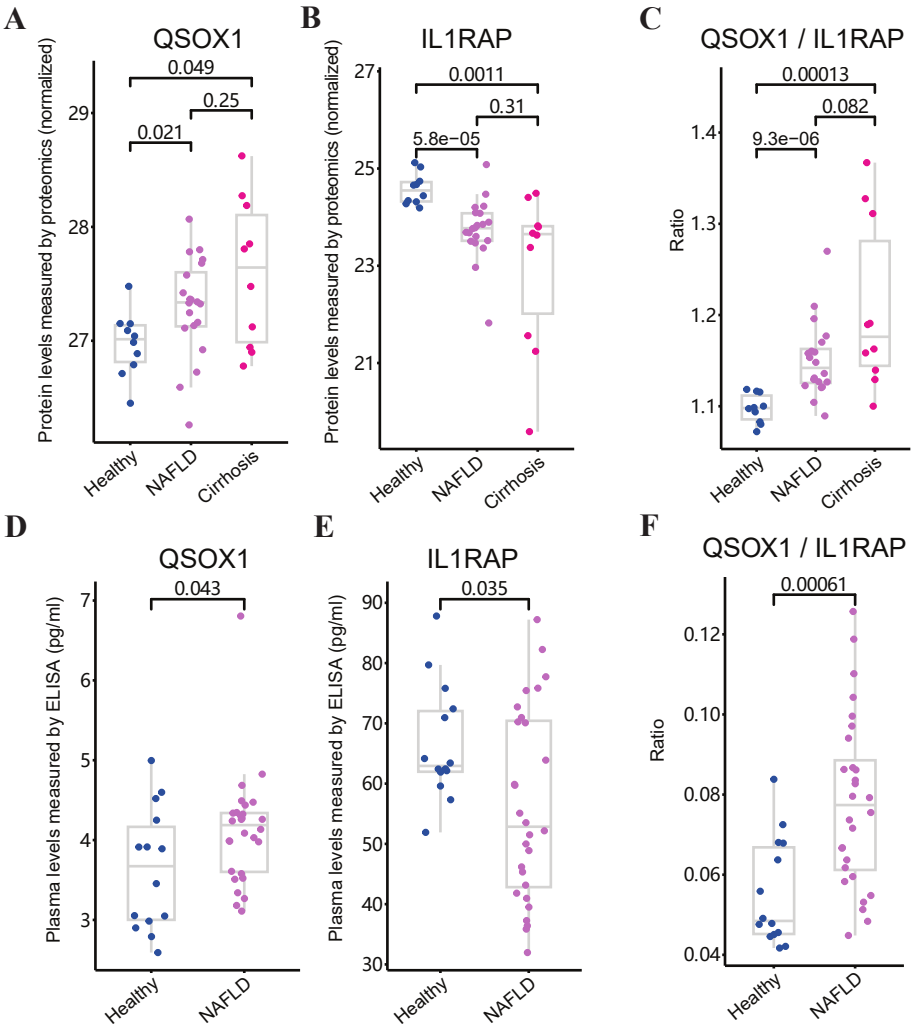
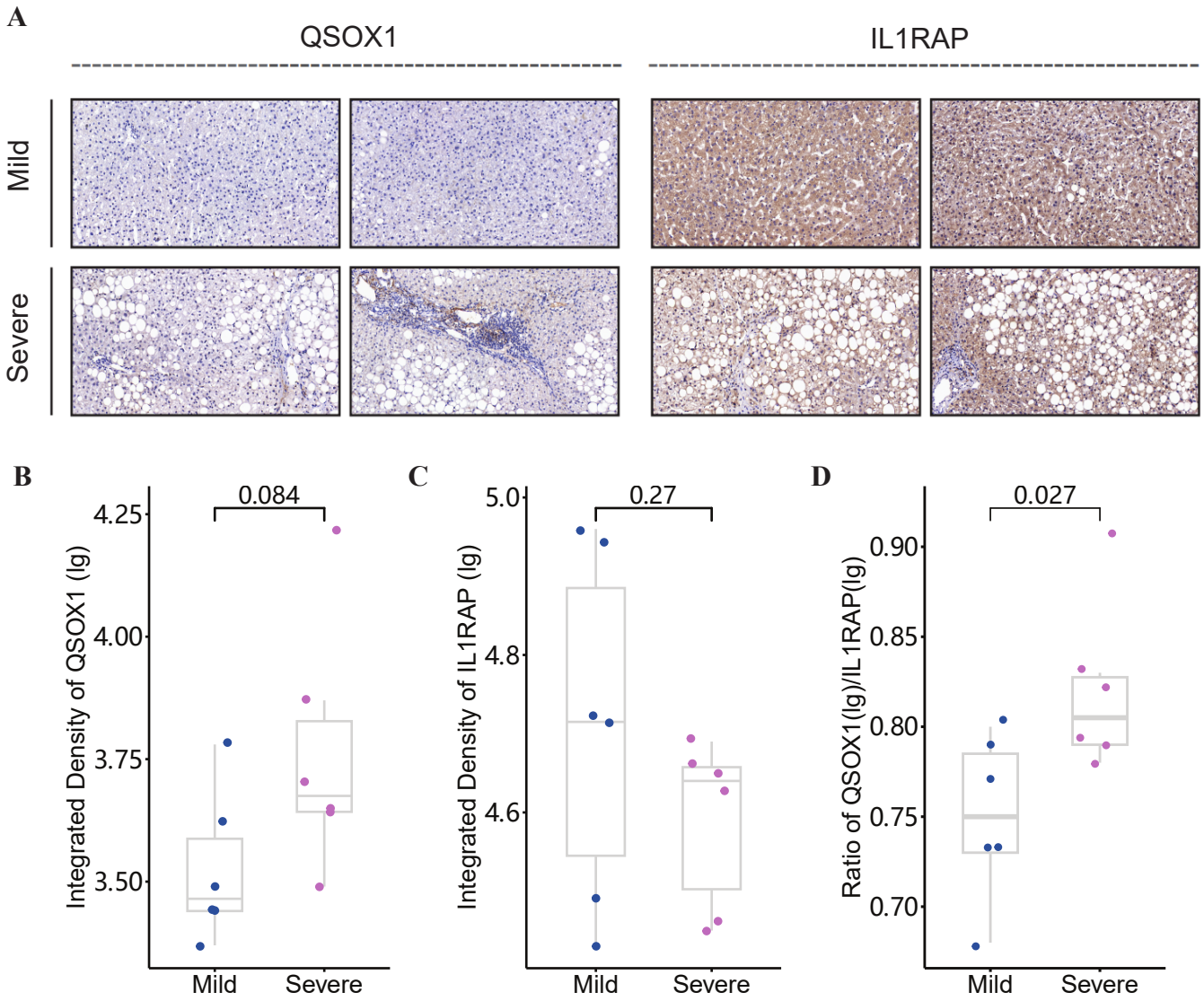

A
QSOX1
Protein levels measured by proteomics (normalized)
0.049
0.25
0.021
29
28
27
Healthy
NAFLD
Cirrhosis
B
IL1RAP
Protein levels measured by proteomics (normalized)
0.0011
0.31
5.8e−05
27
25
23
21
Healthy
NAFLD
Cirrhosis
C
QSOX1 / IL1RAP
Ratio
0.00013
0.082
9.3e−06
1.4
1.3
1.2
1.1
Healthy
NAFLD
Cirrhosis
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

Supplementary File

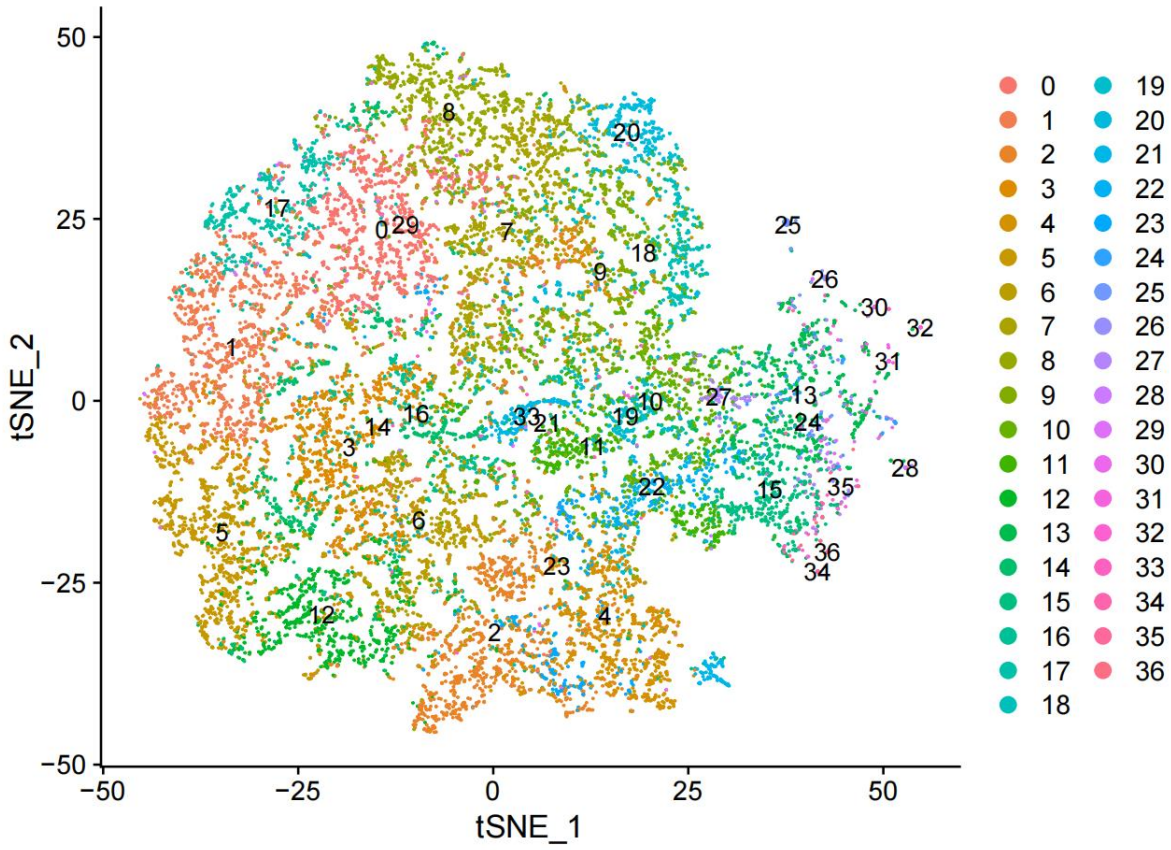
A gene-based clustering approach reveals QSOX1/IL1RAP as promising biomarkers for the severity of non-alcoholic fatty liver disease

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

This supplementary file contains: Supplementary Figure S1-S5

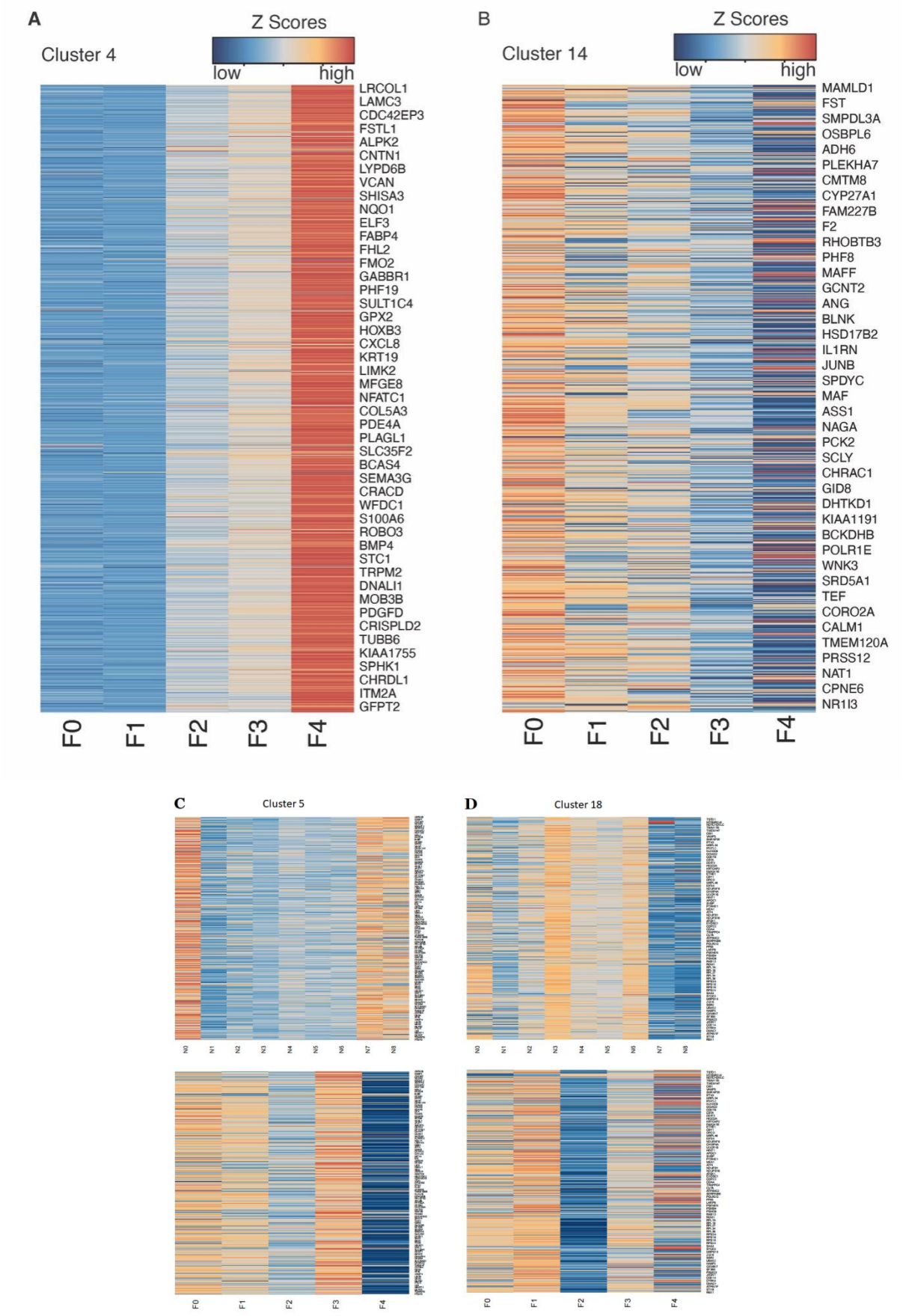
Supplementary Figure S1



Supplementary Figure S1. t-SNE visualization of gene clusters

Visualization of gene expression profiling clusters across NAFLD progression with the t-distributed stochastic neighbor embedding (t-SNE) statistical method. Each dot represents one protein coding gene (n = 17,946). A graph-based clustering approach was used. The dimensions of reduction were set to 1:20 and visualized with a resolution of 2.3.

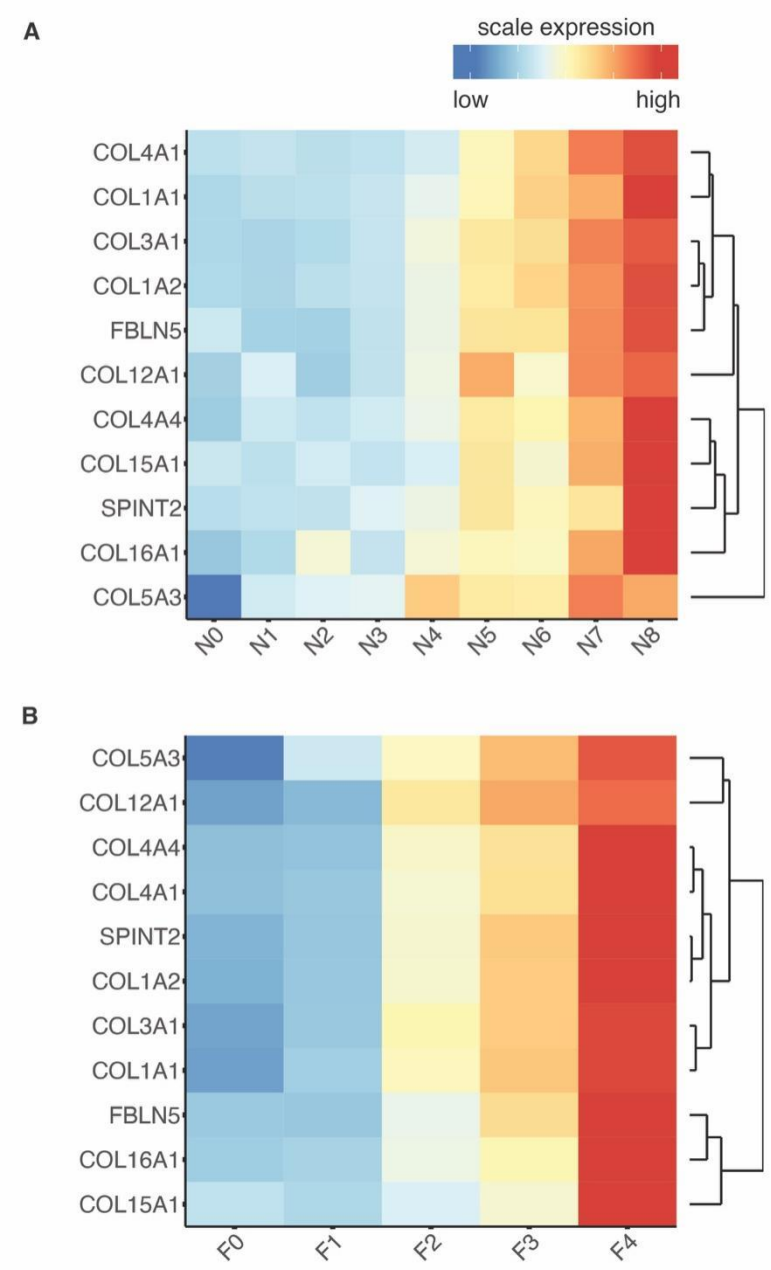
Supplementary Figure S2



Supplementary Figure S2 Heatmap presentation of gene expression profile along NAFLD progression.

A. Heatmap presentation of 1021 up-regulated genes in cluster 4 associated with increasing fibrosis scores.
B. Heatmap presentation of 643 down-regulated genes in cluster 14 associated with increasing fibrosis scores.
C and D. Contrary to genes in cluster 4 and 14, C and D displayed gene clusters with chaotic gene expression patterns associated with both NAS scores and fibrosis scores.

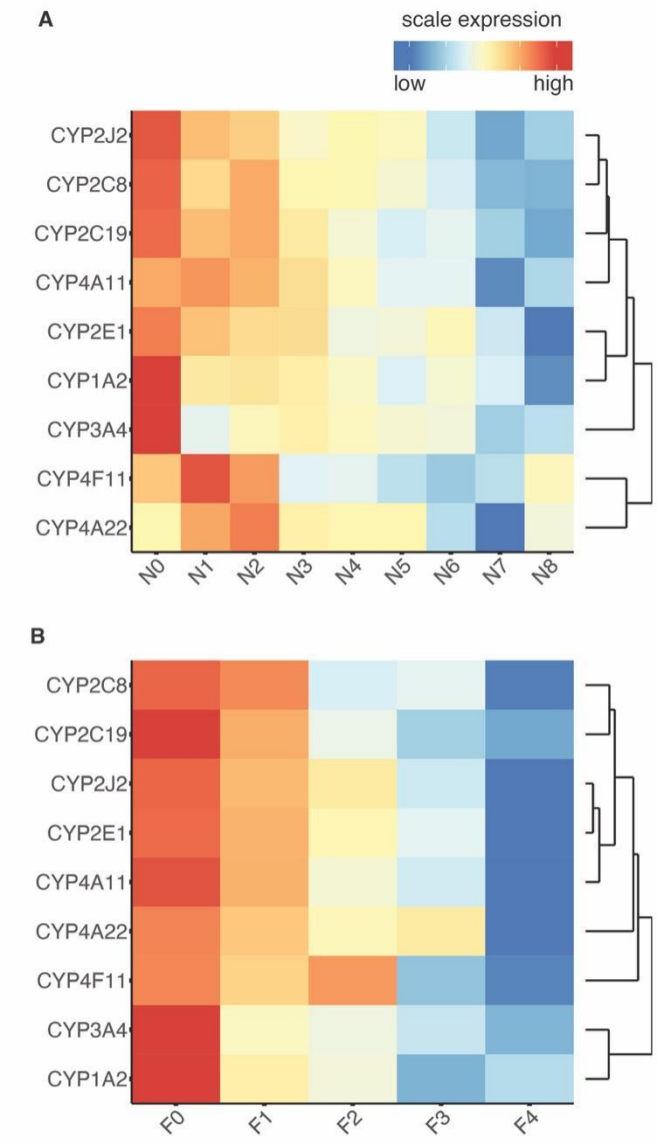
Supplementary Figure S3



Supplementary Figure S3 Heatmap presentation of ECM gene expression profile

A. Scale gene expression profile of multiple genes involved in the ECM process according to NAS scores. B. Scale gene expression profile of multiple genes involved in the ECM process according to fibrosis scores. Genes were clustered based on profile similarity. Genes expression level was scaled for heatmap presentation (also see the NAFLD-DB).

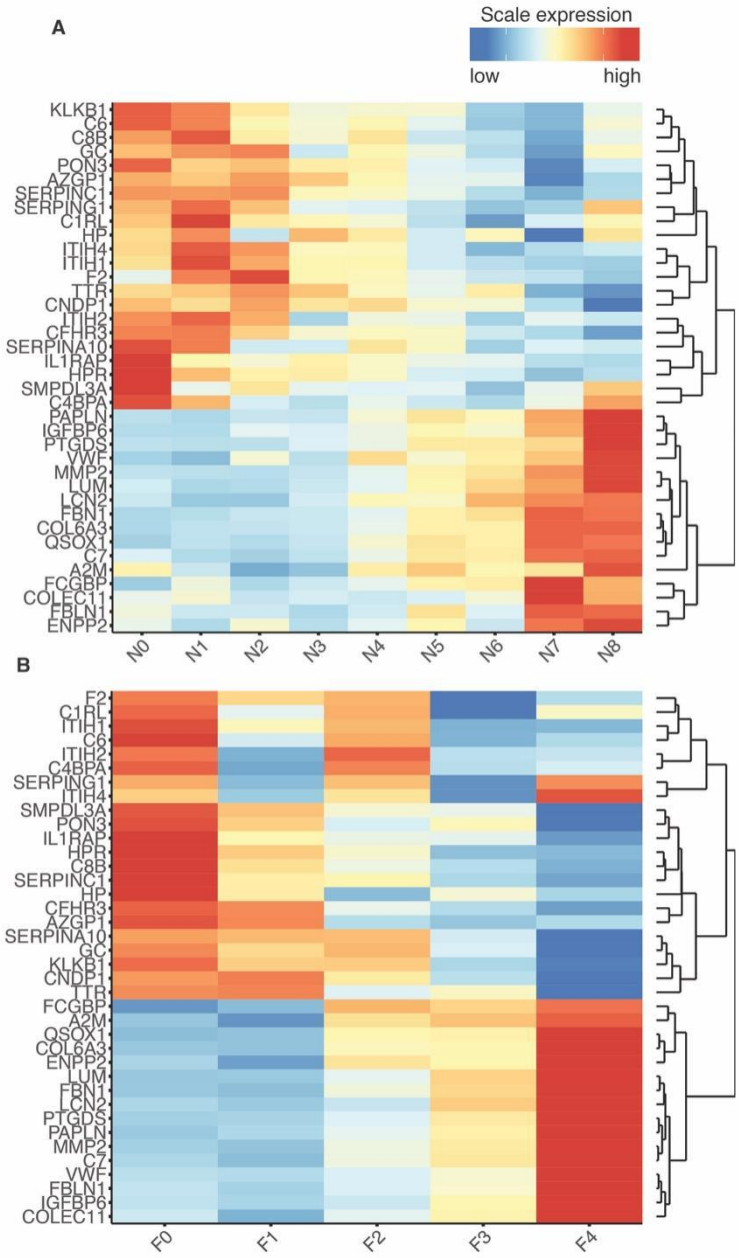
Supplementary Figure S4



Supplementary Figure S4 Heatmap presentation of Cytochrome P450 superfamily gene expression profile

A. Scale gene expression profile of multiple genes involved in the Cytochrome P450 superfamily according to NAS scores. B. Scale gene expression profile of multiple genes involved in the Cytochrome P450 superfamily according to fibrosis scores. Genes were clustered based on profile similarity. Genes expression level was scaled for heatmap presentation (also see the NAFLD-DB).

Supplementary Figure S5



Supplementary Figure S5 Heatmap presentation of expression profile for 32 biomarker genes

A. Scale gene expression profile of 38 biomarker genes (16 up-regulation, 22 down-regulation) according to NAS scores. B. Scale gene expression profile of 38 biomarker genes according to fibrosis scores. Genes were clustered based on profile similarity. Genes expression level was scaled for heatmap presentation (also see the NAFLD-DB). This figure is related to Figure 2E.

List of Supplementary Tables

Supplementary Table 1

The RNA-seq data of human liver samples.

Supplementary Table 2

GO enrichment results for genes in cluster 4 (up-regulation) and cluster 14 (down regulation).

Supplementary Table 3

List of secreting protein-encoding genes in cluster 4, and 14 of RNA-seq analysis. Representing PMID supporting that the candidate gene as potential biomarker for NAFLD was listed. Note: this is a noncomprehensive list.

Supplementary Table 4

Performance Characteristics of QSOX1, IL1RAP, and the QSOX1/IL1RAP ratio in proteomics data and the results of ELISA.

Supplementary Table 5

Metadata for NAFLD patients and control participants involved in the ELISA validation study.