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 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.

Results: By machine learning, we divided global gene expression into modules, such that transcription factors in a module regulate a set of 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

Abbreviations

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

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].

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 to 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.

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 cell type composition of tissue samples assayed in bulk, has been used for transcriptome and DNA methylation assays [3,4]. For ATAC-seq, Li and associates [5] 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

79 ATAC-seq on the liver of rats fed or withdrawn from a high-fat diet (HFD). We
80 first analyzed the composition of cell types in the liver and the gene
81 expression separately by cell type, followed by experimental validation of
82 ATAC-seq-based deconvolution. Then, to explore global TF regulation in each
83 cell type, we developed a new statistical method, identifying major modules
84 with each characterized by a set of biological processes. We also developed a
85 statistical method to detect core genes for the individual biological processes.
86 Taking snATAC-seq data of an *in vivo* experimental model as the input, our
87 data-driven workflow proceeds top-down from global TF regulation to gene,
88 i.e., step-wise search of major TF regulation modules, biological processes
89 and core genes. The core genes thus identified overlapped well with the
90 previously known NAFLD genes, providing experimental validation to our
91 newly developed statistical methods.
92

93 **Results**

94 **Cell type composition of the liver**

95 Spontaneously Hypertensive Rats (SHR) fed HFD for 4 weeks developed fatty
96 liver, and those fed HFD for 8 weeks further developed inflammation and
97 premature fibrosis (steatohepatitis) (Figure 1A; Table S1). Under the washout
98 condition, where 4 weeks of HFD is followed by 4 weeks of a normal diet,
99 fatty liver was partially ameliorated. Along with rats only fed a normal diet,
100 five dietary conditions were examined.

101 In order to catalog cell types in the liver, we performed snATAC-seq
102 on one animal per dietary condition. We obtained a total of 14,486 nuclei
103 (~3000 per sample), which were grouped into 16 clusters by the similarity of
104 chromatin opening. Hepatocytes (seven clusters), endothelial cells (three
105 clusters), stellate cells (two clusters) and white blood cells (four clusters)
106 separated clearly (Figures 1B and 1C). The clusters for white blood cells
107 corresponded to inflammatory macrophages, non-inflammatory

108 macrophages, T cells and NK cells combined, and B cells. Cholangiocytes were
109 included in the hepatocyte clusters, and neutrophils and Kupffer cells were
110 included in the macrophage clusters.

111 To measure the changes in cell type population through dietary
112 intervention, we performed bulk ATAC-seq on four animals per condition.
113 Using snATAC-seq as reference, we inferred the cell type composition in the
114 bulk ATAC-seq samples by deconvolution. Beforehand, using the five samples
115 assayed in both single-nucleus and bulk ATAC-seq, the accuracy of
116 deconvolution for ATAC-seq was validated (Figure 1D). The difference of the
117 inferred proportion from the true proportion counted by snATAC-seq had
118 average of zero and standard deviation of 0.035, indicating high accuracy.
119 Compared with the normal diet condition, hepatocytes significantly decreased
120 after 8 weeks of HFD (Figure 1E). Inflammatory macrophages largely
121 increased after 4 weeks of HFD and further increased after 8 weeks of HFD.
122 Non-inflammatory macrophages also increased after 8 weeks of HFD. In
123 addition, B cells increased under HFD conditions. In accordance, previous
124 studies of HFD-induced NAFLD animal models reported an increase in liver-
125 resident macrophages [6], recruitment of monocyte-derived macrophages
126 [7], and an increase in intrahepatic B cells [8,9]. Under the washout condition,
127 the cell type composition returned to the same as the normal diet condition.
128 Here, we elucidated how the quantity of cell types changed through dietary
129 intervention, and we next investigate the qualitative change.
130

131 **Differential gene expression under HFD in each cell type**

132 We analyzed cell type-specific differential gene expression by comparing the
133 nuclei of the same cell type between dietary conditions (Table S2; Dataset
134 S1). Hereafter, clusters for inflammatory and non-inflammatory macrophages
135 are combined and regarded as the macrophage cell type. T cells, NK cells,
136 and B cells were not analyzed because the number of nuclei was limited. The

137 biological processes characteristic of the differentially expressed genes were
138 cell type-dependent (Figure 2). Steroid metabolism and fatty acid metabolism
139 genes were expressed in hepatocytes differently under HFD and washout
140 compared with the normal diet. As of inflammation genes, the expression
141 altered in macrophages after 4 weeks of HFD and washout. The inflammation
142 genes and apoptosis genes were expressed in hepatocytes differently after 8
143 weeks of HFD. Genes for the actin filament-based process were expressed in
144 stellate cells differently after 8 weeks of HFD.

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

158

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

160 To elucidate the global picture of TF regulation, we inferred the regulatory
161 impact of all TFs on all genes and extracted major modules of the regulatory
162 relation. By snATAC-seq, we measured in all nuclei of each cell type the
163 genome-wide binding of TFs and the expression of genes, and inferred the
164 regulation of the latter by the former by applying machine learning. The
165 regulatory relation was divided into modules. A module comprises a subset

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

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

183 Module 2 of hepatocytes, Module 1 of endothelial cells, and Module 1
184 of macrophages were characterized by the binding of AP-1 family TFs (namely,
185 FOS, JUND, BACH1 and BACH2), SMARCC1 and NFKB1, and by genes for
186 TNFα signaling via NF-κB. The AP-1 TFs respond to cytokine stimuli [10].
187 SMARCC1 is a member of the SWI/SNF family of proteins, which function as
188 a helicase and interacts with NF-κB [11].

189 Module 3 of hepatocytes was characterized by the binding of TCF/LEF
190 family TFs (namely, LEF1 and TCF7L1) and by genes in the farnesoid X
191 receptor pathway for bile acid synthesis and genes for the glutamine family
192 amino acid biosynthesis. Within the liver lobule, a unit that assembles to form
193 the liver, a central vein is located in the center, the portal triad (portal vein,
194 hepatic artery, and bile duct) is located in the periphery, and the metabolism

195 of hepatocyte in between differs between the pericentral and periportal zones.
196 The glutamine family amino acid biosynthesis genes include *Glul* and *Ass1*,
197 which are highly expressed in the pericentral and periportal zones,
198 respectively [12]. In addition, the LEF1 TF binds to β-catenin protein and
199 activates Wnt signaling pathway [13], and liver-specific β-catenin KO mice
200 exhibited defective cholesterol and bile acid metabolism in the liver [14].
201 Moreover, Wnt-activated genes are highly expressed in the pericentral zone.
202 Thus, LEF1 TF presumably regulates bile acid metabolism and zonated gene
203 expression in the liver.
204 Module 2 of endothelial cells was characterized by the DNA binding of
205 AHCTF1 (aka ELYS) and ZNF740 and by genes for angiogenesis. AHCTF1 is a
206 DNA-binding protein that is required in mitosis [15]. ZNF740 activates
207 angiogenesis in pulmonary artery endothelial cells of rats [16].
208 Module 1 of stellate cells was characterized by SOX9 TF binding and
209 genes in the semaphorin-plexin signaling pathway.
210 Module 2 of macrophages was characterized by the binding of Maf
211 family TFs (namely, MAFA, MAFB and MAFG) and by genes for complement
212 system activation. In *Mafb*-deficient macrophages of mice, C1q production
213 decreased [17].
214 Module 3 of macrophages was characterized by the binding of IRF
215 family TFs (namely, IRF1, IRF2 and IRF7), which regulate immune responses,
216 and by genes for angiogenesis. IRF1 contributes to the commitment of pro-
217 inflammatory M1 macrophages, which produce angiogenic stimulators [18].
218
219 **Core genes in specific biological processes**
220 In the previous section we discovered major modules of TF regulation and
221 then searched across databases, such as Gene Ontology, for biological
222 processes that characterize the modules. We equated a biological process
223 with its comprising genes, and defined them as a gene set (GS). A GS

224 characterizes a module if it significantly overlaps with the genes of the module.
225 Here, we aim to find "core genes" of a biological process. In preparation, we
226 computed the activity of a GS in single nuclei as the "GS activity score" by
227 taking the average binding/expression level of TFs/genes in the set or, more
228 precisely, the first principal component. In hepatocytes, regarding the
229 biological process of TNFα signaling via NF-κB, its GS activity score was the
230 highest among nuclei from 8 weeks of HFD (Figure 7A, top panel).

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

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

249 In endothelial cells, the GS activity score for TNFα signaling via NF-
250 κB (Figure 7B) was higher under HFD conditions. The positively correlated
251 core genes were *Socs3* and *Pecam1*. The negatively correlated core genes
252 were *Tlr2*, *Irs2*, *Tlr4*, *Il15*, *Ccr5* and *Cxcr4* (receptor for the chemoattractant

253 CXCL12 [19]).

254 In macrophages, the GS activity score for TNFα signaling via NF-κB
255 (Figure 7C) was higher after 4 weeks of HFD. TF binding that was positively
256 correlated and had abundant protein-protein interactions included JUN and
257 KLF6. The positively correlated core genes were *Cd44*, *Fn1* and *Cxcr4*. KLF6
258 regulates the switching of macrophages between pro- and anti-inflammatory
259 [20]. In accordance, *Lyz2* and *Marco*, the marker genes for inflammatory and
260 non-inflammatory macrophages, showed a correlation of 0.39 and −0.49,
261 respectively, with KLF6 binding in our single-nucleus data.

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

272 To investigate lobular zonation genes in hepatocytes, the GS was
273 taken from the list of zonation genes in normal mice (q-value < 10^{-6} in Table
274 S3 in [12]). The GS activity score showed a modest difference between
275 dietary conditions (Figure S2B), which agrees with zonation being a normal
276 phenomenon unrelated to NAFLD. When classified by the correlation with the
277 GS activity score, the genes with correlation coefficient > 0.3 were all
278 pericentral and those with correlation coefficient < −0.3 were all periportal,
279 showing that the GS activity score rediscovered zonation.

280 In stellate cells, the GS activity score for semaphorin-plexin signaling
281 was higher at 8 weeks of HFD (Figure S2C). The positively correlated core

282 gene was *Nrp2*. The negatively correlated core genes were *Nrp1*, *Plxnd1*,
283 semaphorin genes (*Sema3d*, *Sema3e* and *Sema6d*) and *Ntn1*.
284

285 **Discussion**

286 By performing single-nucleus and bulk ATAC-seq on the liver of rats fed or
287 withdrawn from the HFD, we studied the transition in tissue composition of
288 cell types and cell type-specific gene regulation in a rat model of NAFLD. The
289 accuracy of ATAC-seq-based deconvolution was validated and used for
290 analyzing cell type composition. In accordance with the pathological
291 progression from fatty liver to steatohepatitis that could have occurred
292 between 4 and 8 weeks of the HFD intervention in our rat model, the
293 proportion of inflammatory macrophages dramatically increased. In contrast,
294 after 8 weeks of HFD, the proportion of hepatocytes largely decreased. In the
295 washout condition, where 4 weeks of HFD was followed by 4 weeks of normal
296 diet, the cell type composition returned similar to rats only fed with the
297 normal diet.

298 From the global analysis of TF binding and gene expression in single
299 nuclei, we identified major modules of TF regulation (Figure 8), which could
300 be shared between different cell types or be specific to certain cell types. One
301 module shared among hepatocytes, endothelial cells, and macrophages was
302 characterized by the binding of AP-1 TFs and a biological process of TNFα
303 signaling via NF-κB. For many of the discovered modules, TFs in a given
304 module were known to regulate biological processes assignable to a set of
305 genes in the same module, suggesting the validity of our module discovery
306 algorithm. Finally, for the biological processes that emerged in TF regulation
307 analysis, by incorporating known protein-protein interactions, we could
308 identify core genes, many of which overlap with previously implicated NAFLD
309 genes (see below).

310 In each cell type, we can decompose the global variation of TF
311 regulation into major modules, where the TFs belonging to a given module
312 appear to regulate a set of genes that belong to the same module. Using the
313 data for TF binding and gene expression measured over nuclei, we extract
314 modules by applying the machine learning technique. In the discovered
315 modules, the TFs and the function of regulated genes agree with the
316 literature: e.g., the AP-1 TFs that regulate inflammation genes (Module 2 of
317 hepatocytes, Module 1 of endothelial cells and Module 1 of macrophages) are
318 known to respond to cytokine stimuli [10], and LEF1 that regulates zonation
319 (Module 3 of hepatocytes) is also known to activate Wnt signaling [13].
320 Previous snATAC-seq studies have characterized the difference between cell
321 types [21] using TF binding. In contrast, we illustrated how TFs control the
322 variation of cells within one cell type, for example, the hepatocytes with
323 various degrees of inflammation.
324 One novel method developed in this article is the computation of the
325 GS activity score in single cells to indicate the activity of GS, as if it was the
326 expression of a specific gene. In experiments, gene expressions can be
327 directly measured. The activity of a pathway is of interest but not directly
328 measurable. One way would be to take the average expression of genes in
329 that pathway, but this is not optimal for the following reasons. A simple
330 average would give equal weight to both core and periphery genes in the
331 pathway, increasing noise from the latter. Moreover, there are activator and
332 suppressor genes that contribute to the pathway in opposite directions.
333 Instead of simple average, we used the first principal component to deal with
334 these issues. The principal component approach was adapted from the PLAGE
335 program [22], which is used to compute pathway activity scores in bulk tissue
336 specimens. The GS activity score could improve the understanding of
337 biological processes. For example, *Abcg1* and *Sqle* correlated positively and
338 negatively, respectively, with the activity score for the steroid metabolism

339 pathway (Figure S2A). As this score became higher under the HFD, the GS
340 activity represented by the score presumably works in response to the
341 overdose of cholesterol contained in the diet. Thus, a positive correlation of
342 *Abcg1* (for lipid efflux) and a negative correlation of *Sqle* (for sterol
343 biosynthesis) are reasonable. Moreover, the cell type-specific nature of the
344 GS activity score could unravel cell type-specific dynamics; the inflammation
345 process peaked under 4 weeks of HFD in macrophages and under 8 weeks of
346 HFD in hepatocytes (Figure 7). The GS activity score can also be applied to
347 single-cell and single-nucleus RNA-seq data.

348 Another novel method in this article is finding the core genes of
349 biological processes. Given a GS comprising known genes for the process, we
350 identified core genes as those that show strong correlation with the GS
351 activity score as well as abundant protein-protein interactions with the GS.
352 We obtained evidence supporting the former from single-nucleus experiment
353 data and the latter from a public database. We sought literature indicating
354 that core genes cause NAFLD or serve as biomarkers for NAFLD and found
355 indications for a large part of the core genes (see below): for the TNFα
356 signaling via NF-κB process, 4 of 5 in hepatocytes, 4 of 8 in endothelial cells,
357 and 1 of 3 in macrophages; for steroid metabolism 3 of 9 in hepatocytes; for
358 semaphorin-plexin signaling 3 of 7 in stellate cells. The overlap with known
359 NAFLD genes suggests the biological validity of our data-driven approach,
360 using snATAC-seq data from a rat model of dietary intervention as input,
361 extracting modules, biological processes, and then core genes. The non-
362 overlapping genes can be candidates for future functional study of NAFLD.
363 Regarding TNFα signaling via NF-κB in hepatocytes, in animal studies, *Tnf* KO
364 mouse was protected from NAFLD [23], *Nfkb1* KO mouse exacerbated NAFLD
365 [24], and *Il1r1* KO mouse had less hepatic triacylglycerol [25]. CXCL12 (aka
366 SDF1) induced migration of stellate cells [26], endothelial cells [19] and $CD4^+$
367 T cell [27]. In human biomarker studies, higher circulating TNFα level was

368 associated with NAFLD [28] and future NAFLD [29]. In NAFLD patients,
369 soluble IL1R1 was higher (Ajmera et al., 2017). Regarding TNFα signaling via
370 NF-κB in endothelial cells, in animal studies, *Pecam1* KO mouse developed
371 NASH [30], whereas *Tlr4* KO mouse [31] and *Il15* KO mouse [32] were
372 protected from NAFLD. *Ccr5* antagonist eased NAFLD in mice [33]. Regarding
373 TNFα signaling via NF-κB in macrophages, *Cd44* KO mouse had less liver
374 injury under NAFLD inducing diet, and NAFLD patients had elevated serum
375 soluble CD44 [34]. Regarding steroid metabolism in hepatocytes, *Scd1* KO
376 mouse exacerbated NAFLD [35], and *Acox2* KO mouse developed NAFLD [36].
377 NAFLD patients had lower serum APOA1 [37,38]. Regarding semaphorin-
378 plexin signaling in stellate cells, in animal studies, *Nrp2* conditional KO
379 reduced liver fibrosis in mice [39]. *Nrp1* KO attenuated cell motility and
380 overexpression increased motility [40]. *Sema3e* KO eased liver fibrosis in
381 mice [41].

382 In search of drug targets or biomarkers of NAFLD, we looked up
383 proteins whose subcellular localization is extracellular or plasma membrane
384 proteins [42] among the genes strongly upregulated under a dietary
385 intervention (Table S2) or the core genes positively correlated with a NAFLD-
386 related activity. A number of genes have been reported as NAFLD biomarkers
387 in humans. For hepatocytes, we found *Il1r1* [43], *Il1r2* and *Tnf* [28]. For
388 endothelial cells, there was *Hgf* [44–46]. *Pecam1* was listed for both
389 endothelial and stellate cells. For stellate cells, there were *Faslg* [47] and
390 *Nrp2*. For macrophages, there were RT1-Db1, *Cd44* [34], *Cxcr4* and *Fn1*.

391 This study has several limitations. Firstly, it was difficult to distinguish
392 TFs with very similar motifs. A specific TF was regarded to bind a specific site
393 if the chromatin was open near the site and if the DNA sequence of the site
394 matched with the known TF-binding motif. However, instead, a different TF
395 with a similar motif could have bound there. Secondly, the interpretation of
396 gene expression requires caution. Gene expression can be quantified by the

397 amount of either A) mRNA transcription per time, B) mRNA in the nucleus,
398 C) mRNA in the cell, D) protein translation per time, or E) protein.
399 Measurement for bulk tissue is conducted by ATAC-seq (A), RNA-seq (C),
400 Ribo-seq (D), or Western blotting (E), and measurement for single
401 nucleus/cell is performed by single-nucleus ATAC-seq (A), single-nucleus
402 RNA-seq (B), or single-cell RNA-seq (C). While our measurement, A, could
403 reflect the state of gene regulation, it could differ from E, which reflects the
404 state of metabolism, and the fold-change value would differ among the types
405 of gene expression.
406
407 **Conclusions**
408 By performing single-nucleus and bulk ATAC-seq, we analyzed the transition
409 of cell type composition and cell type-specific gene expression in a rat model
410 of NAFLD. Using novel statistical methods, we elucidated a global picture of
411 *in vivo* TF regulation in each cell type as a set of modules and discovered core
412 genes for NAFLD-relevant biological processes. The core genes can serve as
413 candidates for novel NAFLD genes and our statistical methods can be applied
414 to elucidate the TF regulation in other biological systems.
415

416 **Methods**

417 **Animal experiment**

418 Male SHR/Izm strain rats were fed a normal diet (SP diet, Funabashi Farm,
419 Japan) or a high-fat atherogenic diet (24% fat, 15% protein, 5% cholesterol,
420 2% cholic acid). Compared to models based on nutritional deficiency (such
421 as methionine and choline-deficient diet) or chemical-induced models (such
422 as carbon tetrachloride), HFD models resemble more human NAFLD, both
423 pathophysiologically and phenotypically [48]. SHR can mimic NAFLD under a
424 hypertension background. We used the liver's largest lobe (left lateral lobe)
425 for pathology imaging and the right segment of the second largest lobe (left
426 median lobe) for genomics analysis. All animal experiments conformed to the
427 Guidelines for Animal Experiments of the National Center for Global Health
428 and Medicine (NCGM) and were approved by the Animal Research Committee
429 of NCGM (permission number 21009).

430

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

432 Nuclei were isolated from flash-frozen liver tissue according to the 10x
433 Genomics protocol CG000212 (samples m154207 and m154211) or using the
434 Minute Single Nucleus Isolation Kit for Tissues/Cells SN-047 (samples
435 m167108, m167203 and m168101). For each sample, the sequencing library
436 was prepared using the Chromium Next GEM Single Cell ATAC Reagent Kits
437 v1.1 and sequenced using the Novaseq 6000 sequencer.

438

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

440 Bulk tissue ATAC-seq was performed according to the Omni-ATAC protocol
441 [49] using flash-frozen liver tissue. The library was sequenced using the
442 NextSeq sequencer.

443

444 **Measuring triglycerides in liver tissue**

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

454

455 **Measuring total cholesterol in liver tissue**

456 Total cholesterol in liver tissue was measured enzymatically using the Total
457 Cholesterol and Cholesteryl Ester Colorimetric Assay Kit II (BioVision, K623-
458 100). Tissue homogenate was obtained by dispersing 15 mg tissue in 600 ul
459 $CHCl_3$:IPA:NP-40 (7:11:0.1) at 2200 rpm for 1 min, using a handheld
460 disperser (KINEMATICA, PT 1300 D). The homogenate was then centrifuged
461 for 5 min at 15,000 *g*. Thereafter, we transferred 200 ul of the liquid phase
462 to a new tube, dried with air-flow-in at 40 C for 10 min using a centrifugal
463 concentrator (TOMY Digital Biology, CC-105), and dissolved the dried lipids
464 with 400 ul Cholesterol Assay Buffer. Total cholesterol was measured in a
465 colorimetric assay according to the manual.

466

467 **Data cleaning for snATAC-seq**

468 Data cleaning was conducted separately by samples. ATAC-seq fragments
469 were assigned to single nuclei, and nuclei of good quality were retained as
470 described below. In principle, one nucleus and one barcoded bead are
471 contained in one droplet of the Chromium assay; thus, ATAC-seq fragments
472 from the same nucleus are identifiable by the shared unique barcode. Using

473 the Cell Ranger ATAC software (version 2.0.0), we mapped the paired-end
474 reads to the mRatBN7.2/rn7 rat reference genome and identified ATAC-seq
475 fragments, which were assigned to single barcodes.

476 Occasionally in the Chromium assay, more than one barcoded beads
477 are contained in one droplet, in which case, excessive ATAC-seq fragments
478 would be assigned in common to the barcodes. Such cases were detected,
479 and the barcodes were unified and renamed using the bap2 software (version
480 0.6.7) [50]. In the downstream analysis, we can assume that one barcode
481 corresponds to one droplet.

482 We managed the snATAC-seq data using the ArchR package (version
483 1.0.1) [51] of the R software (version 4.1.1). ATAC-seq fragments from the
484 autosomes and chromosome X were retained. We retained nuclei that had
485 abundant and enriched fragments around transcription start sites (TSS) as
486 follows: ≥1000 fragments within 1 kb to TSS, the proportion of fragments
487 within 1 kb to TSS being ≥7.5% (samples m154207, m154211 and m167203)
488 or ≥15% (samples m167108 and m168101), and Tn5 insertions (which occur
489 at both ends of a fragment) within 25 bp to TSS occurring ≥4 times frequently
490 than the insertions at 1900-2000 bp from TSS.

491 For each nucleus, the proportion of fragment counts from a
492 chromosome (e.g., chromosome 1) is approximately constant, being
493 proportional to the chromosome length, but could deviate if a droplet is
494 contaminated by free DNA from broken nuclei [52]. Therefore, we
495 represented the proportion by a matrix indexed by nuclei (rows) and
496 chromosomes (columns). To standardize the matrix, we subtracted column-
497 wise averages and divided by the standard deviation. By taking the sum of
498 the absolute values of a row, we defined the deviation score of a nucleus. In
499 order to omit contaminated droplets, we discarded nuclei with a score ≥40.

500 Since Tn5 inserts less efficiently into nucleosome-associated DNA,
501 ATAC-seq fragment size distribution shows periodicity [21]. Using the

502 program provided by Cusanovich et al.
503 (https://atlas.gs.washington.edu/mouse-atac/docs/), for each nucleus, we
504 computed the spectral densities for the 100-300 bp range, which is the pitch
505 of nucleosomes. In order to omit nuclei with an aberrant chromatin state, we
506 discarded nuclei below the 1st or above the 99th percentile in this score.
507 When two or more nuclei are contained in one droplet, called a
508 multiplet, those nuclei share the same barcode, violating the one-nucleus-to-
509 one-barcode assumption. We removed the barcodes for multiplets in two
510 ways. First, for the set of ATAC-seq fragments assigned to a barcode, we
511 inferred the number of homologous chromosomes from which the fragments
512 were derived. The number is analogous to ploidy, but differs because the
513 container is not a cell, but a droplet. By counting overlaps among the
514 fragments, the number of homologous chromosomes was inferred either 2*n*,
515 4*n* or ≥6*n*. Since the majority of hepatocytes in rat liver are diploids or
516 tetraploids [53] and other cell types in liver are diploids, ≥6*n* droplets were
517 regarded as possible multiplets and discarded. Second, we detected doublets
518 formed by two heterogeneous cell types using ArchR. Droplets with doublet
519 enrichment score >10 were discarded.
520

521 **Chromatin accessibility peak discovery and nucleus clustering for**
522 **snATAC-seq**

523 The Tn5 insertions observed in ATAC-seq occur at open chromatin regions,
524 such as promoters and enhancers, and the genome-wide profile of Tn5
525 insertions characterizes the types of cells. We clustered the nucleus based on
526 the similarity of Tn5 insertions. Firstly, we represented the Tn5 insertions in
527 snATAC-seq as a "tile matrix": a binary matrix indexed by tiles of 500 bp
528 width across the genome (rows) and nuclei (columns), coding the occurrence
529 (1) or non-occurrence (0) of Tn5 insertions. Using ArchR, we obtained a
530 matrix with 5,230,329 rows and 14,615 columns. We then grouped the nuclei

according to the similarity of corresponding columns (see next section). For each group, pseudo-bulk ATAC-seq datasets were generated, and Tn5 insertion peaks were computed using ArchR and the MACS2 program (version 2.1.4) [54]. We obtained 599,035 chromatin accessibility peaks of width 501 bp.

Secondly, we represented the Tn5 insertions as a "peak matrix": a binary matrix indexed by peaks (rows) and nuclei (columns), coding the occurrence or non-occurrence of Tn5 insertions. Again, we grouped the nuclei according to the similarity of corresponding columns. The nuclei clustering based on the peak matrix was used in downstream analysis, as it would be more accurate than the clustering based on the tile matrix.

Algorithm for dimensionality reduction and nucleus clustering

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 dimensionality reduction, we used latent semantic indexing [55] 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

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

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 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 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) [56].
4. Take 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 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) [57] and optimizing modularity with the Leiden algorithm (version 0.8.4) [58]. By merging the clusters, define coarse clustering that clearly separates in the VD coordinates. In our data, the coarse clusters corresponded to hepatocytes, stellate cells, endothelial cells and white blood cells.
7. Compute sample batch correction factor γ_{ij} in (2) by incorporating the coarse clustering of nuclei.
8. Compute the matrix $(g_i \cdot d_j \cdot f_{ij} \cdot \gamma_{ij})$ and centralize each row by subtracting the average.
9. Compute truncated SVD, UDV^T, for the 30 largest singular values.
10. Take the truncated SVD for the non-noise singular values, and project the nuclei to the VD coordinates. We took the eight largest singular values.
11. Identify clusters of nuclei by the shared nearest neighbor modularity optimization.

Cell type assignment for snATAC-seq

For each nuclei cluster obtained above, we assigned a cell type based on

614 chromatin accessibility at known marker genes. For each cluster, chromatin
615 accessibility at the TSS and gene body of the genes were visually inspected
616 using the browser of ArchR. We used the following marker genes in the rat:

617 - *Alb*, *Apoc3*, *Cyp2e1* and *Cyp2f4* for hepatocyte
618 - *Epcam* and *Krt19* for cholangiocyte
619 - *Stab2* for endothelial cell
620 - *Dcn*, *Des*, *Lrat* and *Acta2* for stellate cell
621 - *Csf3r* for granulocyte
622 - *Elane*, *Ly6g6c*, *Ly6g6d*, *Ly6g6e* and *Mpo* for neutrophil
623 - *Csf1r* for macrophage (including Kupffer cell)
624 - *Lyz2* for inflammatory macrophage
625 - *Marco* for non-inflammatory macrophage
626 - *Cd3g* and *Gata3* for T cell
627 - *Gzma* and *Prf1* for cytotoxic T cell and NK cell
628 - *Ebf1* for B cell

629

630 **Data cleaning for bulk ATAC-seq**

631 According to the ATAC-seq data processing workflow [59], we trimmed reads,
632 mapped reads to the reference genome, removed mitochondrial DNA reads,
633 subsampled to equate the library complexity, and removed PCR duplicates.

634

635 **Inferring cell type composition of bulk tissue samples**

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

642 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) [5] 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 [60]. We compute the correction factor

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

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 [61], we solve

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

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

where σ^2 is the parameter for error variance. Peaks with large variability e_i^2 are allowed to have a larger variance [62]. 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 indicators of dietary condition, with the normal diet being the reference. Using Bonferroni's correction for multiple testing of seven cell types and three conditions (apart from normal diet), the significance level was set to $0.05/[7 \times 3]$.

Quantifying gene expression in snATAC-seq

For each nucleus, the expression level of a gene was inferred from the number of Tn5 insertions in the gene body and vicinity. The targeted biological phenomenon is the chromatin opening of the gene rather than the amount of mRNA in a nucleus or cell. Using ArchR, we computed the Gene Score Matrix, which is indexed by genes (rows) and nuclei (columns) and represents the insertion counts. As the matrix is analogous to the count matrix obtained in single-cell RNA-seq, we applied library size normalization of the R package Linnorm (version 2.16.0) [63] and imputation of the R package SAVER (version 1.1.2) [64]. Before imputation, 47% of the matrix elements were

727 zero, and after imputation, 0.2% were set to the lowest value (0.001). We
728 excluded low expression genes, where the lowest value occurred in ≥1% of
729 nuclei. We took log2 of the matrix and applied quantile normalization. The
730 obtained matrix was used for inspecting cell type marker genes. We applied
731 SVD and subtracted the first two components from the matrix, which captured
732 the sample batch effect. Although the two components also captured major
733 cell type differences, we anticipated that the removal would not harm
734 downstream analysis, which compares nuclei of the same cell type between
735 samples.
736
737 **Quantifying TF binding in snATAC-seq**
738 If the DNA sequence of a chromatin accessibility peak includes the motif of a
739 TF, Tn5 insertion in the peak suggests the binding of the corresponding TF.
740 For each nucleus, by aggregating over all peaks, we can infer the genome-
741 wide binding of the TF. This is obtained as the chromVAR [65] deviation z-
742 score in ArchR. The CIS-BP database of TF binding motifs [66] was used. In
743 the same way as we did for the Gene Score Matrix, we applied quantile
744 normalization to the Motif Matrix and subtracted the first two components of
745 SVD.
746
747 **Gene set enrichment analysis of differentially expressed genes under**
748 **HFD**
749 To seek which biological processes and/or pathways are differentially
750 expressed under a particular dietary intervention, we performed gene set
751 enrichment analysis. The target GS's were obtained from the Molecular
752 Signatures Database (MSigDB) (version 7.4) [67] using the R package
753 msigdbr (version 7.4.1). Among the datasets in MSigDB, we used Gene
754 Ontology for Biological Process, Hallmark gene sets, KEGG pathways,
755 Reactome pathways, and WikiPathways. Using the sample batch-corrected

Gene Score Matrix, we compared nuclei from a pair of dietary conditions, for example, the hepatocytes from 4 weeks of HFD *vs.* the normal diet. The statistical significance of gene set enrichment was tested using the R package PADOG (version 1.34.0) [68] and by performing permutation for 10,000 times or 20 times the number of GS's in a dataset, whichever was larger. Within each dataset, the GS's that attained a false discover rate <0.05 were considered statistically significant. Significant GS's were visually inspected using the EnrichmentMap App (version 3.3.3) of the Cytoscape software (version 3.8.2).

Discovering major modules of TF regulation and assigning biological processes

To elucidate a global picture of TF regulation, we inferred the regulatory impacts of all TFs on all genes and then extracted major modules of the regulatory relationship between TFs and genes. For each gene, we computed its regulatory TFs using the R package GENIE3 (version 1.14.0) [69], which implements a random forest algorithm and thus can model nonlinear regulation. The algorithm aimed to predict the expression level of the gene based on the binding of TFs and dietary conditions. GENIE3 quantified the impacts of regulators by the proportion of explained variance. By combining the results from all genes, we obtained a regulator-regulatee matrix, in which the rows correspond to TFs and dietary conditions, the columns correspond to genes, and an element represents the importance of a TF to a gene.

Next, we applied the nonnegative matrix factorization implemented in the R package NMF (version 0.23.0) [70] to the regulator-regulatee matrix. We used the Brunet algorithm [71] with 100 random starts. The number of factors was chosen based on data as the inflection point of residual error [72]. In the output of NMF, the membership of TFs (regulators) and genes (regulatees) to a factor was defined by nonnegative weight. We called the

785 factors as "modules" of TF regulation.

786 We discovered biological processes that characterize a module by
787 applying gene set enrichment analysis to genes belonging to the module. For
788 each module of TF regulation, we tested if heavily weighted genes were
789 enriched in some GS. For each GS (see above section), we computed the P-
790 value and *t*-statistics to test if genes in the given GS had a larger weight than
791 other genes. To account for multiple testing by the GS's, we performed 100
792 trials of shuffling gene names and computed family-wise error rates. The GS's
793 that attained a family-wise error rate of <0.05 were considered statistically
794 significant.

795

796 **GS activity score in single nuclei**

797 To infer the activity of a GS in single nuclei, we computed the "average" level
798 of TF binding and gene expression, with regards to the TFs/genes included in
799 the GS. Formally, we defined the weight of TFs/genes as one if it was included
800 in the GS and as zero otherwise. We then performed SVD for the TF binding
801 and gene expression matrix incorporating the weight. In the Motif Matrix (see
802 above), each TF (row) was normalized by subtracting the mean and being
803 divided by the standard deviation and then multiplied by the weight of the TF.
804 Similarly, in the Gene Score Matrix (see above), each gene (row) was
805 normalized by subtracting the mean and being divided by the standard
806 deviation and then multiplied by the weight of the gene. By stacking the two
807 matrices, we obtained a large matrix, in which the rows correspond to TFs
808 and genes and the columns correspond to nuclei. Finally, SVD was applied
809 and the first right-singular vector was defined as the GS activity score in
810 single nuclei.

811 Although the abovementioned activity score can always be computed,
812 it would be meaningless if the TFs/genes in the GS were not co-expressed.
813 Therefore, we conducted a statistical test to exclude such meaningless GS.

814 We checked whether the TFs/genes in the GS were mutually correlated more
815 than by chance. We computed the null distribution of the first singular value
816 by performing 100 trials of the SVD procedure but with the weight of TFs
817 permutated and the weight of genes permutated and verified whether the
818 actual first singular value could attain an empirical P-value of <0.05.
819 The principal component approach was adapted from the PLAGE
820 program [22] that was developed to compute the pathway activity score of a
821 GS in bulk tissue specimens. Instead, we applied to nuclei. Compared with
822 the number of specimens, the large number of nuclei enabled significance
823 testing by permutation.
824

Availability of data and materials

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

Competing interests

The authors declare no competing interests.

Funding

This work was supported by the NCGM Intramural Research Fund (20A1013) and by JSPS KAKENHI [grant number JP16K07218].

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.

846 **REFERENCES**

847

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

850 2. Chu AL, Schilling JD, King KR, Feldstein AE. The Power of Single Cell Analysis for the
851 Study of Liver Pathobiology. Hepatology [Internet]. 2020;hep.31485-23. Available from:
852 https://aasldpubs.onlinelibrary.wiley.com/doi/epdf/10.1002/hep.31485

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

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

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

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

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

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

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

873 10. Hess J, Angel P, Schorpp-Kistner M. AP-1 subunits: quarrel and harmony among
874 siblings. J Cell Sci. 2004;117:5965-73.

875 11. Zinzalla G. A New Way Forward in Cancer Drug Discovery: Inhibiting the SWI/SNF
876 Chromatin Remodelling Complex. Chembiochem. 2016;17:677-82.

877 12. Halpern KB, Shenhav R, Matcovitch-Natan O, Tóth B, Lemze D, Golan M, et al. Single-
878 cell spatial reconstruction reveals global division of labour in the mammalian liver. Nature.
879 2017;542:352-6.

880 13. Sun J, Weis WI. Biochemical and Structural Characterization of β-Catenin Interactions
881 with Nonphosphorylated and CK2-Phosphorylated Lef-1. J Mol Biol. 2011;405:519-30.

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

885 15. Rasala BA, Orjalo AV, Shen Z, Briggs S, Forbes DJ. ELYS is a dual
886 nucleoporin/kinetochore protein required for nuclear pore assembly and proper cell division.
887 Proc National Acad Sci. 2006;103:17801-6.

888 16. Yu X, Chen X, Zheng XD, Zhang J, Zhao X, Liu Y, et al. Growth Differentiation Factor 11
889 Promotes Abnormal Proliferation and Angiogenesis of Pulmonary Artery Endothelial Cells.
890 Hypertension. 2018;71:729-41.

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

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

896 19. Salcedo R, Wasserman K, Young HA, Grimm MC, Howard OMZ, Anver MR, et al. Vascular
897 Endothelial Growth Factor and Basic Fibroblast Growth Factor Induce Expression of CXCR4
898 on Human Endothelial Cells In Vivo Neovascularization Induced by Stromal-Derived Factor-
899 1α. Am J Pathology. 1999;154:1125-35.

900 20. Date D, Das R, Narla G, Simon DI, Jain MK, Mahabeleshwar GH. Kruppel-like
901 Transcription Factor 6 Regulates Inflammatory Macrophage Polarization*. J Biol Chem.
902 2014;289:10318-29.

903 21. Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, et al. A
904 Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell. 2018;174:1309-
905 1324.e18.

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

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

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

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

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

919 27. Boujedidi H, Robert O, Bignon A, Cassard-Doulcier A-M, Renoud M-L, Gary-Gouy H, et
920 al. CXCR4 dysfunction in non-alcoholic steatohepatitis in mice and patients. Clin Sci.
921 2014;128:257-67.

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

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

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

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

934 32. Cepero-Donates Y, Lacraz G, Ghobadi F, Rakotoarivelo V, Orkhis S, Mayhue M, et al.
935 Interleukin-15-mediated inflammation promotes non-alcoholic fatty liver disease. Cytokine.
936 2016;82:102-11.

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

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

943 35. Li ZZ, Berk M, McIntyre TM, Feldstein AE. Hepatic Lipid Partitioning and Liver Damage in
944 Nonalcoholic Fatty Liver Disease ROLE OF STEAROYL-CoA DESATURASE*. J Biol Chem.
945 2009;284:5637-44.

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

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

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

953 39. Rigotti FDA, Wiedmann L, Hubert MO, Vacca M, Hasan SS, Moll I, et al. Semaphorin 3C
954 exacerbates liver fibrosis. Biorxiv. 2022;2021.07.29.454292.

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

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

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

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

967 44. Koutsogiannis D, Summers K, George B, Adams P, Marotta P, Chakrabarti S.
968 Identification of Serum Biomarkers in End Stage Liver Disease. Open Biomarkers J.
969 2010;3:1-6.

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

973 46. An X, Liu J, Zheng X, Dou Z, Li Y, Suo Y, et al. Serum HGF ,PCIII and PLT are
974 Noninvasive Markers for the diagnosis of nonalcoholic fatty liver disease. 2019;

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

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

980 49. Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S,
981 et al. An improved ATAC-seq protocol reduces background and enables interrogation of
982 frozen tissues. Nature Methods [Internet]. 2017;14:959-62. Available from: https://sci-
983 hub.tw/10.1038/nmeth.4396

984 50. Lareau CA, Ma S, Duarte FM, Buenrostro JD. Inference and effects of barcode
985 multiplets in droplet-based single-cell assays. Nature Communications. 2020;1-9.

986 51. Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, et al. ArchR is a
987 scalable software package for integrative single-cell chromatin accessibility analysis. Nature
988 Genetics. 2021;1-31.

989 52. Orchard P, Kyono Y, Hensley J, Kitzman JO, Parker SCJ. Quantification, Dynamic
990 Visualization, and Validation of Bias in ATAC-Seq Data with ataqv. Cell Systems.
991 2020;10:298-306.e4.

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

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

998 55. Berry MW, Browne M. Understanding Search Engines. 2nd edition. SIAM; 2005.

999 56. Baglama J, Reichel L. Augmented Implicitly Restarted Lanczos Bidiagonalization
1000 Methods. Siam J Sci Comput. 2005;27:19-42.

1001 57. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated
1002 analysis of multimodal single-cell data. Cell. 2021;184:3573-3587.e29.

1003 58. Traag VA, Waltman L, Eck NJ van. From Louvain to Leiden: guaranteeing well-connected
1004 communities. Sci Rep-uk. 2019;9:5233.

1005 59. Reske JJ, Wilson MR, Chandler RL. ATAC-seq normalization method can significantly
1006 affect differential accessibility analysis and interpretation. Epigenetics & chromatin. 2020;1-
1007 17.

1008 60. Jew B, Alvarez M, Rahmani E, Miao Z, Ko A, Garske KM, et al. Accurate estimation of
1009 cell composition in bulk expression through robust integration of single-cell information.
1010 Nature Communications. 2020;1-11.

1011 61. Teschendorff AE, Breeze CE, Zheng SC, Beck S. A comparison of reference-based
1012 algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies.
1013 BMC Bioinformatics. 2017;18:105.

1014 62. Wilson DR, Jin C, Ibrahim JG, Sun W. ICeD-T Provides Accurate Estimates of Immune
1015 Cell Abundance in Tumor Samples by Allowing for Aberrant Gene Expression Patterns. J Am
1016 Stat Assoc. 2019;115:1-23.

1017 63. Yip SH, Wang P, Kocher J-PA, Sham PC, Wang J. Linnorm: improved statistical analysis
1018 for single cell RNA-seq expression data. Nucleic Acids Research. 2017;45:e179.

1019 64. Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, et al. SAVER: gene
1020 expression recovery for single-cell RNA sequencing. Nature Methods. 2018;1-11.

1021 65. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. chromVAR: inferring transcription-
1022 factor-associated accessibility from single-cell epigenomic data. Nat Methods. 2017;14:975-
1023 8.

1024 66. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al.
1025 Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell.
1026 2014;158:1431-43.

1027 67. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene
1028 set enrichment analysis: A knowledge-based approach for interpreting genome-wide
1029 expression profiles. Proceedings of the National Academy of Sciences of the United States
1030 of America. 2005;102:15545-50.

1031 68. Tarca AL, Draghici S, Bhatti G, Romero R. Down-weighting overlapping genes improves
1032 gene set analysis. Bmc Bioinformatics. 2012;13:136.

1033 69. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring Regulatory Networks from
1034 Expression Data Using Tree-Based Methods. Plos One. 2010;5:e12776.

1035 70. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC
1036 Bioinformatics. 2010;11:367.

1037 71. Brunet J-P, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern
1038 discovery using matrix factorization. Proc National Acad Sci. 2004;101:4164-9.

1039 72. Hutchins LN, Murphy SM, Singh P, Graber JH. Position-dependent motif characterization
1040 using non-negative matrix factorization. Bioinformatics. 2008;24:2684-90.

1041

1042

1043 **FIGURE LEGENDS**

1044 **Figure 1**

1045 Cell types observed in the liver and their composition

1046 (A) Experimental design.

1047 (B) UMAP plot of the observed nuclei. The number of observed nuclei is shown

1048 in parentheses.

1049 (C) Expression level of cell type marker genes in nuclei (vertical axis) is

1050 compared between nuclei clusters (horizonal axis) in violin plots.

1051 (D) Validation of deconvolution based on ATAC-seq using the samples assayed

1052 in both single-nucleus and bulk ATAC-seq. The true cell type proportion

1053 counted in snATAC-seq (horizontal axis) is compared with the proportion

1054 inferred based on bulk ATAC-seq (vertical axis).

1055 (E) Proportion of nuclei from a specific cell type (vertical axis) is compared

1056 between dietary conditions (horizontal axis). The proportion was measured

1057 by bulk tissue ATAC-seq (n=4 samples per condition).

1058

1059 **Figure 2**

1060 Biological processes that were differentially expressed under dietary

1061 interventions

1062 The heatmap shows whether a biological process (row) was differentially

1063 expressed in a cell type under a dietary intervention compared to the normal

1064 diet (column). The coloring indicates statistical significance of gene set

1065 enrichment analysis: red if significant adjusting for multiple testing, brown if

1066 nominally significant at $P<0.01$, light brown if $P<0.05$, and white otherwise.

1067

1068 **Figure 3**

1069 Major modules of TF regulation in hepatocytes

1070 (A) The central heatmap shows the regulatory impact of TF binding (columns)

1071 on genes (rows). Major modules are highlighted by red rectangles. The

1072 heatmap at the bottom shows the membership of TFs (columns) to modules
1073 (rows). The heatmap on the right shows the membership of genes (rows) to
1074 modules (columns). The heatmaps only include the TFs/genes with
1075 membership ranking in the top five and membership weight more than half
1076 of the top.
1077 (B) Gene set enrichment analysis for genes belonging to the modules.
1078 Excessive overlap between genes belonging to a module and the GS for a
1079 biological process (horizontal axis) was tested, and the P-value was plotted
1080 (vertical axis). Bars are colored black if statistically significant, adjusting for
1081 multiple testing, and gray otherwise. Biological processes that were
1082 significant for at least one module are shown.
1083
1084 **Figure 4**
1085 Major modules of TF regulation in endothelial cells
1086 See legend in Figure 3.
1087
1088 **Figure 5**
1089 Major modules of TF regulation in stellate cells
1090 See legend in Figure 3.
1091
1092 **Figure 6**
1093 Major modules of TF regulation in macrophages
1094 See legend in Figure 3.
1095
1096 **Figure 7**
1097 Core genes for the TNFα signaling via NF-κB processes are extracted based
1098 on correlation with GS activity score and protein-protein interaction
1099 (A) Analysis in hepatocytes using the GS for TNFα signaling via NF-κB. In the
1100 top panel, the GS activity score in single nuclei (vertical axis) is compared

1101 between dietary conditions (horizontal axis) in violin plots. Box plots within
1102 the violins indicate quartiles. The score distribution under a diet intervention
1103 was compared with that under normal diet condition by the area under the
1104 receiver operating characteristic curve: * and ** indicate larger than 0.75
1105 and 0.9, respectively. In the remaining panels, TFs and genes are plotted by
1106 the correlation with the GS activity score (horizontal axis) and the total
1107 amount of protein-protein interactions with the GS (vertical axis). The points
1108 for TFs/genes having protein-protein interactions within the top 5% are
1109 colored red, and others are colored blue. The TFs/genes in the GS (middle
1110 panel) and not in the GS (bottom panel) are plotted separately. TFs are
1111 labeled in all capital letters. The GS was taken from the MSigDB Hallmark
1112 database. Protein-protein interactions were taken from the STRING database
1113 (version 11.0).
1114 (B) Analysis in endothelial cells.
1115 (C) Analysis in macrophages.
1116
1117 **Figure 8**
1118 Modules of global TF regulation, biological processes, and core genes were
1119 detected by the data-driven approach
1120 See text for the full list of core genes.
1121

Figure 1

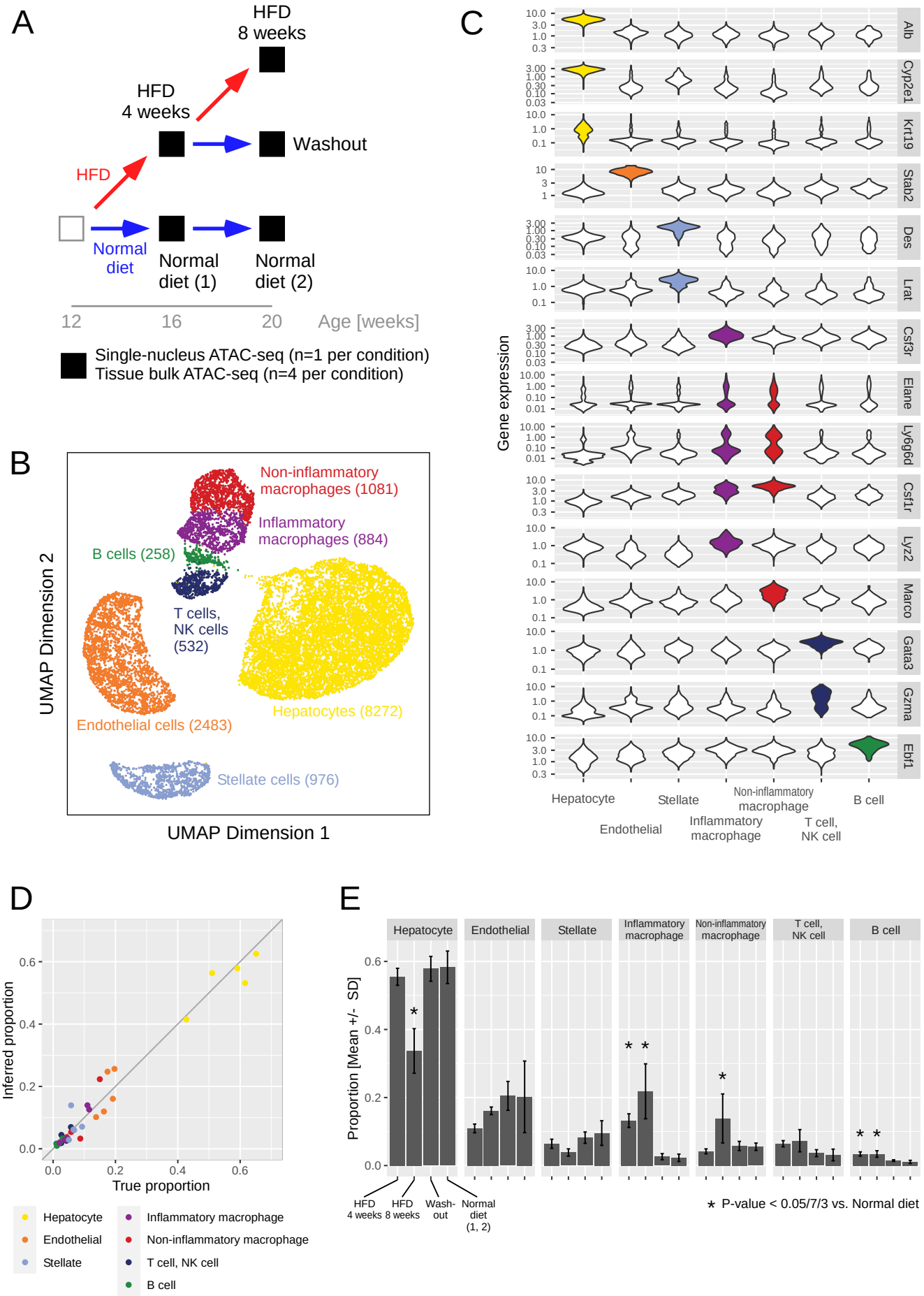
A
HFD
8 weeks
HFD
4 weeks
Washout
HFD
Normal
diet
Normal
diet (1)
Normal
diet (2)
12
16
20
Age [weeks]
Single-nucleus ATAC-seq (n=1 per condition)
Tissue bulk ATAC-seq (n=4 per condition)
B
Non-inflammatory
macrophages (1081)
Inflammatory
macrophages (884)
B cells (258)
T cells,
NK cells
(532)
Hepatocytes (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

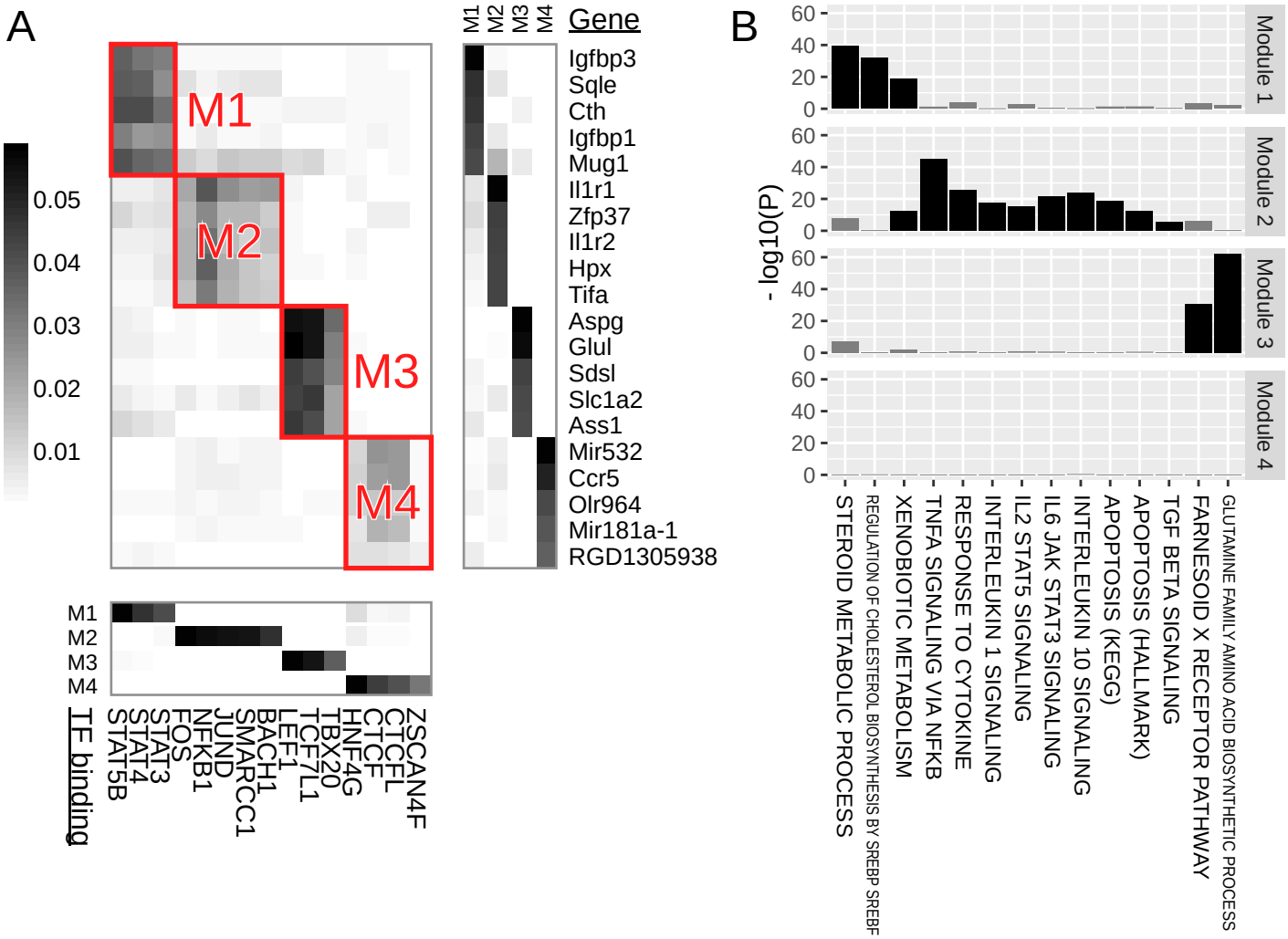

Figure 4. Endothelial

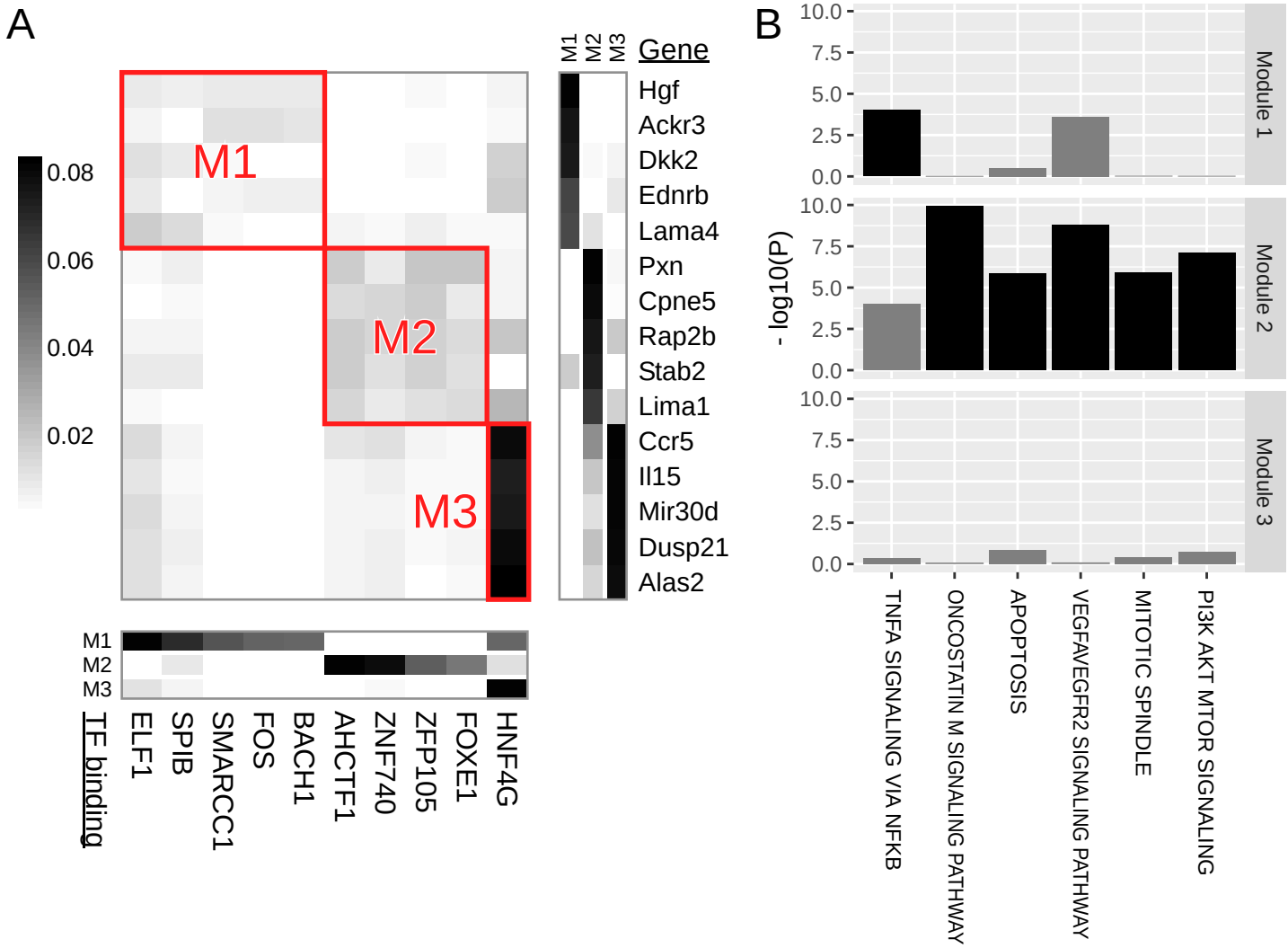

Figure 5. Stellate

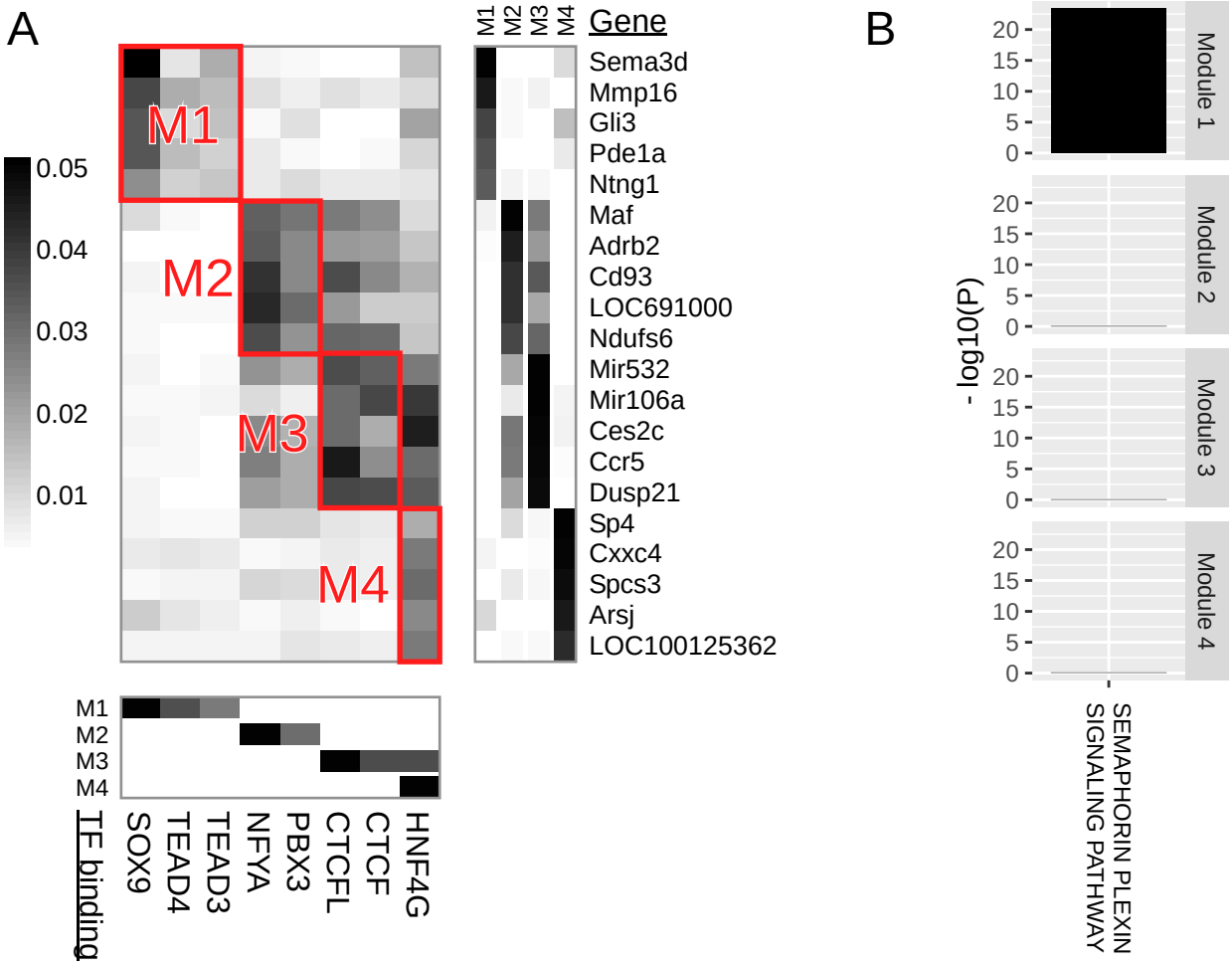

Figure 6. Macrophage

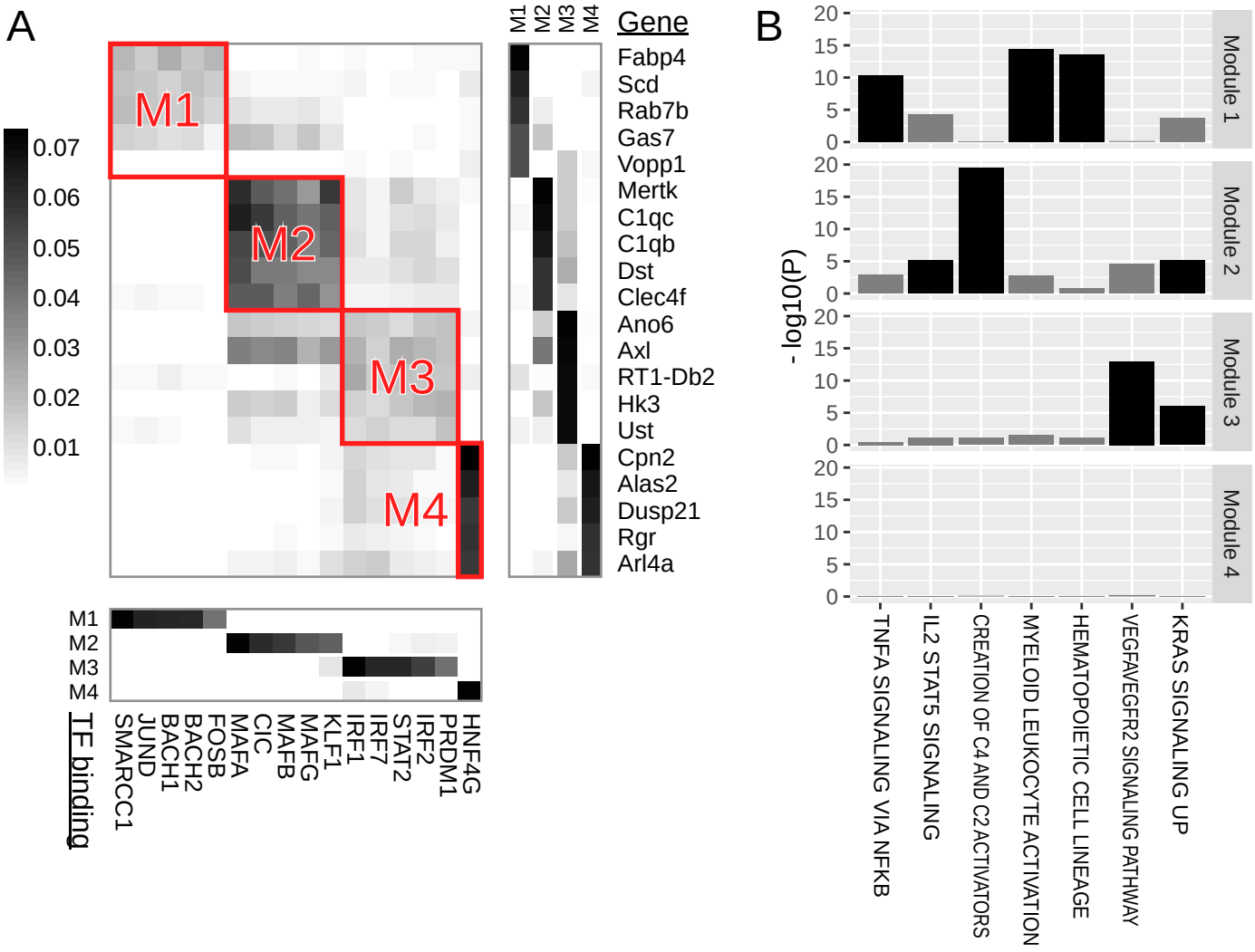

Figure 7

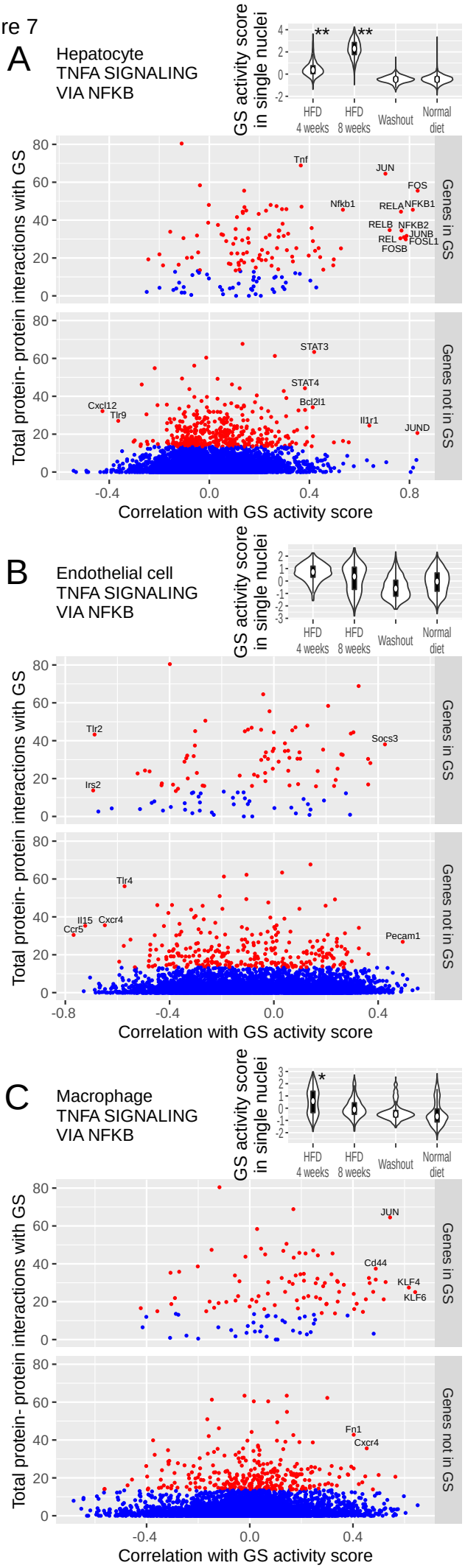
A
Hepatocyte
TNFA SIGNALING
VIA NFKB
GS activity score in single nuclei
HFD 4 weeks
HFD 8 weeks
Washout
Normal diet
Total protein- protein interactions with GS
Genes in GS
Tnf
JUN
FOS
Nfkb1
RELA
NFKB1
RELB
NFKB2
REL
JUNB
FOSL1
FOSB
Genes not in GS
STAT3
STAT4
Cxcl12
Tlr9
Bcl2l1
Il1r1
JUND
Correlation 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	Modules (TFs therein)	Biological processes	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	