Single-nucleus ATAC-seq elucidates major modules of gene regulation in the development of non-alcoholic fatty liver disease

Fumihiko Takeuchi [1 2] * , Yi-Qiang Liang [1] , Hana Shimizu-Furusawa [3] , Masato Isono [1] , Mia Yang Ang [1 4] , Kotaro Mori [2] , Taizo Mori [5] , Eiji Kakazu [5] , Sachiyo Yoshio [5] , Norihiro Kato [1 2 4]

[1] Department of Gene Diagnostics and Therapeutics, Research Institute, National Center for Global Health and Medicine, Japan.

[2] Medical Genomics Center, Research Institute, National Center for Global Health and Medicine, Japan.

[3] Department of Hygiene and Public Health, School of Medicine, Teikyo University, Japan.

[4] Department of Clinical Genome Informatics, Graduate School of Medicine, The University of Tokyo, Japan.

[5] Department of Liver Diseases, The Research Center for Hepatitis and Immunology, National Center for Global Health and Medicine, Japan.

* Correspondence: fumihiko@takeuchi.name

Abstract

Background: Non-alcoholic fatty liver disease (NAFLD) develops from fatty liver to steatohepatitis during which multiple cell types may play different roles. Aiming to understand the tissue composition of cell types, their gene expression and global gene regulation in the development of NAFLD, we performed single-nucleus and bulk ATAC-seq on the liver of rats fed with a high-fat diet (HFD).

Methods: Male Spontaneously Hypertensive Rats were fed a normal diet or a HFD. Rats fed HFD for 4 weeks developed fatty liver, and those fed HFD for 8 weeks further developed steatohepatitis. Under the washout condition, where 4 weeks of HFD is followed by 4 weeks of a normal diet, fatty liver was partially ameliorated. For each dietary condition, we performed single-nucleus ATAC-seq on one animal and bulk ATAC-seq on four animals.

Results: In accordance with the pathological progression from fatty liver to steatohepatitis, the proportion of inflammatory macrophages dramatically increased. By machine learning, we divided the global gene expression into modules, such that transcription factors in a module regulate the genes in the same module. Consequently, many of the modules rediscovered known regulatory relationship between the transcription factors and biological processes. For the discovered biological processes, we searched core genes, which were defined as genes central regarding co-expression and protein-protein interaction. A large part of the core genes overlapped with previously implicated NAFLD genes.

Conclusions: Single-nucleus ATAC-seq combined with data-driven statistical analysis help elucidate the global gene regulation *in vivo* as a combination of modules and discover core genes of the relevant biological processes.

Keywords

Non-alcoholic fatty liver disease, single-nucleus ATAC-seq, transcription factor, gene regulation

Background

Non-alcoholic fatty liver disease (NAFLD) is a continuum of liver abnormalities ranging from non-alcoholic fatty liver (NAFL) to non-alcoholic steatohepatitis (NASH). It has a global prevalence of ~25% [1]. NAFLD can gradually develop from fatty liver, steatohepatitis and cirrhosis to hepatic cancer. No approved drugs are available for NAFLD, and biomarkers are currently insufficient. The liver is composed of multiple cell types, including hepatocytes, cholangiocytes, stellate cells, endothelial cells, and white blood cells, each of which plays a different role in disease development. Recent single-cell RNA-seq studies have uncovered cell type-specific gene expression changes in NAFLD [2–4].

A key upstream event in gene expression regulation is transcription factors (TFs) binding to DNA. The conventional ChIP-seq assay has a methodological weak point since it is limited to measuring the binding of one TF in bulk of cells. In contrast, single-nucleus ATAC-seq (snATAC-seq) realized simultaneously measuring the binding of all TFs in single cells. In ATAC-seq, open chromatin sites are observable and indicate 1) TF binding at the nearby motifs on the genome and 2) expression of a gene located close to the site. Since previous single-cell analyses of NAFLD were not ATAC-seq but all RNA-seq, TFs binding to chromatin could not be measured.

The tissue composition of cell types changes during disease development in various diseases and affects the outcome. Cell type counting of solid tissue became feasible through the single-cell assay, which is still expensive. Instead, deconvolution, the statistical inference of the cell type composition of tissue samples assayed in bulk, has been used for transcriptome and DNA methylation assays [5, 6]. For ATAC-seq, Li and associates [7] previously proposed an algorithm for deconvolution, but no experimental validation has been done with single-cell data.

Aiming to understand the cell type composition and global TF regulation in NAFLD development, we performed single-nucleus and bulk

81 ATAC-seq on the liver of rats fed or withdrawn from a high-fat diet (HFD). We
82 first analyzed the composition of cell types in the liver and the gene
83 expression separately by cell type. Then, to explore global TF regulation in
84 each cell type, we developed a new statistical method that decomposes TF
85 regulation into major modules, each being characterized by biological
86 processes. We also developed a statistical method to detect core genes for
87 the individual biological processes. Taking snATAC-seq data of an *in vivo*
88 experimental model as the input, our data-driven reductionist workflow
89 proceeds from global TF regulation, TF regulation modules, biological
90 processes to core genes. The core genes thus identified overlapped well with
91 the previously known NAFLD genes, providing experimental validation to our
92 newly developed statistical methods.
93
94 **Results**
95 **Cell type composition of the liver**
96 Spontaneously Hypertensive Rats (SHR) fed HFD for 4 weeks developed fatty
97 liver, and those fed HFD for 8 weeks further developed inflammation and
98 premature fibrosis (steatohepatitis) (Figure 1A; Table S1). Under the washout
99 condition, where 4 weeks of HFD is followed by 4 weeks of a normal diet,
100 fatty liver was partially ameliorated. Along with control rats only fed a normal
101 diet, five dietary conditions were examined.
102 In order to catalog cell types in the liver, we performed snATAC-seq
103 on one animal per dietary condition, five animals in total. We obtained a total
104 of 14,486 nuclei (~3000 per sample), which were grouped into 16 clusters
105 by the similarity of chromatin opening. Hepatocytes (seven clusters),
106 endothelial cells (three clusters), stellate cells (two clusters) and white blood
107 cells (four clusters) separated clearly (Figures 1B and 1C). The clusters for
108 white blood cells corresponded to inflammatory macrophages, non-
109 inflammatory macrophages, T cells and NK cells combined, and B cells.

110 Cholangiocytes were included in the hepatocyte clusters, and neutrophils and
111 Kupffer cells were included in the macrophage clusters.
112 To measure the changes in cell type population through dietary
113 intervention, we performed bulk ATAC-seq on four animals per condition, 20
114 animals in total. Using snATAC-seq as reference, we inferred the cell type
115 composition in the bulk ATAC-seq samples by deconvolution. Beforehand,
116 using the five samples assayed in both single-nucleus and bulk ATAC-seq, the
117 accuracy of deconvolution for ATAC-seq was validated (Figure 1D). The
118 difference of the inferred proportion from the true proportion counted by
119 snATAC-seq had average of zero and standard deviation of 0.035, indicating
120 high accuracy. By using four animals per condition, the standard error would
121 become 0.018. Compared with the normal diet condition, hepatocytes
122 significantly decreased after 8 weeks of HFD (Figure 1E). Inflammatory
123 macrophages largely increased after 4 weeks of HFD and further increased
124 after 8 weeks of HFD. Non-inflammatory macrophages also increased after 8
125 weeks of HFD. Additionally, B cells increased under HFD conditions. In
126 accordance, previous studies of HFD-induced NAFLD animal models reported
127 an increase in liver-resident macrophages [8], recruitment of monocyte-
128 derived macrophages [9], and an increase in intrahepatic B cells [10, 11].
129 Under the washout condition, the cell type composition returned to the same
130 as the normal diet condition. Here, we elucidated how the quantity of cell
131 types changed through dietary intervention, and we next investigate the
132 qualitative change.
133
134 **Differential gene expression under HFD in each cell type**
135 We analyzed cell type-specific differential gene expression by comparing the
136 nuclei of the same cell type between dietary conditions (Table S2; Dataset
137 S1). Hereafter, clusters for inflammatory and non-inflammatory macrophages
138 are combined and regarded as the macrophage cell type. T cells, NK cells,

139 and B cells were not analyzed because the number of nuclei was limited. The
140 biological processes characteristic of the differentially expressed genes were
141 cell type-dependent (Figure 2). Steroid metabolism and fatty acid metabolism
142 genes were expressed in hepatocytes differently under HFD and washout
143 compared with the normal diet. As of inflammation genes, the expression was
144 altered in macrophages after 4 weeks of HFD and washout. The inflammation
145 genes and apoptosis genes were expressed in hepatocytes differently after 8
146 weeks of HFD. Genes for the actin filament-based process were expressed in
147 stellate cells differently after 8 weeks of HFD.

148 Conventional differential gene expression analysis using bulk tissue
149 samples was limited in detecting 1) changes in minor cell types, 2) changes
150 in a cell type with a low basal expression of the gene relative to other cell
151 types, and 3) concurrent changes in gene expression and cell population that
152 mutually cancel, for example, downregulated genes in a cell type that is
153 increasing in population. For the comparison between the 4 weeks of HFD and
154 the normal diet, cell type-specific differential gene expression was cross-
155 checked with bulk tissue transcriptome analysis, and the results were found
156 to replicate under the abovementioned limitations (Figure S1). The
157 differential gene expression was replicated in hepatocytes, but to a lesser
158 degree in minor cell types. For macrophages, whose proportion increased
159 after 4 weeks of HFD, the upregulated genes were replicated in bulk
160 experiments but the downregulated genes were not.

161

162 **Major modules of TF regulation and associated biological processes**

163 To elucidate the global picture of TF regulation, we inferred the regulatory
164 impact of all TFs on all genes and extracted major modules of the regulatory
165 relation. By snATAC-seq, we measured in all nuclei of each cell type (from
166 rats of all dietary conditions) the genome-wide binding of TFs and the
167 expression of genes, and inferred the regulation of the latter by the former

168 by applying machine learning. The regulatory relation was divided into
169 modules. A module comprises a subset of TFs and a subset of genes, such
170 that TFs belonging to a module regulate genes belonging to the same module
171 but do not regulate genes belonging to other modules. In hepatocytes, we
172 obtained four modules (Figure 3A). Modules are simply a grouping of
173 TFs/genes detected by machine learning and have no biological meaning by
174 themselves. By applying gene set enrichment analysis to the genes belonging
175 to a module, but not using the information on TFs, we searched for biological
176 processes that characterize the module (Figure 3B). In the following, we
177 describe modules for hepatocytes, endothelial cells (Figure 4), stellate cells
178 (Figure 5), and macrophages (Figure 6). If a biological process was assigned
179 to a module, we sought the literature that supports the regulation of the
180 biological process by the TFs belonging to the module. In five of the seven
181 discovered modules, there was support for the discovered linkage between
182 the TFs and biological processes, suggesting the validity of the module
183 discovery algorithm.

184 Module 1 of hepatocytes was characterized by the binding of STAT
185 family TFs (namely, STAT5B, STAT4 and STAT3) and by genes for steroid
186 metabolism.

187 Module 2 of hepatocytes, Module 1 of endothelial cells, and Module 1
188 of macrophages were characterized by the binding of AP-1 family TFs (namely,
189 FOS, JUND, BACH1 and BACH2), SMARCC1 and NFKB1, and by genes for
190 TNFα signaling via NF-κB. AP-1 TFs respond to cytokine stimuli [12].
191 SMARCC1 is a member of the SWI/SNF family of proteins, which function as
192 a helicase and interacts with NF-κB [13].

193 Module 3 of hepatocytes was characterized by the binding of TCF/LEF
194 family TFs (namely, LEF1 and TCF7L1) and by genes in the farnesoid X
195 receptor pathway for bile acid synthesis and genes for the glutamine family
196 amino acid biosynthesis. Within the liver lobule, a unit that assembles to form

197 the liver, a central vein is located in the center, the portal triad (portal vein,
198 hepatic artery, and bile duct) is located in the periphery, and the metabolism
199 of hepatocytes in between differs between the pericentral and periportal
200 zones. The glutamine family amino acid biosynthesis genes include *Glul* and
201 *Ass1*, which are highly expressed in the pericentral and periportal zones,
202 respectively [14]. Additionally, the LEF1 TF binds to β-catenin protein and
203 activates the Wnt signaling pathway [15], and liver-specific β-catenin KO mice
204 exhibited defective cholesterol and bile acid metabolism in the liver [16].
205 Moreover, Wnt-activated genes are highly expressed in the pericentral zone.
206 Thus, LEF1 TF presumably regulates bile acid metabolism and zonated gene
207 expression in the liver.
208 Module 2 of endothelial cells was characterized by the DNA binding of
209 AHCTF1 (aka ELYS) and ZNF740 and by genes for angiogenesis. AHCTF1 is a
210 DNA-binding protein that is required in mitosis [17]. ZNF740 activates
211 angiogenesis in pulmonary artery endothelial cells of rats [18].
212 Module 1 of stellate cells was characterized by SOX9 TF binding and
213 genes in the semaphorin-plexin signaling pathway.
214 Module 2 of macrophages was characterized by the binding of Maf
215 family TFs (namely, MAFA, MAFB and MAFG) and by genes for complement
216 system activation. In *Mafb*-deficient macrophages of mice, C1q production
217 decreased [19].
218 Module 3 of macrophages was characterized by the binding of IRF
219 family TFs (namely, IRF1, IRF2 and IRF7), which regulate immune responses,
220 and by genes for angiogenesis. IRF1 contributes to the commitment of pro-
221 inflammatory M1 macrophages, which produce angiogenic stimulators [20].
222
223 **Core genes in specific biological processes**
224 In the previous section, we discovered major modules of TF regulation and
225 then searched across databases, such as Gene Ontology, for biological

226 processes that characterize the modules. We equated a biological process
227 with its comprising genes, and defined them as a gene set (GS). A GS
228 characterizes a module if it significantly overlaps with the genes of the module.
229 Here, we aim to find "core genes" of a biological process. In preparation, we
230 compute the activity of a GS in single nuclei as the "GS activity score" by
231 taking the average binding/expression level of TFs/genes in the set or, more
232 precisely, the first principal component. In hepatocytes, regarding the
233 biological process of TNFα signaling via NF-κB, its GS activity score was the
234 highest among nuclei from 8 weeks of HFD (Figure 7A, top panel).

235 Core genes are the genes that are central regarding two connections
236 of the GS, co-expression and protein-protein interaction. For co-expression,
237 we first compute the correlation coefficient between the GS activity score and
238 the expression level of a gene (or the binding of a TF). The absolute value of
239 the correlation coefficient is required to be large for a core gene. Moreover,
240 the total amount of protein-protein interactions between the gene and the GS
241 is required to be large. Therefore, by plotting the correlation coefficient on
242 the horizontal axis and the total protein-protein interactions on the vertical
243 axis, core genes are expected to appear in the top left or top right corner.
244 This approach can be used to prioritize the known genes in the target GS
245 (Figure 7A, middle panel) and to discover novel genes not included in the GS
246 (Figure 7A, bottom panel). In the following, we inspect some of the biological
247 processes that appeared in the previous section.

248 In hepatocytes, the GS activity score for TNFα signaling via NF-κB
249 (Figure 7A) was the highest after 8 weeks of HFD. TF binding that was
250 positively correlated and had abundant protein-protein interactions included
251 FOS/JUN and NFKB1/RELA. The positively correlated core genes were *Tnf*,
252 *Nfkb1*, *Bcl2l1* and *Il1r1*. The negatively correlated core gene was *Cxcl12*.

253 In endothelial cells, the GS activity score for TNFα signaling via NF-
254 κB (Figure 7B) was higher under HFD conditions. The positively correlated

255 core genes were *Socs3* and *Pecam1*. The negatively correlated core genes
256 were *Tlr2*, *Irs2*, *Tlr4*, *Il15*, *Ccr5* and *Cxcr4* (receptor for the chemoattractant
257 CXCL12 [21]).

258 In macrophages, the GS activity score for TNFα signaling via NF-κB
259 (Figure 7C) was higher after 4 weeks of HFD. TF binding that was positively
260 correlated and had abundant protein-protein interactions included JUN and
261 KLF6. The positively correlated core genes were *Cd44*, *Fn1* and *Cxcr4*. KLF6
262 regulates the switching of macrophages between pro- and anti-inflammatory
263 [22]. In accordance, *Lyz2* and *Marco*, the marker genes for inflammatory and
264 non-inflammatory macrophages, showed a correlation of 0.39 and −0.49,
265 respectively, with KLF6 binding in our single-nucleus data.

266 In hepatocytes, the GS activity score for steroid metabolism (Figure
267 S2A) was the highest after 8 weeks of HFD and to a lesser degree after 4
268 weeks of HFD. The positively correlated core genes included *Abcg1*, *Abca1*
269 and *Scd*. The negatively correlated core genes included *Sqle*, *Hmgcr*, *Apoa1*,
270 *Sult2a1*, *Acox2* and *Alb*. Among the 15 cytochrome P450 genes showing a
271 strong negative correlation (correlation smaller than −0.3 and protein-protein
272 interactions within the top 5%), 13 were oxidoreductase of EC 1.14.14.1. For
273 reference, we labeled the binding of steroid sensing TFs, NR1H2/3 (aka
274 LXRα/β), SREBF1 and SREBF2, which were negatively correlated with the GS
275 expression.

276 To investigate lobular zonation genes in hepatocytes, the GS was
277 taken from the list of zonation genes in normal mice (q-value < 10^{-6} in Table
278 S3 in [14]). The GS activity score showed a modest difference between
279 dietary conditions (Figure S2B), which agrees with zonation being a normal
280 phenomenon unrelated to NAFLD. When classified by the correlation with the
281 GS activity score, the genes with correlation coefficient > 0.3 were all
282 pericentral and those with correlation coefficient < −0.3 were all periportal,
283 showing that the GS activity score rediscovered zonation.

284 In stellate cells, the GS activity score for semaphorin-plexin signaling
285 was higher at 8 weeks of HFD (Figure S2C). The positively correlated core
286 gene was *Nrp2*. The negatively correlated core genes were *Nrp1*, *Plxnd1*,
287 semaphorin genes (*Sema3d*, *Sema3e* and *Sema6d*) and *Ntn1*.
288
289 **Discussion**
290 By performing single-nucleus and bulk ATAC-seq on the liver of rats fed or
291 withdrawn from the HFD, we studied the transition in tissue composition of
292 cell types and cell type-specific gene regulation in a rat model of NAFLD. The
293 accuracy of ATAC-seq-based deconvolution was validated and used to analyze
294 cell type composition. In accordance with the pathological progression from
295 fatty liver to steatohepatitis that could have occurred between 4 and 8 weeks
296 of the HFD intervention in our rat model, the proportion of inflammatory
297 macrophages dramatically increased. In contrast, after 8 weeks of HFD, the
298 proportion of hepatocytes largely decreased. In the washout condition, where
299 4 weeks of HFD was followed by 4 weeks of normal diet, the cell type
300 composition returned similar to rats only fed with the normal diet.
301 From the global analysis of TF binding and gene expression in single
302 nuclei, we identified major modules of TF regulation (Figure 8), which could
303 be shared between different cell types or be specific to certain cell types. One
304 module shared among hepatocytes, endothelial cells, and macrophages was
305 characterized by the binding of AP-1 TFs and a biological process of TNFα
306 signaling via NF-κB associated with core genes that appear to be specific to
307 cell types. For many of the discovered modules, TFs in a given module were
308 known to regulate biological processes assignable to the genes in the same
309 module, suggesting the validity of our module discovery algorithm. Finally,
310 for the biological processes that emerged in TF regulation analysis, by
311 incorporating known protein-protein interactions, we could identify core
312 genes, many of which overlap with previously implicated NAFLD genes (see

313 below).

314 We consider that the gene regulatory machinery of a cell is modular
315 and some of the modules would become active when perturbed by HFD. As
316 an analogy, a car consists of many machinery parts, and on a cold sunny day
317 the engine and air conditioning are active but the wiper is not. To discover
318 gene regulatory modules that respond to HFD stimulus, we integratively
319 analyzed the TF binding and gene expression in nuclei derived from rats under
320 various dietary conditions. In each cell type, we used machine learning to
321 decompose the global variation of TF regulation into major modules, where
322 the TFs belonging to a given module appear to regulate the genes that belong
323 to the same module. In the modules identified as such, the TFs and the
324 function of regulated genes agree with the literature: e.g., the AP-1 TFs that
325 regulate inflammation genes (Module 2 of hepatocytes, Module 1 of
326 endothelial cells and Module 1 of macrophages) are known to respond to
327 cytokine stimuli [12], and LEF1 that regulates zonation (Module 3 of
328 hepatocytes) is known to activate Wnt signaling [15]. Previous snATAC-seq
329 studies have characterized the difference between cell types [23] using TF
330 binding. In contrast, we illustrated how TFs control the states of cells
331 belonging to one cell type (e.g., the hepatocytes with various degrees of
332 inflammation) possibly in a cell-type specific manner (e.g., a distinct set of
333 core genes and other target genes associated with AP-1 TFs in different cell
334 types).

335 One novel method developed in this article is the computation of the
336 GS activity score in single cells to indicate the activity of a GS, as if it were
337 the expression of a representative gene. In experiments, gene expression
338 can be directly measured. The activity of a pathway is of interest but not
339 directly measurable. One way would be to take the average expression of
340 genes in that pathway, but this is not optimal for the following reasons. A
341 simple average would give equal weight to both core and periphery genes in

342 the pathway, increasing noise from the latter. Moreover, activator and
343 suppressor genes contribute to the pathway in opposite directions. Instead of
344 a simple average, we used the first principal component to deal with these
345 issues. The principal component approach was adapted from the PLAGE
346 program [24], which is used to compute pathway activity scores in bulk tissue
347 specimens. The GS activity score could improve the understanding of
348 biological processes. For example, *Abcg1* and *Sqle* correlated positively and
349 negatively, respectively, with the activity score for the steroid metabolism
350 pathway (Figure S2A). As this score became higher under the HFD, the GS
351 activity represented by the score presumably works in response to the
352 overdose of cholesterol contained in the diet. Thus, a positive correlation of
353 *Abcg1* (for lipid efflux) and a negative correlation of *Sqle* (for sterol
354 biosynthesis) are reasonable. Moreover, the cell type-specific nature of the
355 GS activity score could unravel cell type-specific dynamics; the inflammation
356 process peaked under 4 weeks of HFD in macrophages and under 8 weeks of
357 HFD in hepatocytes (Figure 7). The GS activity score can also be applied to
358 single-cell and single-nucleus RNA-seq data.

359 Another novel method in this article is finding the core genes of
360 biological processes. Given a GS comprising known genes for the process, we
361 identified core genes as those that show a strong correlation with the GS
362 activity score as well as abundant protein-protein interactions with the GS.
363 We obtained evidence supporting the former from single-nucleus experiment
364 data and the latter from a public database. We sought literature indicating
365 that core genes cause NAFLD or serve as biomarkers for NAFLD and found
366 indications for a large part of the core genes (see below): for the TNFα
367 signaling via NF-κB process, 4 of 5 in hepatocytes, 4 of 8 in endothelial cells,
368 and 1 of 3 in macrophages; for steroid metabolism 3 of 9 in hepatocytes; for
369 semaphorin-plexin signaling 3 of 7 in stellate cells. The presence of
370 substantial overlap between the current findings and known NAFLD genes

371 provides a proof-of-concept for our data-driven approach, using snATAC-seq
372 data from a rat model of dietary intervention as input, extracting modules,
373 biological processes, and then core genes. Regarding TNFα signaling via NF-
374 κB in hepatocytes, in animal studies, *Tnf* KO mouse was protected from
375 NAFLD [25], *Nfkb1* KO mouse exacerbated NAFLD [26], and *Il1r1* KO mouse
376 had less hepatic triacylglycerol [27]. CXCL12 (aka SDF1) induced migration
377 of stellate cells [28], endothelial cells [21] and $CD4^+$ T cells [29]. In human
378 biomarker studies, higher circulating TNFα level was associated with NAFLD
379 [30] and future NAFLD [31]. In NAFLD patients, soluble IL1R1 was higher
380 (Ajmera et al., 2017). Regarding TNFα signaling via NF-κB in endothelial cells,
381 in animal studies, *Pecam1* KO mouse developed NASH [32], whereas *Tlr4* KO
382 mouse [33] and *Il15* KO mouse [34] were protected from NAFLD. *Ccr5*
383 antagonist eased NAFLD in mice [35]. Regarding TNFα signaling via NF-κB in
384 macrophages, *Cd44* KO mouse had less liver injury under NAFLD inducing
385 diet, and NAFLD patients had elevated serum soluble CD44 [36]. Regarding
386 steroid metabolism in hepatocytes, *Scd1* KO mouse exacerbated NAFLD [37],
387 and *Acox2* KO mouse developed NAFLD [38]. NAFLD patients had lower
388 serum APOA1 [39, 40]. Regarding semaphorin-plexin signaling in stellate
389 cells, in animal studies, *Nrp2* conditional KO reduced liver fibrosis in mice
390 [41]. *Nrp1* KO attenuated cell motility and overexpression increased motility
391 [42]. *Sema3e* KO eased liver fibrosis in mice [43].
392 In search of drug targets or biomarkers of NAFLD, we looked up
393 proteins whose subcellular localization is extracellular or plasma membrane
394 proteins [44] among the genes strongly upregulated under a dietary
395 intervention (Table S2) or the core genes positively correlated with a NAFLD-
396 related activity. Previous reports of NAFLD biomarkers in humans are cited.
397 For hepatocytes, we found *Il1r1* [45], *Il1r2* and *Tnf* [30]. For endothelial cells,
398 there was *Hgf* [46–48]. *Pecam1* was listed for both endothelial and stellate
399 cells. For stellate cells, there were *Faslg* [49] and *Nrp2*. For macrophages,

400 there were RT1-Db1, *Cd44* [36], *Cxcr4* and *Fn1*. We also searched DGIdb
401 (2022-Feb) for existing drugs that can potentially be repurposed and found
402 antagonist/antibody/inhibitor of upregulated or positively correlated genes,
403 *Bcl2l1*, *Cxcr4*, *Hgf*, *Il1r1* and *Tnf*, and agonist of downregulated or negatively
404 correlated genes, *Ccr5*, *Cxcr4*, *Tlr2* and *Tlr4* (Table S3). *Cxcr4* was positively
405 correlated with TNFα signaling via NF-κB in macrophages but negatively
406 correlated in endothelial cells.

407 This study has several limitations. Firstly, it was difficult to distinguish
408 TFs with very similar motifs. A specific TF was regarded to bind a specific site
409 if the chromatin was open near the site and if the DNA sequence of the site
410 matched with the known TF-binding motif. However, instead, a different TF
411 with a similar motif could have bound there. Secondly, the interpretation of
412 gene expression requires caution. Gene expression can be quantified by the
413 amount of either A) mRNA transcription per time, B) mRNA in the nucleus,
414 C) mRNA in the cell, D) protein translation per time, or E) protein.
415 Measurement for bulk tissue is conducted by ATAC-seq (A), RNA-seq (C),
416 Ribo-seq (D), or Western blotting (E), and measurement for single
417 nucleus/cell is performed by single-nucleus ATAC-seq (A), single-nucleus
418 RNA-seq (B), or single-cell RNA-seq (C). While our measurement, A, could
419 reflect the state of gene regulation, it could differ from E, which reflects the
420 state of metabolism, and the fold-change value would differ among the types
421 of gene expression. Thirdly, the TF regulation could differ between rat, mouse
422 and human.

423 To discover TF modules, we used machine learning whose
424 performance can be affected by the choice of hyperparameters. Under a
425 hyperparameter that allows the machine learning model to be too complex
426 and flexible, compared to the amount of data, the model can pick up noise
427 as signal and overfit. We chose hyperparameters based on previous empirical
428 validations (for the number of candidate regulators at each node in GENIE3)

429 or residual error (for the number of factors in NMF) but did not explore
430 hyperparameters extensively. We assessed the adopted hyperparameter post
431 hoc. Under a range of hyperparameters, we randomly divided the hepatocytes
432 into two sets of halves, performed TF module discovery in each set and
433 assessed the consistency. We did not observe overfitting for the
434 hyperparameters adopted in this study (Figure S3). Nevertheless, an
435 assessment using data from an independent experiment is needed.
436
437 **Conclusions**
438 By performing single-nucleus ATAC-seq in a rat model of NAFLD and using
439 novel statistical methods, we elucidated a global picture of *in vivo* TF
440 regulation in each cell type as a set of modules and discovered core genes
441 for NAFLD-relevant biological processes. The core genes can serve as novel
442 candidates responsible for NAFLD and our statistical methods can be applied
443 to elucidate the TF regulation in biological systems other than NAFLD.

Methods

Animal experiment

We used male rats of SHR/Izm strain, which is an inbred genetic model of essential hypertension. The strain was supplied by Japan SLC. Rats were weaned at 4 weeks after birth and placed on a normal diet (SP diet, Funabashi Farm, Japan). Adult rats were fed the normal diet or a high-fat atherogenic diet (24% fat, 15% protein, 5% cholesterol, 2% cholic acid) according to the study design. Compared to models based on nutritional deficiency (such as methionine and choline-deficient diet) or chemical-induced models (such as carbon tetrachloride), HFD models resemble more human NAFLD, both pathophysiologically and phenotypically [50]. SHR can mimic NAFLD under a hypertension background. The rats were killed under pentobarbital anesthesia (200 mg/kg via intraperitoneal infusion), and the organs were excised and immediately frozen at −70 C. We used the liver's largest lobe (left lateral lobe) for pathology imaging and the right segment of the second largest lobe (left median lobe) for genomics analysis. The measured outcomes were genomic and biochemical assays of the liver, pathology imaging of the liver, biochemical assays of the serum, organ and body weight, and blood pressure. The experimental unit was the individual animal. No animal was excluded during the experiment. The allocation of animals to dietary conditions was not randomized. We controlled the dietary conditions but did not control for other possible confounders. The group allocation was known to all experimenters except for the pathologist, who inspected unlabeled preparations. All animal experiments conformed to the Guidelines for Animal Experiments of the National Center for Global Health and Medicine (NCGM), and were performed in the animal facility of NCGM. The animal experiments and the protocol were registered and approved by the Animal Research Committee of NCGM (permission number 21009).

473 **Library preparation and sequencing for snATAC-seq**

474 Nuclei were isolated from flash-frozen liver tissue according to the 10x
475 Genomics protocol CG000212 (samples m154207 and m154211) or using the
476 Minute Single Nucleus Isolation Kit for Tissues/Cells (Invent Biotechnologies,
477 SN-047) (samples m167108, m167203 and m168101). For each sample, the
478 sequencing library was prepared using the Chromium Next GEM Single Cell
479 ATAC Reagent Kits v1.1 (10x Genomics, CG000209) and sequenced using the
480 Novaseq 6000 sequencer (Illumina).

481

482 **Library preparation and sequencing for bulk tissue ATAC-seq**

483 Bulk tissue ATAC-seq was performed according to the Omni-ATAC protocol
484 [51] using flash-frozen liver tissue. The library was sequenced using the
485 NextSeq sequencer (Illumina).

486

487 **Measuring triglycerides in liver tissue**

488 Triglyceride in the liver tissue was measured enzymatically using the
489 Adipogenesis Colorimetric/Fluorometric Assay Kit (BioVision, K610-100).
490 Tissue homogenate was obtained by dispersing 10 mg tissue in 1 ml lipid
491 extraction solution under 2200 rpm for 1 min, using a handheld disperser
492 (KINEMATICA, PT 1300 D). For fatty liver tissues, we prepared 1/3 dilution of
493 the homogenate by adding lipid extraction solution and examined both the
494 original and the dilution. The triglyceride was measured in a colorimetric
495 assay according to the manual. Triglyceride quantity in mol unit was
496 converted to g unit by multiplying 885.7.

497

498 **Measuring total cholesterol in liver tissue**

499 Total cholesterol in the liver tissue was measured enzymatically using the
500 Total Cholesterol and Cholesteryl Ester Colorimetric Assay Kit II (BioVision,
501 K623-100). Tissue homogenate was obtained by dispersing 15 mg tissue in

502 600 ul $CHCl_3$:IPA:NP-40 (7:11:0.1) at 2200 rpm for 1 min, using a handheld
503 disperser (KINEMATICA, PT 1300 D). The homogenate was then centrifuged
504 for 5 min at 15,000 *g*. Thereafter, we transferred 200 ul of the liquid phase
505 to a new tube, dried with air-flow-in at 40 C for 10 min using a centrifugal
506 concentrator (TOMY Digital Biology, CC-105), and dissolved the dried lipids
507 with 400 ul Cholesterol Assay Buffer. Total cholesterol was measured in a
508 colorimetric assay according to the manual.
509
510 **Data cleaning for snATAC-seq**
511 Data cleaning was conducted separately by samples. ATAC-seq fragments
512 were assigned to single nuclei, and nuclei of good quality were retained as
513 described below. In principle, one nucleus and one barcoded bead are
514 contained in one droplet of the Chromium assay; thus, ATAC-seq fragments
515 from the same nucleus are identifiable by the shared unique barcode. Using
516 the Cell Ranger ATAC software (version 2.0.0), we mapped the paired-end
517 reads to the mRatBN7.2/rn7 rat reference genome and identified ATAC-seq
518 fragments, which were assigned to single barcodes.
519 Occasionally in the Chromium assay, more than one barcoded beads
520 are contained in one droplet, in which case, excessive ATAC-seq fragments
521 would be assigned in common to the barcodes. Such cases were detected,
522 and the barcodes were unified and renamed using the bap2 software (version
523 0.6.7) [52]. In the downstream analysis, we can assume that one barcode
524 corresponds to one droplet.
525 We managed the snATAC-seq data using the ArchR package (version
526 1.0.1) [53] of the R software (version 4.1.1). ATAC-seq fragments from the
527 autosomes and chromosome X were retained. We retained nuclei that had
528 abundant and enriched fragments around transcription start sites (TSS) as
529 follows: ≥1000 fragments within 1 kb to TSS, the proportion of fragments
530 within 1 kb to TSS being ≥7.5% (samples m154207, m154211 and m167203)

531 or ≥15% (samples m167108 and m168101), and Tn5 insertions (which occur
532 at both ends of a fragment) within 25 bp to TSS occurring ≥4 times frequently
533 than the insertions at 1900–2000 bp from TSS.

534 For each nucleus, the proportion of fragment counts from a
535 chromosome (e.g., chromosome 1) is approximately constant, being
536 proportional to the chromosome length, but could deviate if a droplet is
537 contaminated by free DNA from broken nuclei [54]. Therefore, we
538 represented the proportion by a matrix indexed by nuclei (rows) and
539 chromosomes (columns). To standardize the matrix, we subtracted column-
540 wise averages and divided by the standard deviation. By taking the sum of
541 the absolute values of a row, we defined the deviation score of a nucleus. In
542 order to omit contaminated droplets, we discarded nuclei with a score ≥40.

543 Since Tn5 inserts less efficiently into nucleosome-associated DNA,
544 ATAC-seq fragment size distribution shows periodicity [23]. Using the
545 program provided by Cusanovich et al.
546 (https://atlas.gs.washington.edu/mouse-atac/docs/), for each nucleus, we
547 computed the spectral densities for the 100–300 bp range, which is the pitch
548 of nucleosomes. In order to omit nuclei with an aberrant chromatin state, we
549 discarded nuclei below the 1st or above the 99th percentile in this score.

550 When two or more nuclei are contained in one droplet, called a
551 multiplet, those nuclei share the same barcode, violating the one-nucleus-to-
552 one-barcode assumption. We removed the barcodes for multiplets in two
553 ways. First, for the set of ATAC-seq fragments assigned to a barcode, we
554 inferred the number of homologous chromosomes from which the fragments
555 were derived. The number is analogous to ploidy, but differs because the
556 container is not a cell, but a droplet. By counting overlaps among the
557 fragments, the number of homologous chromosomes was estimated to be
558 either 2*n*, 4*n* or ≥6*n* using the R package scPloidy (version 0.3.0). Since the
559 majority of hepatocytes in rat liver are diploids or tetraploids [55] and other

560 cell types in the liver are diploids, ≥6*n* droplets were regarded as possible
561 multiplets and discarded. Second, we detected doublets formed by two
562 heterogeneous cell types using ArchR. Droplets with doublet enrichment
563 score >10 were discarded.
564
565 **Chromatin accessibility peak discovery and nucleus clustering for**
566 **snATAC-seq**
567 The Tn5 insertions observed in ATAC-seq occur at open chromatin regions,
568 such as promoters and enhancers, and the genome-wide profile of Tn5
569 insertions characterizes the types of cells. We clustered the nuclei based on
570 the similarity of Tn5 insertions. Firstly, we represented the Tn5 insertions in
571 snATAC-seq as a "tile matrix": a binary matrix indexed by tiles of 500 bp
572 width across the genome (rows) and nuclei (columns), coding the occurrence
573 (1) or non-occurrence (0) of Tn5 insertions. Using ArchR, we obtained a
574 matrix with 5,230,329 rows and 14,615 columns. We then grouped the nuclei
575 according to the similarity of corresponding columns (see next section). For
576 each group, pseudo-bulk ATAC-seq datasets were generated, and Tn5
577 insertion peaks were computed using ArchR and the MACS2 program (version
578 2.1.4) [56]. We obtained 599,035 chromatin accessibility peaks of width 501
579 bp.
580 Secondly, we represented the Tn5 insertions as a "peak matrix": a
581 binary matrix indexed by peaks (rows) and nuclei (columns), coding the
582 occurrence or non-occurrence of Tn5 insertions. Again, we grouped the nuclei
583 according to the similarity of corresponding columns. The nuclei clustering
584 based on the peak matrix was used in downstream analysis, as it would be
585 more accurate than the clustering based on the tile matrix.
586
587 **Algorithm for dimensionality reduction and nucleus clustering**
588 The tile and peak matrices are large binary matrices indexed by genomic

positions (rows) and nuclei (columns). Since the column vectors are extremely high dimensional, we need to reduce the dimensions in order to properly analyze the similarity among nuclei. For the dimensionality reduction, we used latent semantic indexing [57] and singular value decomposition (SVD). The dimensionality reduction was performed for all samples combined. Since experimental conditions, including Tn5:DNA ratio, could differ among the samples, we introduced a sample batch correction step. We applied the same algorithm to the tile and peak matrices.

The input binary matrix (f_{ij}) with n columns was treated as the term-document matrix in the latent semantic indexing literature. We defined the inverse document frequency (IDF) as

$$g_i = \log_2\left(\frac{n}{\sum_j f_{ij}}\right).$$

For the *i*-th row of the matrix weighted by IDF, the variance becomes

$$g_i{}^2\left(\frac{\sum_j f_{ij}}{n}\right)\left(1-\frac{\sum_j f_{ij}}{n}\right)$$

and ranges between zero and one. In order to restrict noise, we retained rows with IDF-weighted variance >0.4 of the tile matrix and >0.1 of the peak matrix. We defined the document normalization factor as

$$d_j = \frac{1}{\sqrt{\sum_i f_{ij}{}^2}}.$$

We applied sample batch correction independently to the rows of the matrix. For row i, the normalized elements $d_j \cdot f_{ij}$ vary among nuclei j, but the average across many nuclei in a sample is anticipated to be approximately constant. The difference in the averages among samples was regarded as a batch effect and was corrected by scaling. For the nucleus j belonging to the sample s, the correction factor is

$$\gamma_{ij} = \frac{\mathrm{Mean}\left[\{d_{j'} \cdot f_{ij'} : 1 \le j' \le n\}\right]}{\mathrm{Mean}\left[\{d_{j'} \cdot f_{ij'} : j' \text{ belongs to sample } s\}\right]}. \quad (1)$$

When the denominator equals zero, γ_{ij} is set as one. The above correction

requires no knowledge of cell types. If cell type classification of nuclei is available, the correction can be refined. For the nucleus j of cell type c belonging to sample s, the correction factor becomes

$$\gamma_{ij} = \frac{\mathrm{Mean}\left[\{d_{j'} \cdot f_{ij'} : j' \text{ is cell type } c\}\right]}{\mathrm{Mean}\left[\{d_{j'} \cdot f_{ij'} : j' \text{ is cell type } c \text{ and belongs to sample } s\}\right]}. \quad (2)$$

Since we had no knowledge of cell type in advance, we first applied the batch correction in (1) and then inferred coarse cell type classification. Using the coarse classification, we applied the batch correction in (2). To avoid the inflation of γ_{ij} under small denominator, for each nucleus j, we computed the square of the third quartile of γ_{ij} divided by the second quartile, and imposed this upper limit.

Dimensionality reduction and nucleus clustering were performed in the following steps:

1. Compute the sample batch correction factor γ_{ij} in (1).
2. Compute the matrix $(g_i \cdot d_j \cdot f_{ij} \cdot \gamma_{ij})$ and centralize each row by subtracting the average.
3. Compute truncated SVD, UDV^T, for the 30 largest singular values using the R package irlba (version 2.3.3) [58].
4. Take the truncated SVD for the two largest singular values, project the nuclei to the VD coordinates, and detect and discard outlier nuclei. Outliers were detected based on the sum of the squared distance to the five nearest neighbors.
5. Take the truncated SVD for the non-noise singular values, and project the nuclei to the VD coordinates. As the 11th and later singular values bottomed out and the corresponding columns of V appeared randomly distributed in our data, we took the ten largest singular values.
6. Identify clusters of nuclei by computing the shared nearest-neighbor graph with the R package Seurat (version 4.0.0) [59] and optimizing the modularity with the Leiden algorithm (version 0.8.4) [60]. By merging

643 the clusters, define coarse clustering that clearly separates in the VD
644 coordinates. In our data, the coarse clusters corresponded to hepatocytes,
645 stellate cells, endothelial cells and white blood cells.
646 7. Compute the sample batch correction factor γ_{ij} in (2) by incorporating
647 the coarse clustering of nuclei.
648 8. Compute the matrix $(g_i \cdot d_j \cdot f_{ij} \cdot \gamma_{ij})$ and centralize each row by
649 subtracting the average.
650 9. Compute truncated SVD, UDV^T, for the 30 largest singular values.
651 10. Take the truncated SVD for the non-noise singular values, and project the
652 nuclei to the VD coordinates. We took the eight largest singular values.
653 11. Identify clusters of nuclei by the shared nearest-neighbor modularity
654 optimization.
655
656 **Cell type assignment for snATAC-seq**
657 For each nuclei cluster obtained above, we assigned a cell type based on
658 chromatin accessibility at known marker genes. For each cluster, chromatin
659 accessibility at the TSS and gene body of the genes were visually inspected
660 using the browser of ArchR. We used the following marker genes in the rat:
661 - *Alb*, *Apoc3*, *Cyp2e1* and *Cyp2f4* for hepatocyte
662 - *Epcam* and *Krt19* for cholangiocyte
663 - *Stab2* for endothelial cell
664 - *Dcn*, *Des*, *Lrat* and *Acta2* for stellate cell
665 - *Csf3r* for granulocyte
666 - *Elane*, *Ly6g6c*, *Ly6g6d*, *Ly6g6e* and *Mpo* for neutrophil
667 - *Csf1r* for macrophage (including Kupffer cell)
668 - *Lyz2* for inflammatory macrophage
669 - *Marco* for non-inflammatory macrophage
670 - *Cd3g* and *Gata3* for T cell
671 - *Gzma* and *Prf1* for cytotoxic T cell and NK cell

- *Ebf1* for B cell

Data cleaning for bulk ATAC-seq

According to the ATAC-seq data processing workflow [61], we trimmed reads, mapped reads to the reference genome, removed mitochondrial DNA reads, subsampled to equate the library complexity, and removed PCR duplicates.

Inferring cell type composition of bulk tissue samples

We inferred the cell type composition of liver tissue samples by comparing their bulk ATAC-seq data with the reference data generated by snATAC-seq. As references, pseudo-bulk ATAC-seq data of pure cell types were computed from snATAC-seq results. We combined several published deconvolution methods. Let the indices be i for a chromatin accessibility peak, j for a cell type, and k for a sample, either assayed in bulk, in single nuclei or both.

By adding Tn5 insertion counts of snATAC-seq over a_{jk} nuclei of cell type j from sample k, we obtain Tn5 insertion count u_{ijk} at peak i. The counts across all peaks are represented as a vector $\boldsymbol{u}_{jk}$ in bold font, in which the index i is dropped. We quantile normalize the vectors $\boldsymbol{u}_{jk}$ and obtain vectors $\boldsymbol{v}_{jk}$. Although quantile normalization is not a linear transformation, we can regard the total number of nuclei in $\boldsymbol{v}_{jk}$ as

$$b_{jk} = \frac{\sum_i v_{ijk}}{\sum_i u_{ijk}} \cdot a_{jk}.$$

We factorize the count as

$$b_{jk} \approx c_j \cdot d_k,$$

where c_j represents how many nuclei are included in the normalized ATAC-seq library for pure cell type j and d_k adjusts for experimental batch effect between samples. Specifically, we solve a linear regression whose dependent variable is the logarithm of b_{jk} and independent variables are the indicator functions of j and k, and take the exponential of the obtained coefficients. For the solution to be unique, we set $d_1 = 1$. We represent by C the diagonal

matrix with c_j in the diagonal. The values of c_j for hepatocytes were half the values for other cell types due to higher ploidy of hepatocytes.

We used the DeconPeaker software (version 1.0) [7] to select the peaks (rows) that discriminate cell types. DeconPeaker selects discriminating peaks based on *t*-tests while minimizing the conditional number. Let $\tilde{\boldsymbol{v}}_{jk}$ be the vector truncated from $\boldsymbol{v}_{jk}$ accordingly. The vectors $\tilde{\boldsymbol{v}}_{jk}$ across the samples from the same cell type j are approximately equal, and their average becomes the reference vector $\boldsymbol{w}_j$ for chromatin accessibility:

$$w_{ij} = \mathrm{mean}_k\{\tilde{v}_{ijk}\}.$$

We represent by W the matrix whose columns are $\boldsymbol{w}_j$. By regarding the samples as replicates, the error variance for peak i was estimated as

$$e_i^2 = \mathrm{mean}_j\left\{\mathrm{var}_k\{\tilde{v}_{ijk}\}\right\}.$$

For samples assayed in snATAC-seq, we can generate pseudo-bulk chromatin accessibility for mixtures based on W. For sample k, the proportion of cell type j is $a_{jk}/(\sum_{j'} a_{j'k})$, and the mixture becomes

$$\boldsymbol{x}_k = \sum_j \frac{a_{jk}}{(\sum_{j'} a_{j'k})} \cdot \frac{1}{c_j} \cdot \boldsymbol{w}_j.$$

Here, the vector is scaled to one nucleus (under the experimental condition of sample $k = 1$). In other words, each element is the average count of Tn5 insertions at the site in one nucleus.

The bulk ATAC-seq results of mixture samples are quantile normalized and represented as vector $\boldsymbol{y}_k$ for sample k. The elements y_{ik} are taken only for the peaks i that were selected by DeconPeaker. Five samples were assayed both in bulk and in single nuclei, for which we have vectors $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$. Although, in principle, the two vectors should be proportional, the ratio y_{ik}/x_{ik} actually varies by peaks, due to measurement differences between different assay technologies. Correction of between-peak variability can improve deconvolution [62]. We compute the correction factor as

$$f_i = \exp\left(\text{mean}_k\left\{\log\left(\frac{y_{ik}}{x_{ik}}\right)\right\}\right).$$

To avoid over-correction, we cap f_i to range between ${Q_1}^2/Q_2$ and ${Q_3}^2/Q_2$, where Q_1, Q_2, Q_3 are the quartiles of the original value. We represent by F the diagonal matrix with f_i in the diagonal. A bulk ATAC-seq experiment $\boldsymbol{y}_k$ after correction becomes $F^{-1}\boldsymbol{y}_k$. Notice that the vector is scaled to one nucleus.

For deconvolution of a bulk sample k, we first compute $\boldsymbol{\theta}_k$ that approximates

$$F^{-1}\boldsymbol{y}_k \approx W\boldsymbol{\theta}_k$$

under $\theta_{jk} \geq 0$. Using robust linear regression [63], we solve

$$F^{-1}\boldsymbol{y}_k = W\boldsymbol{\theta}_k + \boldsymbol{\varepsilon},$$

$$\varepsilon_i \sim N\left(0, e_i^2\sigma^2\right),$$

where σ^2 is the parameter for error variance. Peaks with large variability e_i^2 are allowed to have a larger variance [64]. To avoid over-correction, we cap e_i^2 to range between ${Q_1}^2/Q_2$ and ${Q_3}^2/Q_2$, where Q_1, Q_2, Q_3 are the quartiles of the original value. The negative elements of the estimate $\widehat{\boldsymbol{\theta}_k}$ are set to zero. When computing $\boldsymbol{\theta}_k$, the independent variable $F^{-1}\boldsymbol{y}_k$ was scaled to one nucleus, and the regressors $\boldsymbol{w}_j$ were scale to c_j nuclei. We multiply $\widehat{\boldsymbol{\theta}_k}$ by C and obtain $C\widehat{\boldsymbol{\theta}_k}$, which is the fractional number of nuclei from each cell types included per nucleus of sample k, which in spirit is the cell type composition. Although the sum of the elements in $C\widehat{\boldsymbol{\theta}_k}$ was approximately equal to one, in order to be precise, we rescale to obtain the cell type composition for sample k as

$$\frac{C\widehat{\boldsymbol{\theta}_k}}{\sum_j c_j\widehat{\theta_{jk}}}.$$

Comparing cell type composition among dietary conditions

For each cell type, the imputed composition in bulk samples was compared among dietary conditions. We performed multiple regression where the dependent variable was the composition and independent variables were

756 indicators of dietary condition, with the normal diet being the reference. Using
757 Bonferroni's correction for multiple testing of seven cell types and three
758 conditions (apart from normal diet), the significance level was set to
759 $0.05/[7 \times 3]$.
760
761 **Quantifying gene expression in snATAC-seq**
762 For each nucleus, the expression level of a gene was inferred from the number
763 of Tn5 insertions in the gene body and vicinity. The targeted biological
764 phenomenon is the chromatin opening of the gene rather than the amount of
765 mRNA in a nucleus or cell. Using ArchR, we computed the Gene Score Matrix,
766 which is indexed by genes (rows) and nuclei (columns) and represents the
767 insertion counts. As the matrix is analogous to the count matrix obtained in
768 single-cell RNA-seq, we applied library size normalization of the R package
769 Linnorm (version 2.16.0) [65] and imputation of the R package SAVER
770 (version 1.1.2) [66]. Before imputation, 47% of the matrix elements were
771 zero, and after imputation, 0.2% were set to the lowest value (0.001). We
772 excluded low expression genes, where the lowest value occurred in ≥1% of
773 nuclei. We took log2 of the matrix and applied quantile normalization. The
774 obtained matrix was used for inspecting cell type marker genes. We applied
775 SVD and subtracted the first two components from the matrix, which captured
776 the sample batch effect. Although the two components also captured major
777 cell type differences, we anticipated that the removal would not harm
778 downstream analysis, which compares nuclei of the same cell type between
779 samples.
780
781 **Quantifying TF binding in snATAC-seq**
782 If the DNA sequence of a chromatin accessibility peak includes the motif of a
783 TF, Tn5 insertion in the peak suggests the binding of the corresponding TF.
784 For each nucleus, by aggregating over all peaks, we can infer the genome-

785 wide binding of the TF. This is obtained as the chromVAR [67] deviation z-
786 score in ArchR. The CIS-BP database of TF binding motifs [68] was used. In
787 the same way as we did for the Gene Score Matrix, we applied quantile
788 normalization to the Motif Matrix and subtracted the first two components of
789 SVD.
790
791 **Gene set enrichment analysis of differentially expressed genes under**
792 **HFD**
793 To seek which biological processes and/or pathways are differentially
794 expressed under a particular dietary intervention, we performed gene set
795 enrichment analysis. The target GS's were obtained from the Molecular
796 Signatures Database (MSigDB) (version 7.4) [69] using the R package
797 msigdbr (version 7.4.1). Among the datasets in MSigDB, we used Gene
798 Ontology for Biological Process, Hallmark gene sets, KEGG pathways,
799 Reactome pathways, and WikiPathways. Using the sample batch-corrected
800 Gene Score Matrix, we compared nuclei from a pair of dietary conditions, for
801 example, the hepatocytes from 4 weeks of HFD *vs.* the normal diet. The
802 statistical significance of gene set enrichment was tested using the R package
803 PADOG (version 1.34.0) [70] and by performing permutation for 10,000
804 times or 20 times the number of GS's in a dataset, whichever was larger.
805 Within each dataset, the GS's that attained a false discover rate <0.05 were
806 considered statistically significant. Significant GS's were visually inspected
807 using the EnrichmentMap App (version 3.3.3) of the Cytoscape software
808 (version 3.8.2).
809
810 **Discovering major modules of TF regulation and assigning biological**
811 **processes**
812 To elucidate a global picture of TF regulation, we inferred the regulatory
813 impacts of all TFs on all genes and then extracted major modules of the

814 regulatory relationship between TFs and genes. For each gene, we computed
815 its regulatory TFs using the R package GENIE3 (version 1.14.0) [71], which
816 implements a random forest algorithm and thus can model nonlinear
817 regulation. The algorithm aims to predict the expression level of the gene
818 based on the binding of TFs and dietary conditions. As for the number of
819 candidate regulators randomly selected at each node of the random forests,
820 we chose the square root of the number of predictors, as suggested in the
821 GENIE3 publication. GENIE3 quantified the impacts of regulators by the
822 proportion of explained variance. By combining the results from all genes, we
823 obtained a regulator-regulatee matrix, in which the rows correspond to TFs
824 and dietary conditions, the columns correspond to genes, and an element
825 represents the importance of a TF to a gene.

826 Next, we applied the nonnegative matrix factorization implemented in
827 the R package NMF (version 0.23.0) [72] to the regulator-regulatee matrix.
828 We used the Brunet algorithm [73] with 100 random starts. The number of
829 factors was chosen based on data as the inflection point of residual error [74].
830 In the output of NMF, the membership of TFs (regulators) and genes
831 (regulatees) to a factor was defined by nonnegative weight. We called the
832 factors as "modules" of TF regulation.

833 We discovered biological processes that characterize a module by
834 applying gene set enrichment analysis to genes belonging to the module. For
835 each module of TF regulation, we tested if heavily weighted genes were
836 enriched in some GS. For each GS (see above section), we computed the P-
837 value and *t*-statistics to test if genes in the given GS had a larger weight than
838 other genes. To account for multiple testing by the GS's, we performed 100
839 trials of permuting gene names and computed family-wise error rates. The
840 GS's that attained a family-wise error rate of <0.05 were considered
841 statistically significant. The permutation test does not require the assumption
842 of normality.

843

844 **GS activity score in single nuclei**

845 To infer the activity of a GS in single nuclei, we computed the "average" level
846 of TF binding and gene expression, with regard to the TFs/genes included in
847 the GS. Formally, we defined the weight of TFs/genes as one if it was included
848 in the GS and as zero otherwise. We then performed SVD for the TF binding
849 and gene expression matrix incorporating the weight. In the Motif Matrix (see
850 above), each TF (row) was standardized by subtracting the mean and being
851 divided by the standard deviation and then multiplied by the weight of the TF.
852 Similarly, in the Gene Score Matrix (see above), each gene (row) was
853 standardized by subtracting the mean and being divided by the standard
854 deviation and then multiplied by the weight of the gene. By stacking the two
855 matrices, we obtained a large matrix, in which the rows correspond to TFs
856 and genes and the columns correspond to nuclei. Finally, SVD was applied
857 and the first right-singular vector was defined as the GS activity score in
858 single nuclei.

859 Although the abovementioned activity score can always be computed,
860 it would be meaningless if the TFs/genes in the GS were not co-expressed.
861 Therefore, we conducted a statistical test to exclude such meaningless GS.
862 We checked whether the TFs/genes in the GS were mutually correlated more
863 than by chance. We computed the null distribution of the first singular value
864 by performing 100 trials of the SVD procedure but with the weight of TFs
865 permutated and the weight of genes permutated and verified whether the
866 actual first singular value could attain an empirical P-value of <0.05. The
867 permutation test does not require the assumption of normality.

868 The principal component approach was adapted from the PLAGE
869 program [24] that was developed to compute the pathway activity score of a
870 GS in bulk tissue specimens. Instead, we applied to nuclei. Compared with
871 the number of specimens, the large number of nuclei enabled significance

testing by permutation.

874 **Abbreviations**

875 GS, gene set; IDF, inverse document frequency; NAFL, non-alcoholic fatty
876 liver; NAFLD, non-alcoholic fatty liver disease; NASH, non-alcoholic
877 steatohepatitis; snATAC-seq, single-nucleus ATAC-seq; SVD, singular value
878 decomposition; TF, transcription factor

879

880 **Declarations**

881 **Ethics approval and consent to participate**

882 Not applicable.

883 **Consent for publication**

884 Not applicable.

885 **Availability of data and materials**

886 The single-nucleus and bulk ATAC-seq raw data generated during this study
887 are available at DDBJ (DRA DRA014511 and DRA014458 within BioProject
888 PRJDB13870). Datasets after cleaning are available at figshare (DOI
889 10.6084/m9.figshare.20236509). The code generated during this study is
890 available at GitHub https://github.com/fumi-
891 github/rat_singlecell_liver_ArchR

892 **Competing interests**

893 The authors declare no competing interests.

894 **Funding**

895 This work was supported by the NCGM Intramural Research Fund (20A1013)
896 and by JSPS KAKENHI [grant number JP16K07218]. The funding body had no
897 role in the design of the study and collection, analysis and interpretation of
898 data and in writing the manuscript.

899 **Authors' contributions**

900 Conceptualization, F.T., Y-Q.L., H.S-F. and N.K.;

901 Software, F.T.;

902 Formal Analysis, F.T.;

Investigation, F.T., Y-Q.L., H.S-F., M.I., M.Y.A. and E.K.;

Writing – Original Draft, F.T.;

Writing – Review & Editing, all authors;

Supervision, N.K.

All authors have read and approved the manuscript.

Acknowledgments

The authors thank the Center for Omics and Bioinformatics, Graduate School of Frontier Sciences, University of Tokyo and Azenta for their assistance in snATAC-seq. The authors thank Enago (www.enago.jp) for the English language review.

915 **FIGURE LEGENDS**

916 **Figure 1**

917 Cell types observed in the liver and their composition

918 (A) Experimental design.

919 (B) UMAP plot of the observed nuclei. The number of observed nuclei is shown
920 in parentheses.

921 (C) Expression level of cell type marker genes in nuclei (vertical axis) is
922 compared between nuclei clusters (horizonal axis) in violin plots.

923 (D) Validation of deconvolution based on ATAC-seq using the samples assayed
924 in both single-nucleus and bulk ATAC-seq. The true cell type proportion
925 counted in snATAC-seq (horizontal axis) is compared with the proportion
926 inferred based on bulk ATAC-seq (vertical axis).

927 (E) Proportion of nuclei from a specific cell type (vertical axis) is compared
928 between dietary conditions (horizontal axis). The proportion was measured
929 by bulk tissue ATAC-seq (n=4 samples per condition).

930 **Figure 2**

931 Biological processes that were differentially expressed under dietary
932 interventions

933 The heatmap shows whether a biological process (row) was differentially
934 expressed in a cell type under a dietary intervention compared with the
935 normal diet (column). The coloring indicates statistical significance of gene
936 set enrichment analysis: red if significant adjusting for multiple testing,
937 brown if nominally significant at $P<0.01$, light brown if $P<0.05$, and white
938 otherwise.

939 **Figure 3**

940 Major modules of TF regulation in hepatocytes

941 (A) The central heatmap shows the regulatory impact of TF binding (columns)
942 on genes (rows). Major modules are highlighted by red rectangles. The
943 heatmap at the bottom shows the membership of TFs (columns) to modules

944 (rows). The heatmap on the right shows the membership of genes (rows) to
945 modules (columns). The heatmaps only include the TFs/genes with
946 membership ranking in the top five and membership weight more than half
947 of the top.
948 (B) Gene set enrichment analysis for genes belonging to the modules.
949 Excessive overlap between genes belonging to a module and the GS for a
950 biological process (horizontal axis) was tested, and the P-value was plotted
951 (vertical axis). Bars are colored black if statistically significant, adjusting for
952 multiple testing, and gray otherwise. Biological processes that were
953 significant for at least one module are shown.
954 **Figure 4**
955 Major modules of TF regulation in endothelial cells
956 See legend in Figure 3.
957 **Figure 5**
958 Major modules of TF regulation in stellate cells
959 See legend in Figure 3.
960 **Figure 6**
961 Major modules of TF regulation in macrophages
962 See legend in Figure 3.
963 **Figure 7**
964 Core genes for the TNFα signaling via NF-κB processes are extracted based
965 on correlation with GS activity score and protein-protein interaction
966 (A) Analysis in hepatocytes using the GS for TNFα signaling via NF-κB. In the
967 top panel, the GS activity score in single nuclei (vertical axis) is compared
968 between dietary conditions (horizontal axis) in violin plots. Box plots within
969 the violins indicate quartiles. The score distribution under a diet intervention
970 was compared with that under normal diet condition by the area under the
971 receiver operating characteristic curve: * and ** indicate larger than 0.75
972 and 0.9, respectively. In the remaining panels, TFs and genes are plotted by

973 the correlation with the GS activity score (horizontal axis) and the total
974 amount of protein-protein interactions with the GS (vertical axis). The points
975 for TFs/genes having protein-protein interactions within the top 5% are
976 colored red, and the others are colored blue. The TFs/genes in the GS (middle
977 panel) and not in the GS (bottom panel) are plotted separately; the former
978 have more abundant protein-protein interactions. TFs are labeled in all capital
979 letters. The GS was taken from the MSigDB Hallmark database. Protein-
980 protein interactions were taken from the STRING database (version 11.0).
981 (B) Analysis in endothelial cells.
982 (C) Analysis in macrophages.
983 **Figure 8**
984 Modules of global TF regulation, biological processes, and core genes were
985 detected by the data-driven approach
986 See text for the full list of core genes.
987

988 **REFERENCES**

989

990 1. Friedman SL, Neuschwander-Tetri BA, Rinella M, Sanyal AJ. Mechanisms of
991 NAFLD development and therapeutic strategies. Nat Med. 2018;24:908-22.

992 2. Chu AL, Schilling JD, King KR, Feldstein AE. The Power of Single Cell Analysis for
993 the Study of Liver Pathobiology. Hepatology. 2020;:hep.31485-23.

994 3. Ramachandran P, Dobie R, Wilson-Kanamori JR, Dora EF, Henderson BEP, Luu
995 NT, et al. Resolving the fibrotic niche of human liver cirrhosis at single-cell level.
996 Nature. 2019;575:512-8.

997 4. Xiong X, Kuang H, Ansari S, Liu T, Gong J, Wang S, et al. Landscape of
998 Intercellular Crosstalk in Healthy and NASH Liver Revealed by Single-Cell
999 Secretome Gene Analysis. Molecular cell. 2019;75:644-660.e5.

1000 5. Sturm G, Finotello F, Petitprez F, Zhang JD, Baumbach J, Fridman WH, et al.
1001 Comprehensive evaluation of transcriptome-based cell-type quantification methods
1002 for immuno-oncology. Bioinformatics (Oxford, England). 2019;35:i436-45.

1003 6. Teschendorff AE, Zheng SC. Cell-type deconvolution in epigenome-wide
1004 association studies: a review and recommendations. Epigenomics. 2017;9:757-68.

1005 7. Li H, Sharma A, Luo K, Qin ZS, Sun X, Liu H. DeconPeaker, a Deconvolution Model
1006 to Identify Cell Types Based on Chromatin Accessibility in ATAC-Seq Data of
1007 Mixture Samples. Frontiers in Genetics. 2020;11:392.

1008 8. Kiki I, Altunkaynak BZ, Altunkaynak ME, Vuraler O, Unal D, Kaplan S. Effect of
1009 High Fat Diet on the Volume of Liver and Quantitative Feature of Kupffer Cells in
1010 the Female Rat: A Stereological and Ultrastructural Study. Obes Surg. 2007;17:1381.

1011 9. Zhong L, Huang L, Xue Q, Liu C, Xu K, Shen W, et al. Cell-specific elevation of
1012 Runx2 promotes hepatic infiltration of macrophages by upregulating MCP-1 in high-
1013 fat diet-induced mice NAFLD. J Cell Biochem. 2019;120:11761-74.

1014 10. Zhang F, Jiang WW, Li X, Qiu XY, Wu Z, Chi YJ, et al. Role of intrahepatic B cells
1015 in non-alcoholic fatty liver disease by secreting pro-inflammatory cytokines and
1016 regulating intrahepatic T cells. J Digest Dis. 2016;17:464–74.

1017 11. Barrow F, Khan S, Fredrickson G, Wang H, Dietsche K, Parthiban P, et al.
1018 Microbiota-Driven Activation of Intrahepatic B Cells Aggravates NASH Through
1019 Innate and Adaptive Signaling. Hepatology Baltim Md. 2021;74:704–22.

1020 12. Hess J, Angel P, Schorpp-Kistner M. AP-1 subunits: quarrel and harmony among
1021 siblings. J Cell Sci. 2004;117:5965–73.

1022 13. Zinzalla G. A New Way Forward in Cancer Drug Discovery: Inhibiting the
1023 SWI/SNF Chromatin Remodelling Complex. Chembiochem. 2016;17:677–82.

1024 14. Halpern KB, Shenhav R, Matcovitch-Natan O, Tóth B, Lemze D, Golan M, et al.
1025 Single-cell spatial reconstruction reveals global division of labour in the mammalian
1026 liver. Nature. 2017;542:352–6.

1027 15. Sun J, Weis WI. Biochemical and Structural Characterization of β-Catenin
1028 Interactions with Nonphosphorylated and CK2-Phosphorylated Lef-1. J Mol Biol.
1029 2011;405:519–30.

1030 16. Behari J, Yeh T-H, Krauland L, Otruba W, Cieply B, Hauth B, et al. Liver-Specific
1031 β-Catenin Knockout Mice Exhibit Defective Bile Acid and Cholesterol Homeostasis
1032 and Increased Susceptibility to Diet-Induced Steatohepatitis. Am J Pathology.
1033 2010;176:744–53.

1034 17. Rasala BA, Orjalo AV, Shen Z, Briggs S, Forbes DJ. ELYS is a dual
1035 nucleoporin/kinetochore protein required for nuclear pore assembly and proper cell
1036 division. Proc National Acad Sci. 2006;103:17801–6.

1037 18. Yu X, Chen X, Zheng XD, Zhang J, Zhao X, Liu Y, et al. Growth Differentiation
1038 Factor 11 Promotes Abnormal Proliferation and Angiogenesis of Pulmonary Artery
1039 Endothelial Cells. Hypertension. 2018;71:729–41.

1040 19. Tran MTN, Hamada M, Jeon H, Shiraishi R, Asano K, Hattori M, et al. MafB is a
1041 critical regulator of complement component C1q. Nat Commun. 2017;8:1700.

1042 20. Chistiakov DA, Myasoedova VA, Revin VV, Orekhov AN, Bobryshev YV. The
1043 impact of interferon-regulatory factors to macrophage differentiation and
1044 polarization into M1 and M2. Immunobiology. 2018;223:101-11.

1045 21. Salcedo R, Wasserman K, Young HA, Grimm MC, Howard OMZ, Anver MR, et al.
1046 Vascular Endothelial Growth Factor and Basic Fibroblast Growth Factor Induce
1047 Expression of CXCR4 on Human Endothelial Cells In Vivo Neovascularization
1048 Induced by Stromal-Derived Factor-1α. Am J Pathology. 1999;154:1125-35.

1049 22. Date D, Das R, Narla G, Simon DI, Jain MK, Mahabeleshwar GH. Kruppel-like
1050 Transcription Factor 6 Regulates Inflammatory Macrophage Polarization. J Biol
1051 Chem. 2014;289:10318-29.

1052 23. Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, et al. A
1053 Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell.
1054 2018;174:1309-1324.e18.

1055 24. Tomfohr J, Lu J, Kepler TB. Pathway level analysis of gene expression using
1056 singular value decomposition. Bmc Bioinformatics. 2005;6:225-225.

1057 25. Kakino S, Ohki T, Nakayama H, Yuan X, Otabe S, Hashinaga T, et al. Pivotal Role
1058 of TNF-a in the Development and Progression of Nonalcoholic Fatty Liver Disease
1059 in a Murine Model. Horm Metab Res. 2017;50:80-7.

1060 26. Locatelli I, Sutti S, Vacchiano M, Bozzola C, Albano E. NF-κB1 deficiency
1061 stimulates the progression of non-alcoholic steatohepatitis (NASH) in mice by
1062 promoting NKT-cell-mediated responses. Clin Sci. 2012;124:279-87.

1063 27. Roos B de, Rungapamestry V, Ross K, Rucklidge G, Reid M, Duncan G, et al.
1064 Attenuation of inflammation and cellular stress-related pathways maintains insulin
1065 sensitivity in obese type I interleukin-1 receptor knockout mice on a high-fat diet.
1066 Proteomics. 2009;9:3244-56.

1067 28. Sawitza I, Kordes C, Reister S, Häussinger D. The niche of stellate cells within
1068 rat liver. Hepatology. 2009;50:1617-24.

1069 29. Boujedidi H, Robert O, Bignon A, Cassard-Doulcier A-M, Renoud M-L, Gary-
1070 Gouy H, et al. CXCR4 dysfunction in non-alcoholic steatohepatitis in mice and
1071 patients. Clin Sci. 2014;128:257-67.

1072 30. Potoupni V, Georgiadou M, Chatzigriva E, Polychronidou G, Markou E, Gakis CZ,
1073 et al. Circulating tumor necrosis factor-α levels in non-alcoholic fatty liver disease:
1074 A systematic review and a meta-analysis. J Gastroen Hepatol. 2021;36:3002-14.

1075 31. Seo YY, Cho YK, Bae J-C, Seo MH, Park SE, Rhee E-J, et al. Tumor Necrosis
1076 Factor-α as a Predictor for the Development of Nonalcoholic Fatty Liver Disease:
1077 A 4-Year Follow-Up Study. Endocrinol Metab. 2013;28:41-5.

1078 32. Goel R, Boylan B, Gruman L, Newman PJ, North PE, Newman DK. The
1079 proinflammatory phenotype of PECAM-1-deficient mice results in atherogenic diet-
1080 induced steatohepatitis. Am J Physiol-gastr L. 2007;293:G1205-14.

1081 33. Zhang C, Jia Y, Liu B, Wang G, Zhang Y. TLR4 knockout upregulates the
1082 expression of Mfn2 and PGC-1α in a high-fat diet and ischemia-reperfusion mice
1083 model of liver injury. Life Sci. 2020;254:117762.

1084 34. Cepero-Donates Y, Lacraz G, Ghobadi F, Rakotoarivelo V, Orkhis S, Mayhue M,
1085 et al. Interleukin-15-mediated inflammation promotes non-alcoholic fatty liver
1086 disease. Cytokine. 2016;82:102-11.

1087 35. Pérez-Martínez L, Pérez-Matute P, Aguilera-Lizarraga J, Rubio-Mediavilla S,
1088 Narro J, Recio E, et al. Maraviroc, a CCR5 antagonist, ameliorates the development
1089 of hepatic steatosis in a mouse model of non-alcoholic fatty liver disease (NAFLD).
1090 J Antimicrob Chemoth. 2014;69:1903-10.

1091 36. Patouraux S, Rousseau D, Bonnafous S, Lebeaupin C, Luci C, Canivet CM, et al.
1092 CD44 is a key player in non-alcoholic steatohepatitis. J Hepatol. 2017;67:328-38.

1093 37. Li ZZ, Berk M, McIntyre TM, Feldstein AE. Hepatic Lipid Partitioning and Liver
1094 Damage in Nonalcoholic Fatty Liver Disease ROLE OF STEAROYL-CoA
1095 DESATURASE. J Biol Chem. 2009;284:5637-44.

1096 38. Zhang Y, Lu Z, Zeng W, Zhao J, Zhou X. Two sides of NNMT in alcoholic and
1097 nonalcoholic fatty liver development. J Hepatol. 2020;74:1250-3.

1098 39. Chen C, Li H, Song J, Zhang C, Li M, Mao Y, et al. Role of Apolipoprotein A1 in
1099 PPAR Signaling Pathway for Nonalcoholic Fatty Liver Disease. Ppar Res.
1100 2022;2022:4709300.

1101 40. Yang MH, Sung J, Gwak G-Y. The associations between apolipoprotein B, A1,
1102 and the B/A1 ratio and nonalcoholic fatty liver disease in both normal-weight and
1103 overweight Korean population. J Clin Lipidol. 2016;10:289-98.

1104 41. Rigotti FDA, Wiedmann L, Hubert MO, Vacca M, Hasan SS, Moll I, et al.
1105 Semaphorin 3C exacerbates liver fibrosis. Biorxiv. 2022;:2021.07.29.454292.

1106 42. Cao S, Yaqoob U, Das A, Shergill U, Jagavelu K, Huebert RC, et al. Neuropilin-1
1107 promotes cirrhosis of the rodent and human liver by enhancing PDGF/TGF-β
1108 signaling in hepatic stellate cells. J Clin Invest. 2010;120:2379-94.

1109 43. Yagai T, Miyajima A, Tanaka M. Semaphorin 3E Secreted by Damaged
1110 Hepatocytes Regulates the Sinusoidal Regeneration and Liver Fibrosis during Liver
1111 Regeneration. Am J Pathology. 2014;184:2250-9.

1112 44. Binder JX, Pletscher-Frankild S, Tsafou K, Stolte C, O'Donoghue SI, Schneider
1113 R, et al. COMPARTMENTS: unification and visualization of protein subcellular
1114 localization evidence. Database J Biological Databases Curation. 2014;2014:bau012.

1115 45. Ajmera V, Perito ER, Bass NM, Terrault NA, Yates KP, Gill R, et al. Novel plasma
1116 biomarkers associated with liver disease severity in adults with nonalcoholic fatty
1117 liver disease. Hepatology. 2017;65:65-77.

1118 46. Koutsogiannis D, Summers K, George B, Adams P, Marotta P, Chakrabarti S.
1119 Identification of Serum Biomarkers in End Stage Liver Disease. Open Biomarkers J.
1120 2010;3:1-6.

1121 47. Krawczyk M, Zimmermann S, Hess G, Holz R, Dauer M, Raedle J, et al. Panel of
1122 three novel serum markers predicts liver stiffness and fibrosis stages in patients
1123 with chronic liver disease. Plos One. 2017;12:e0173506.

1124 48. An X, Liu J, Zheng X, Dou Z, Li Y, Suo Y, et al. Serum HGF ,PCIII and PLT are
1125 Noninvasive Markers for the diagnosis of nonalcoholic fatty liver disease. 2019.
1126 https://doi.org/10.21203/rs.2.18035/v1.

1127 49. Page S, Birerdinc A, Estep M, Stepanova M, Afendy A, Petricoin E, et al.
1128 Knowledge-Based Identification of Soluble Biomarkers: Hepatic Fibrosis in NAFLD
1129 as an Example. Plos One. 2013;8:e56009.

1130 50. Herck MAV, Vonghia L, Francque SM. Animal Models of Nonalcoholic Fatty Liver
1131 Disease—A Starter's Guide. Nutrients. 2017;9:1072.

1132 51. Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA,
1133 Vesuna S, et al. An improved ATAC-seq protocol reduces background and enables
1134 interrogation of frozen tissues. Nature Methods. 2017;14:959-62.

1135 52. Lareau CA, Ma S, Duarte FM, Buenrostro JD. Inference and effects of barcode
1136 multiplets in droplet-based single-cell assays. Nature Communications. 2020;11:866.

1137 53. Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, et al.
1138 ArchR is a scalable software package for integrative single-cell chromatin
1139 accessibility analysis. Nature Genetics. 2021;53:403 - 411.

1140 54. Orchard P, Kyono Y, Hensley J, Kitzman JO, Parker SCJ. Quantification,
1141 Dynamic Visualization, and Validation of Bias in ATAC-Seq Data with ataqv. Cell
1142 Systems. 2020;10:298-306.e4.

1143 55. Katsuda T, Hosaka K, Matsuzaki J, Usuba W, Prieto-Vila M, Yamaguchi T, et al.
1144 Transcriptomic Dissection of Hepatocyte Heterogeneity: Linking Ploidy, Zonation,
1145 and Stem/Progenitor Cell Characteristics. Cellular and Molecular Gastroenterology
1146 and Hepatology. 2020;9:161-83.

1147 56. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-
1148 based Analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137-R137.

1149 57. Berry MW, Browne M. Understanding Search Engines. 2nd edition. SIAM; 2005.

1150 58. Baglama J, Reichel L. Augmented Implicitly Restarted Lanczos Bidiagonalization
1151 Methods. Siam J Sci Comput. 2005;27:19-42.

1152 59. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al.
1153 Integrated analysis of multimodal single-cell data. Cell. 2021;184:3573-3587.e29.

1154 60. Traag VA, Waltman L, Eck NJ van. From Louvain to Leiden: guaranteeing well-
1155 connected communities. Sci Rep-uk. 2019;9:5233.

1156 61. Reske JJ, Wilson MR, Chandler RL. ATAC-seq normalization method can
1157 significantly affect differential accessibility analysis and interpretation. Epigenetics &
1158 chromatin. 2020;13:22.

1159 62. Jew B, Alvarez M, Rahmani E, Miao Z, Ko A, Garske KM, et al. Accurate
1160 estimation of cell composition in bulk expression through robust integration of
1161 single-cell information. Nature Communications. 2020;11:1971.

1162 63. Teschendorff AE, Breeze CE, Zheng SC, Beck S. A comparison of reference-
1163 based algorithms for correcting cell-type heterogeneity in Epigenome-Wide
1164 Association Studies. BMC Bioinformatics. 2017;18:105.

1165 64. Wilson DR, Jin C, Ibrahim JG, Sun W. ICeD-T Provides Accurate Estimates of
1166 Immune Cell Abundance in Tumor Samples by Allowing for Aberrant Gene
1167 Expression Patterns. J Am Stat Assoc. 2019;115:1-23.

1168 65. Yip SH, Wang P, Kocher J-PA, Sham PC, Wang J. Linnorm: improved statistical
1169 analysis for single cell RNA-seq expression data. Nucleic Acids Research.
1170 2017;45:e179.

1171 66. Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, et al. SAVER: gene
1172 expression recovery for single-cell RNA sequencing. Nature Methods. 2018;15:539 -
1173 542.

1174 67. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. chromVAR: inferring
1175 transcription-factor-associated accessibility from single-cell epigenomic data. Nat
1176 Methods. 2017;14:975-8.

1177 68. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al.
1178 Determination and Inference of Eukaryotic Transcription Factor Sequence
1179 Specificity. Cell. 2014;158:1431-43.

1180 69. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al.
1181 Gene set enrichment analysis: A knowledge-based approach for interpreting

1182 genome-wide expression profiles. Proceedings of the National Academy of Sciences
1183 of the United States of America. 2005;102:15545-50.

1184 70. Tarca AL, Draghici S, Bhatti G, Romero R. Down-weighting overlapping genes
1185 improves gene set analysis. Bmc Bioinformatics. 2012;13:136.

1186 71. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring Regulatory Networks
1187 from Expression Data Using Tree-Based Methods. Plos One. 2010;5:e12776.

1188 72. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization.
1189 BMC Bioinformatics. 2010;11:367.

1190 73. Brunet J-P, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern
1191 discovery using matrix factorization. Proc National Acad Sci. 2004;101:4164-9.

1192 74. Hutchins LN, Murphy SM, Singh P, Graber JH. Position-dependent motif
1193 characterization using non-negative matrix factorization. Bioinformatics.
1194 2008;24:2684-90.

1195

Figure 1

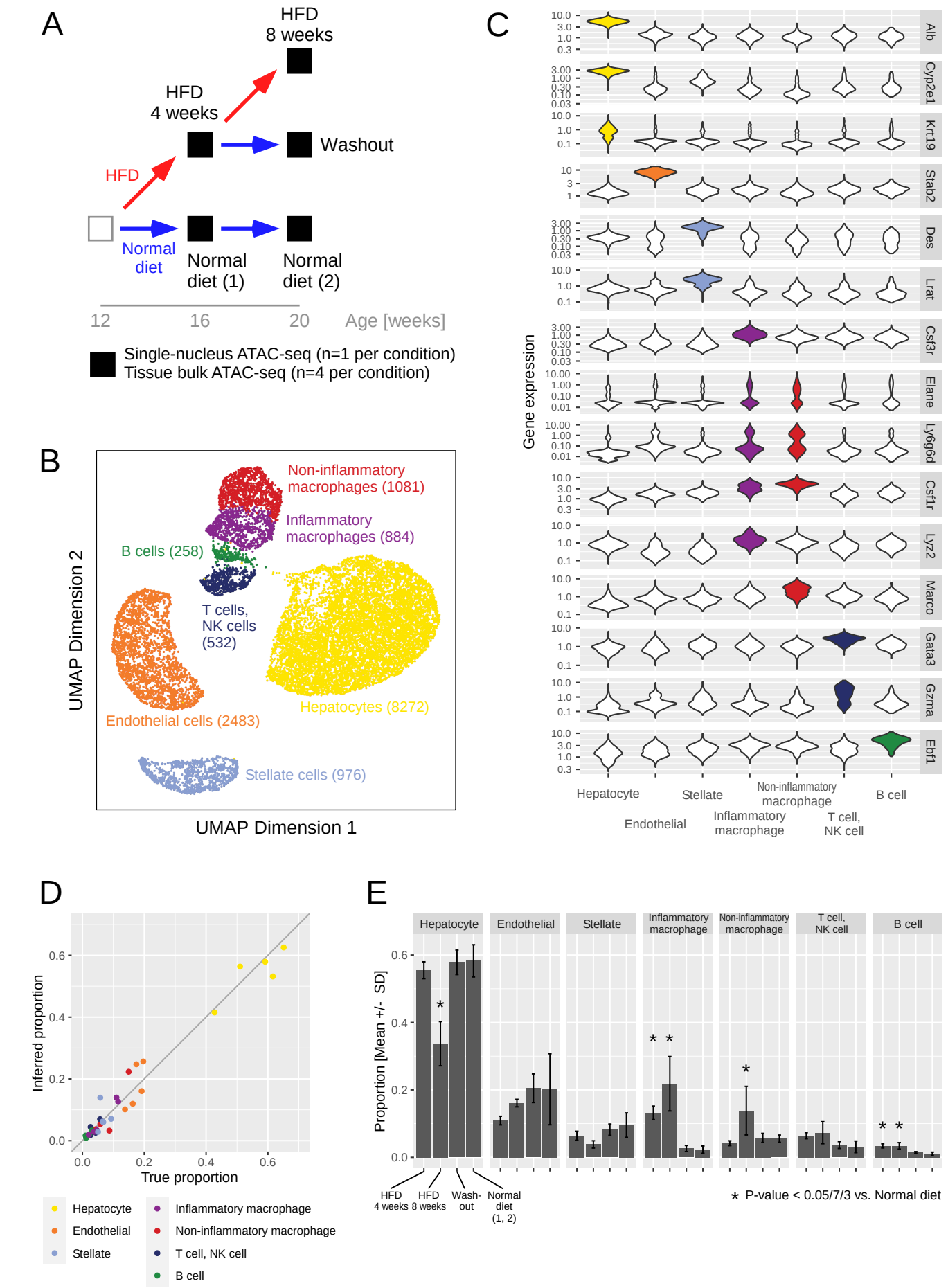
A
HFD
8 weeks
HFD
4 weeks
Washout
HFD
Normal
diet
Normal
diet (1)
Normal
diet (2)
12
16
20
Age [weeks]
Single-nucleus ATAC-seq (n=1 per condition)
Tissue bulk ATAC-seq (n=4 per condition)
B
Non-inflammatory
macrophages (1081)
Inflammatory
macrophages (884)
B cells (258)
T cells,
NK cells
(532)
Hepatocytes (8272)
Endothelial cells (2483)
Stellate cells (976)
UMAP Dimension 2
UMAP Dimension 1
C
Gene expression
Alb
Cyp2e1
Krt19
Stab2
Des
Lrat
Csf3r
Elane
Ly6g6d
Csf1r
Lyz2
Marco
Gata3
Gzma
Ebf1
Hepatocyte
Stellate
Non-inflammatory
macrophage
B cell
Endothelial
Inflammatory
macrophage
T cell,
NK cell
D
Inferred proportion
True proportion
Hepatocyte
Endothelial
Stellate
Inflammatory macrophage
Non-inflammatory macrophage
T cell, NK cell
B cell
E
Hepatocyte
Endothelial
Stellate
Inflammatory
macrophage
Non-inflammatory
macrophage
T cell,
NK cell
B cell
Proportion [Mean +/- SD]
HFD
4 weeks
HFD
8 weeks
Wash-
out
Normal
diet
(1, 2)
* P-value < 0.05/7/3 vs. Normal diet

Figure 2

3	FDR <0.05
2	P <0.01
1	P <0.05

Compared to normal diet.

Database	Gene set	Hepatocyte			Endothelial cell			Stellate cell			Macrophage		
		HFD 4w	HFD 8w	wash out	HFD 4w	HFD 8w	wash out	HFD 4w	HFD 8w	wash out	HFD 4w	HFD 8w	wash out
GOBP	STEROID_METABOLIC_PROCESS	3	3	3	2	2	2	2	0	2	0	2	0
REACTOME	CHOLESTEROL_BIOSYNTHESIS	2	0	3	2	0	1	3	0	0	0	1	0
REACTOME	REGULATION_OF_CHOLESTEROL_BIOSYNTHESIS_BY_SREBP_SREBF	2	0	3	0	0	0	1	0	0	0	0	0
GOBP	FATTY_ACID_METABOLIC_PROCESS	3	2	3	0	1	2	2	0	0	2	1	0
GOBP	LIPID_STORAGE	3	3	3	0	0	1	0	0	0	1	2	0
KEGG	PPAR_SIGNALING_PATHWAY	1	1	3	0	1	0	0	0	0	0	0	0
H	XENOBIOTIC_METABOLISM	2	0	3	0	1	0	0	0	0	0	0	0
H	TNFA_SIGNALING_VIA_NFKB	1	3	0	2	2	0	0	0	0	2	1	0
GOBP	CYTOKINE_PRODUCTION	1	3	0	2	1	0	0	0	0	2	0	3
GOBP	RESPONSE_TO_CYTOKINE	2	3	0	2	0	2	1	1	0	3	0	3
REACTOME	INTERLEUKIN_1_SIGNALING	2	3	0	1	0	0	0	0	0	0	0	0
REACTOME	INTERLEUKIN_10_SIGNALING	1	3	0	2	2	1	3	1	0	0	0	0
REACTOME	INFLAMMASOMES	1	3	0	0	0	0	0	0	0	0	0	0
GOBP	ADAPTIVE_IMMUNE_RESPONSE	0	2	0	1	0	1	0	0	0	3	2	3
KEGG	APOPTOSIS	2	3	0	0	0	0	0	0	0	0	0	0
GOBP	ACTIN_FILAMENT_BASED_PROCESS	1	2	0	0	0	0	0	3	0	1	0	2
GOBP	CELL_MIGRATION	3	3	1	2	1	2	0	3	2	2	2	3
GOBP	INSULIN_LIKE_GROWTH_FACTOR_RECEPTOR_SIGNALING_PATHWAY	1	0	3	0	0	0	0	2	0	0	0	0

Figure 3. Hepatocyte

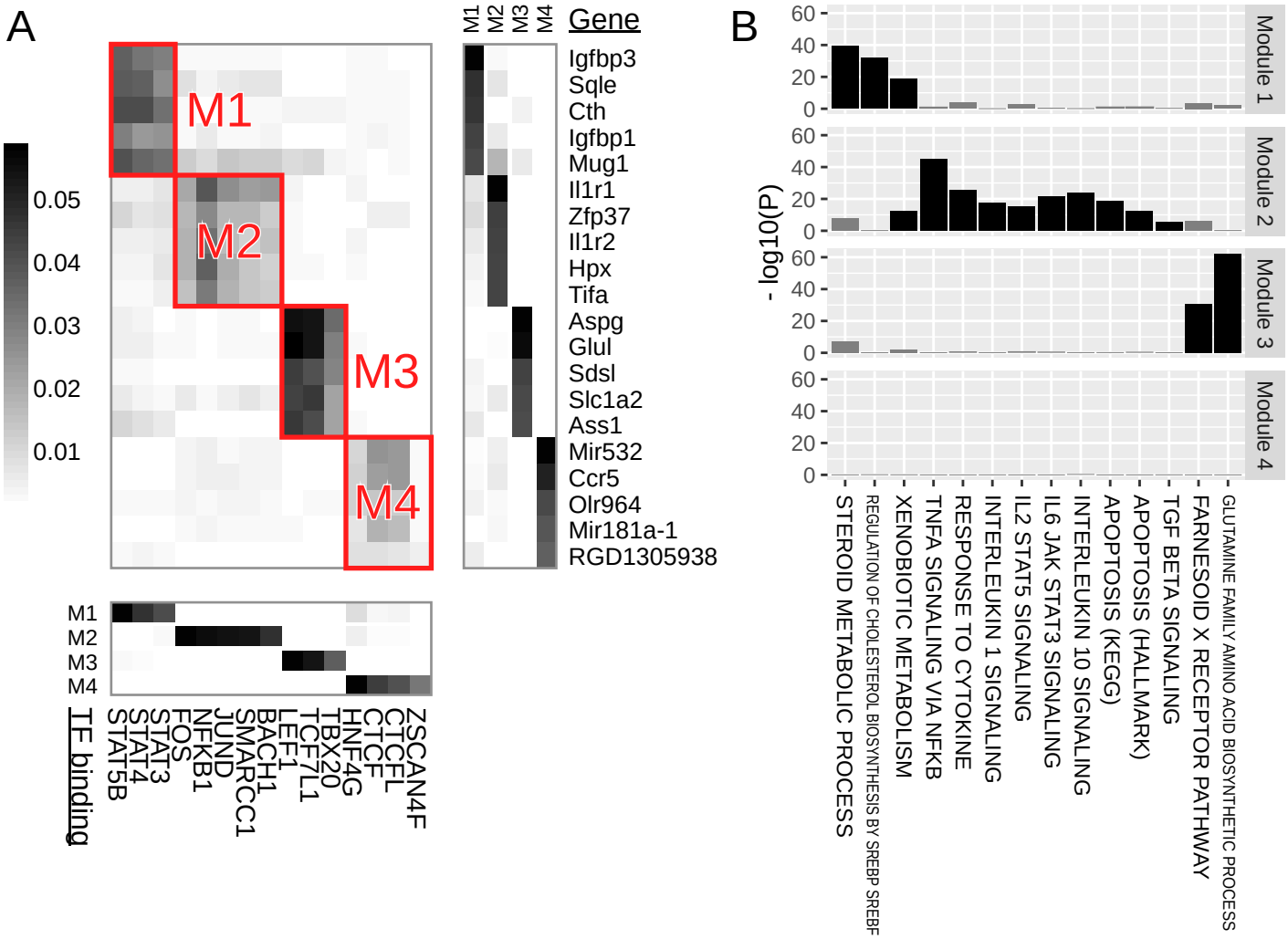

Figure 4. Endothelial

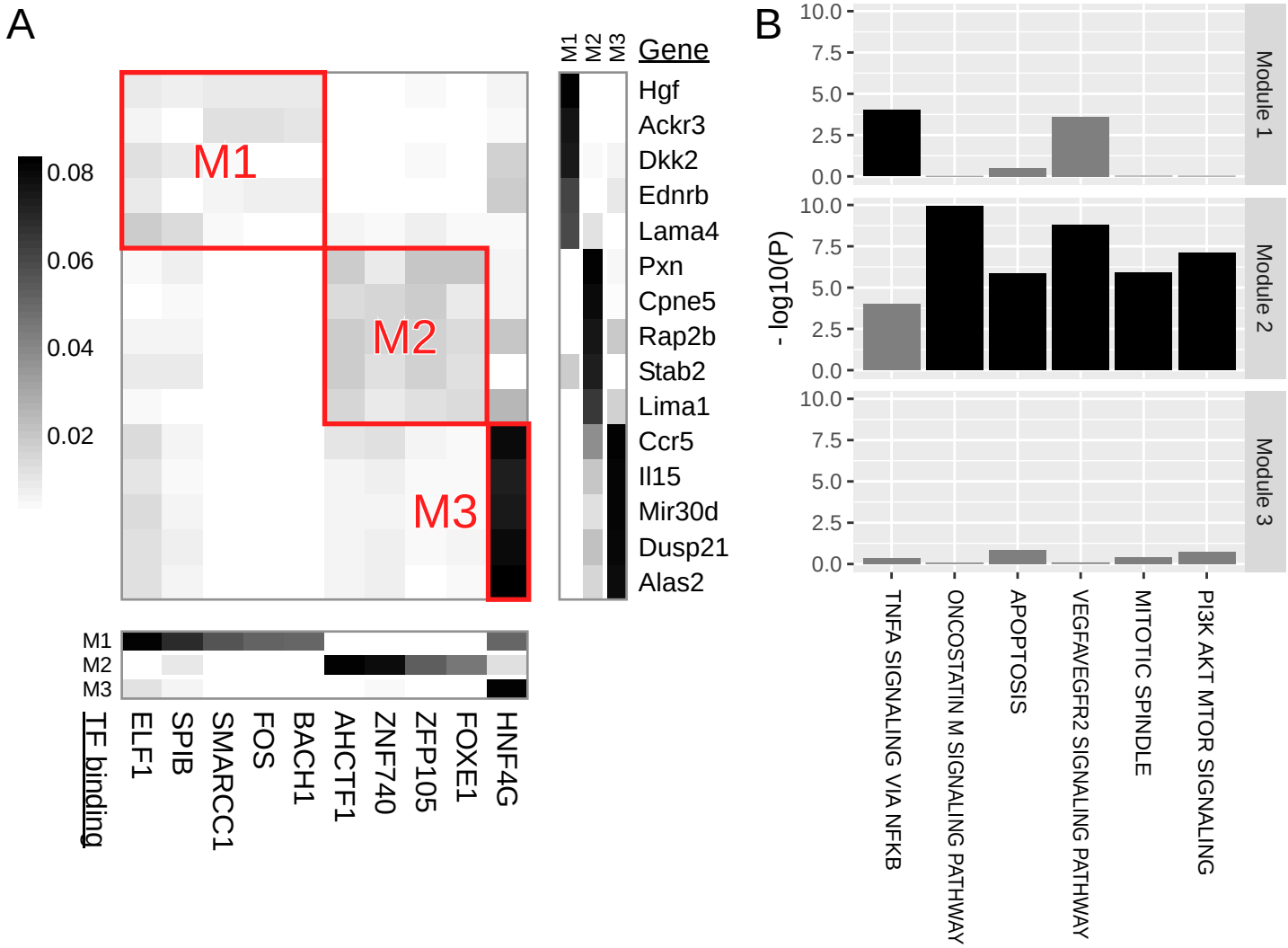

Figure 5. Stellate

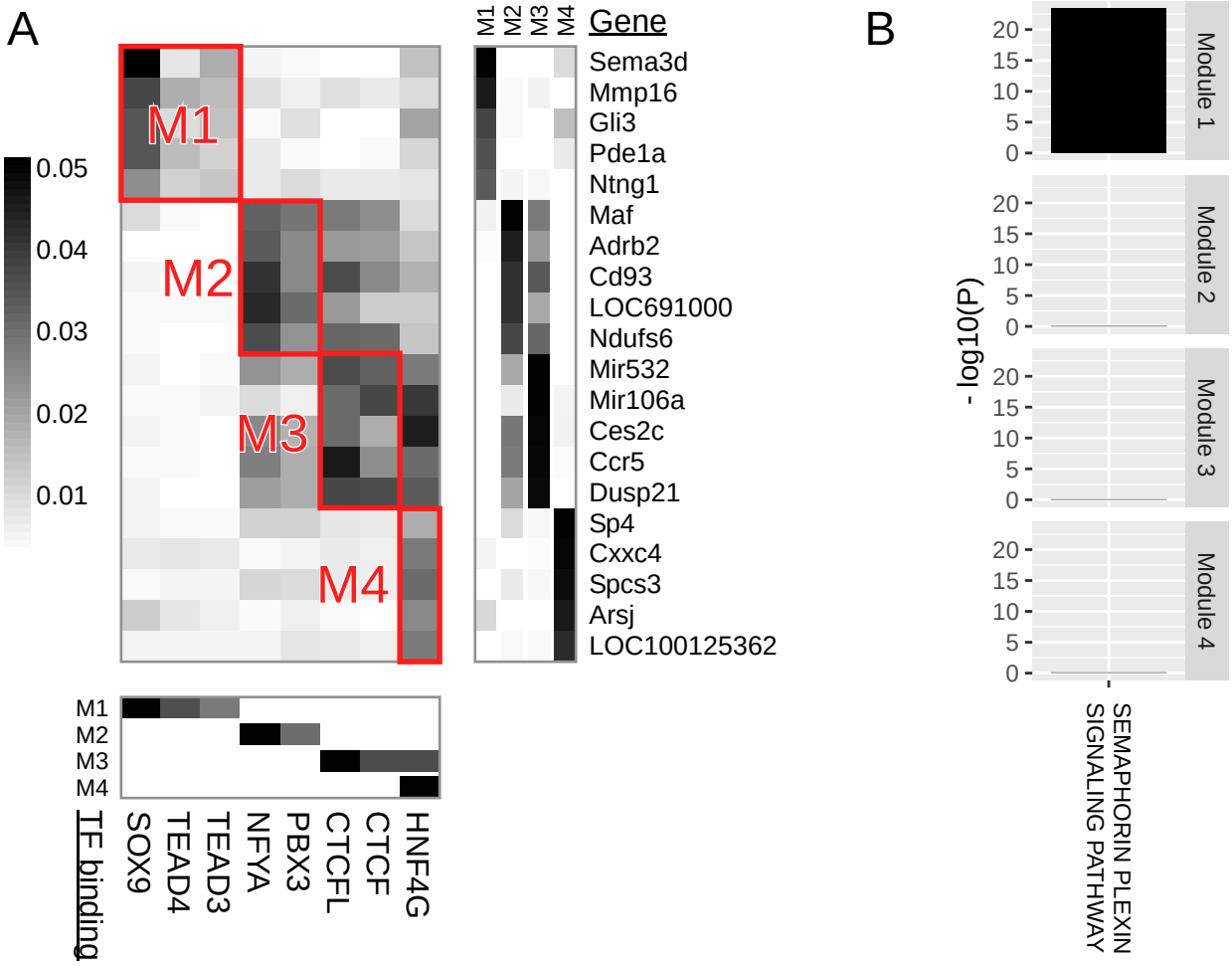

Figure 6. Macrophage

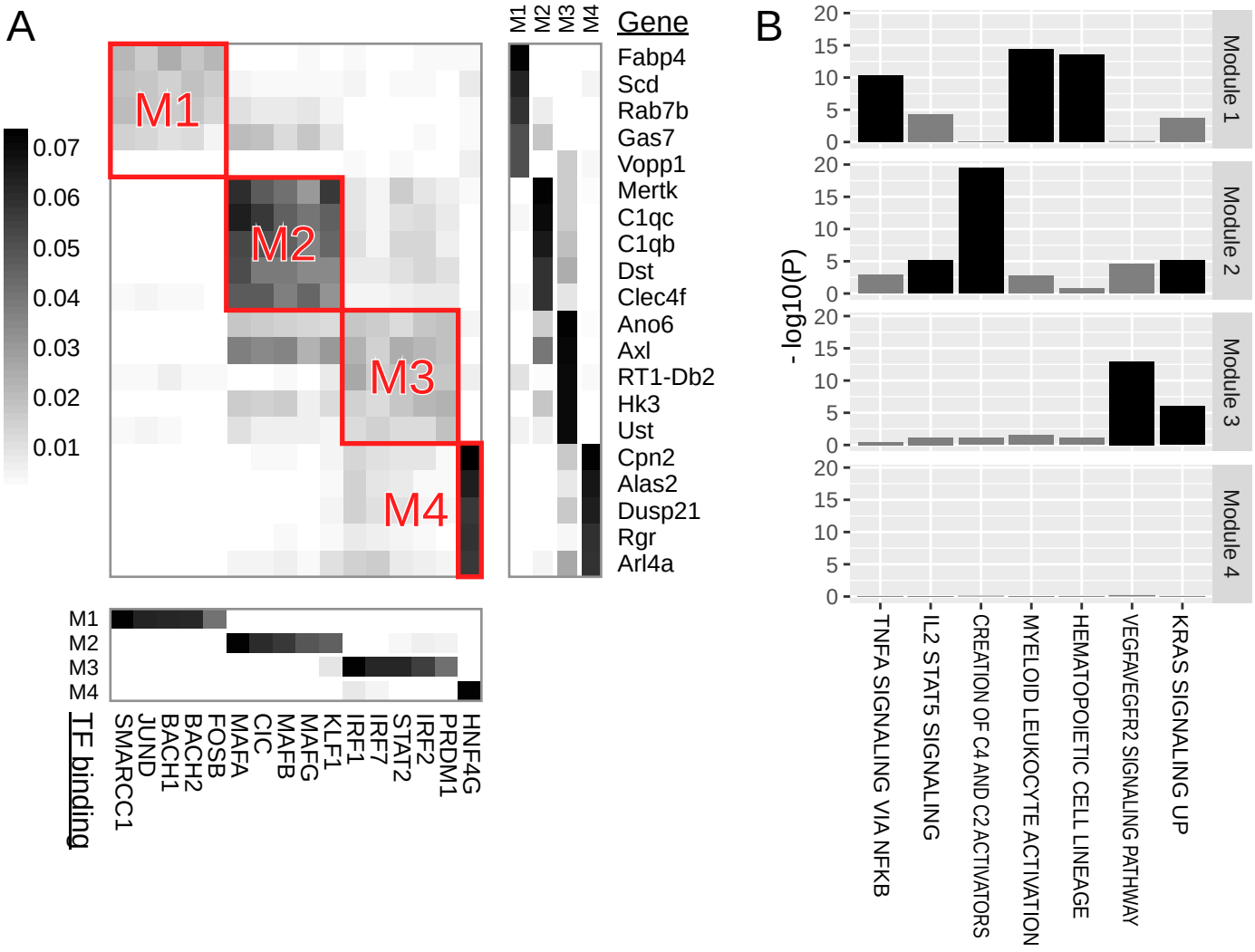

Figure 7

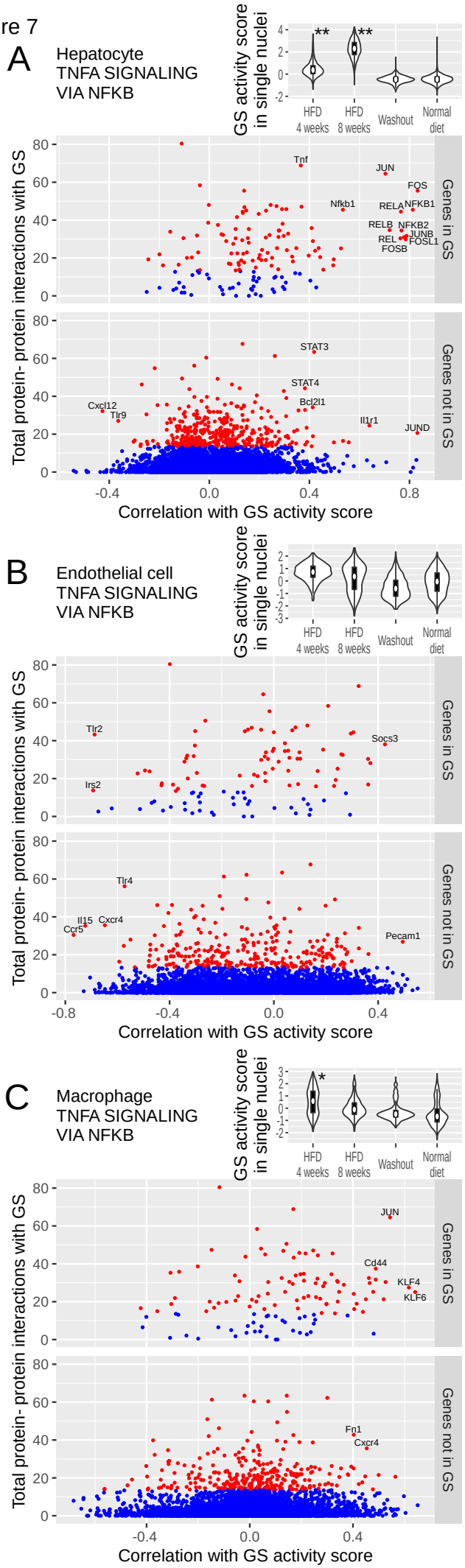
A
Hepatocyte
TNFA SIGNALING
VIA NFKB
GS activity score in single nuclei
HFD 4 weeks
HFD 8 weeks
Washout
Normal diet
Total protein- protein interactions with GS
Genes in GS
Tnf
JUN
FOS
Nfkb1
RELA
NFKB1
RELB
NFKB2
REL
JUNB
FOSL1
FOSB
Genes not in GS
STAT3
STAT4
Cxcl12
Tlr9
Bcl2l1
Il1r1
JUND
Correlation with GS activity score
B
Endothelial cell
TNFA SIGNALING
VIA NFKB
Tlr2
Irs2
Socs3
Tlr4
Il15
Cxcr4
Ccr5
Pecam1
C
Macrophage
TNFA SIGNALING
VIA NFKB
JUN
Cd44
KLF4
KLF6
Fn1
Cxcr4

Figure 8

Global TF regulation in	Machine learning → Modules (TFs therein)	Gene set enrichment analysis → Biological processes	Co-expression, protein-protein interaction → Core genes
Hepatocyte	Module 1 (STAT family)	Steroid metabolism	*Abcg1*, *Sqle*
	Module 2 (AP-1 family)	TNFα signaling via NF-κB	FOS/JUN, *Tnf*
		Apoptosis	FOS/JUN, *Bcl2l1*
	Module 3 (TCF/LEF family)	Lobular zonation	*Glul*, *Cyp7a1*
	Module 4	Undetected	
Endothelial cell	Module 1 (AP-1 family)	TNFα signaling via NF-κB	*Socs3*, *Tlr2*
	Module 2 (AHCTF1, ZNF740)	Angiogenesis	*Pxn*
	Module 3	Undetected	
Stellate cell	Module 1 (SOX9)	Semaphorin-plexin signaling	*Nrp1*, *Nrp2*
	Module 2	Undetected	
	Module 3	Undetected	
	Module 4	Undetected	
Macrophage	Module 1 (AP-1 family)	TNFα signaling via NF-κB	JUN, *Cd44*
	Module 2 (Maf family)	Complement system	*C1qc*, *Fcnb*
	Module 3 (IRF family)	Angiogenesis	*Pak1*, *Kdr*
	Module 4	Undetected	