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

SUMMARY

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. 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. Our statistical methods help elucidate the global gene regulation *in vivo* as a combination of modules and discover core genes of the relevant biological processes.

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

INTRODUCTION

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% (Friedman et al., 2018). 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 (Chu et al., 2020).

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 (Sturm et al., 2019; Teschendorff and Zheng, 2017). For ATAC-seq, Li and associates (Li et al., 2020) previously proposed an algorithm for deconvolution but no experimental validation has been done with single-cell data.

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75 ATAC-seq on the liver of rats fed or withdrawn from a high-fat diet (HFD).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

89 **RESULTS**

90 **Cell type composition of the liver**

91 Spontaneously Hypertensive Rats (SHR) fed HFD for 4 weeks developed fatty
92 liver, and those fed HFD for 8 weeks further developed inflammation and
93 premature fibrosis (steatohepatitis) (Figure 1A; Table S1). Under the washout
94 condition, where 4 weeks of HFD is followed by 4 weeks of a normal diet,
95 fatty liver was partially ameliorated. Along with rats only fed a normal diet,
96 five dietary conditions were examined.
97 In order to catalog cell types in the liver, we performed snATAC-seq
98 on one animal per dietary condition. We obtained a total of 14,486 nuclei
99 (~3000 per sample), which were grouped into 16 clusters by the similarity of
100 chromatin opening. Hepatocytes (seven clusters), endothelial cells (three
101 clusters), stellate cells (two clusters) and white blood cells (four clusters)

separated clearly (Figures 1B and 1C). The clusters for white blood cells corresponded to inflammatory macrophages, non-inflammatory macrophages, T cells and NK cells combined, and B cells. Cholangiocytes were included in the hepatocyte clusters, and neutrophils and Kupffer cells were included in the macrophage clusters.

To measure the changes in cell type population through dietary intervention, we performed bulk ATAC-seq on four animals per condition. Using snATAC-seq as reference, we inferred the cell type composition in the bulk ATAC-seq samples by deconvolution. Beforehand, using the five samples assayed in both single-nucleus and bulk ATAC-seq, the accuracy of deconvolution for ATAC-seq was validated (Figure 1D). The difference of the inferred proportion from the true proportion counted by snATAC-seq had average of zero and standard deviation of 0.035, indicating high accuracy. Compared with the normal diet condition, hepatocytes significantly decreased after 8 weeks of HFD (Figure 1E). Inflammatory macrophages largely increased after 4 weeks of HFD and further increased after 8 weeks of HFD. Non-inflammatory macrophages also increased after 8 weeks of HFD. In addition, B cells increased under HFD conditions. In accordance, previous studies of HFD-induced NAFLD animal models reported an increase in liver-resident macrophages (Kiki et al., 2007), recruitment of monocyte-derived macrophages (Zhong et al., 2019), and an increase in intrahepatic B cells (Barrow et al., 2021; Zhang et al., 2016). Under the washout condition, the cell type composition returned to the same as the normal diet condition. Here, we elucidated how the quantity of cell types changed through dietary intervention, and we next investigate the qualitative change.

Differential gene expression under HFD in each cell type

We analyzed cell type-specific differential gene expression by comparing the nuclei of the same cell type between dietary conditions (Table S2). Hereafter,

131 clusters for inflammatory and non-inflammatory macrophages are combined
132 and regarded as the macrophage cell type. T cells, NK cells, and B cells were
133 not analyzed because the number of nuclei was limited. The biological
134 processes characteristic of the differentially expressed genes were cell type-
135 dependent (Figure 2). Steroid metabolism and fatty acid metabolism genes
136 were expressed in hepatocytes differently under HFD and washout compared
137 with the normal diet. As of inflammation genes, the expression altered in
138 macrophages after 4 weeks of HFD and washout. The inflammation genes
139 and apoptosis genes were expressed in hepatocytes differently after 8 weeks
140 of HFD. Genes for the actin filament-based process were expressed in stellate
141 cells differently after 8 weeks of HFD.
142 Conventional differential gene expression analysis using bulk tissue
143 samples was limited in detecting 1) changes in minor cell types, 2) changes
144 in a cell type with a low basal expression of the gene relative to other cell
145 types, and 3) concurrent changes in gene expression and cell population that
146 mutually cancel, for example, downregulated genes in a cell type that is
147 increasing in population. For the comparison between the 4 weeks of HFD and
148 the normal diet, cell type-specific differential gene expression was cross-
149 checked with bulk tissue transcriptome analysis, and results were found to
150 replicate under the limitations mentioned above (Figure S1). The differential
151 gene expression was replicated in hepatocytes, but to a lesser degree in
152 minor cell types. For macrophages, whose proportion increased after 4 weeks
153 of HFD, the upregulated genes were replicated in bulk experiments but the
154 downregulated genes were not.
155
156 **Major modules of TF regulation and associated biological processes**
157 To elucidate the global picture of TF regulation, we inferred the regulatory
158 impact of all TFs on all genes and extracted major modules of the regulatory
159 relation. By snATAC-seq, we measured in all nuclei of each cell type the

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

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

180 Module 2 of hepatocytes, Module 1 of endothelial cells, and Module 1
181 of macrophages were characterized by the binding of AP-1 family TFs (namely,
182 FOS, JUND, BACH1 and BACH2), SMARCC1 and NFKB1, and by genes for
183 TNFα signaling via NF-κB. The AP-1 TFs respond to cytokine stimuli (Hess et
184 al., 2004). SMARCC1 is a member of the SWI/SNF family of proteins, which
185 function as a helicase and interacts with NF-κB (Zinzalla, 2016).

186 Module 3 of hepatocytes was characterized by the binding of TCF/LEF
187 family TFs (namely, LEF1 and TCF7L1) and by genes in the farnesoid X
188 receptor pathway for bile acid synthesis and genes for the glutamine family

189 amino acid biosynthesis. Within the liver lobule, a unit that assembles to form
190 the liver, a central vein is located in the center, the portal triad (portal vein,
191 hepatic artery, and bile duct) is located in the periphery, and the metabolism
192 of hepatocyte in between differs between the pericentral and periportal zones.
193 The glutamine family amino acid biosynthesis genes include *Glul* and *Ass1*,
194 which are highly expressed in the pericentral and periportal zones,
195 respectively (Halpern et al., 2017). In addition, the LEF1 TF binds to β-catenin
196 protein and activates Wnt signaling pathway (Sun and Weis, 2011), and liver-
197 specific β-catenin KO mice exhibited defective cholesterol and bile acid
198 metabolism in the liver (Behari et al., 2010). Moreover, Wnt-activated genes
199 are highly expressed in the pericentral zone. Thus, LEF1 TF presumably
200 regulates bile acid metabolism and zonated gene expression in the liver.
201 Module 2 of endothelial cells was characterized by the DNA binding of
202 AHCTF1 (aka ELYS) and ZNF740 and by genes for angiogenesis. AHCTF1 is a
203 DNA-binding protein that is required in mitosis (Rasala et al., 2006). ZNF740
204 activates angiogenesis in pulmonary artery endothelial cells of rats (Yu et al.,
205 2018).
206 Module 1 of stellate cells was characterized by SOX9 TF binding and
207 genes in the semaphorin-plexin signaling pathway.
208 Module 2 of macrophages was characterized by the binding of Maf
209 family TFs (namely, MAFA, MAFB and MAFG) and by genes for complement
210 system activation. In *Mafb*-deficient macrophages of mice, C1q production
211 decreased (Tran et al., 2017).
212 Module 3 of macrophages was characterized by the binding of IRF
213 family TFs (namely, IRF1, IRF2 and IRF7), which regulate immune responses,
214 and by genes for angiogenesis. IRF1 contributes to the commitment of pro-
215 inflammatory M1 macrophages, which produce angiogenic stimulators
216 (Chistiakov et al., 2018).
217

218 **Core genes in specific biological processes**

219 In the previous section we discovered major modules of TF regulation and
220 then searched across databases, such as Gene Ontology, for biological
221 processes that characterize the modules. We equated a biological process
222 with its comprising genes, and defined them as a gene set (GS). A GS
223 characterizes a module if it significantly overlaps with the genes of the module.
224 Here, we aim to find "core genes" of a biological process. In preparation, we
225 computed the activity of a GS in single nuclei as the "GS activity score" by
226 taking the average binding/expression level of TFs/genes in the set or, more
227 precisely, the first principal component. In hepatocytes, regarding the
228 biological process of TNFα signaling via NF-κB, its GS activity score was the
229 highest among nuclei from 8 weeks of HFD (Figure 7A, top panel).

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

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

247 *Nfkb1*, *Bcl2l1* and *Il1r1*. The negatively correlated core gene was *Cxcl12*.

248 In endothelial cells, the GS activity score for TNFα signaling via NF-
249 κB (Figure 7B) was higher under HFD conditions. The positively correlated
250 core genes were *Socs3* and *Pecam1*. The negatively correlated core genes
251 were *Tlr2*, *Irs2*, *Tlr4*, *Il15*, *Ccr5* and *Cxcr4* (receptor for the chemoattractant
252 CXCL12 (Salcedo et al., 1999)).

253 In macrophages, the GS activity score for TNFα signaling via NF-κB
254 (Figure 7C) was higher after 4 weeks of HFD. TF binding that was positively
255 correlated and had abundant protein-protein interactions included JUN and
256 KLF6. The positively correlated core genes were *Cd44*, *Fn1* and *Cxcr4*. KLF6
257 regulates the switching of macrophages between pro- and anti-inflammatory
258 (Date et al., 2014). In accordance, *Lyz2* and *Marco*, the marker genes for
259 inflammatory and non-inflammatory macrophages, showed a correlation of
260 0.39 and –0.49, respectively, with KLF6 binding in our single-nucleus data.

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

271 To investigate lobular zonation genes in hepatocytes, the GS was
272 taken from the list of zonation genes in normal mice (q-value < 10^{-6} in Table
273 S3 in (Halpern et al., 2017)). The GS activity score showed a modest
274 difference between dietary conditions (Figure S2B), which agrees with
275 zonation being a normal phenomenon unrelated to NAFLD. When classified

276 by the correlation with the GS activity score, the genes with correlation
277 coefficient > 0.3 were all pericentral and those with correlation coefficient <
278 −0.3 were all periportal, showing that the GS activity score rediscovered
279 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, which could be shared
300 between different cell types or be specific to certain cell types. One module
301 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 (Hess et al., 2004), and LEF1 that
319 regulates zonation (Module 3 of hepatocytes) is also known to activate Wnt
320 signaling (Sun and Weis, 2011). Previous snATAC-seq studies have
321 characterized the difference between cell types (Cusanovich et al., 2018)
322 using TF binding. In contrast, we illustrated how TFs control the variation of
323 cells within one cell type, for example, the hepatocytes with various degrees
324 of inflammation.

325 One novel method developed in this article is the computation of the
326 GS activity score in single cells to indicate the activity of GS, as if it was the
327 expression of a specific gene. In experiments, gene expressions can be
328 directly measured. The activity of a pathway is of interest but not directly
329 measurable. One way would be to take the average expression of genes in
330 that pathway, but this is not optimal for the following reasons. A simple
331 average would give equal weight to both core and periphery genes in the
332 pathway, increasing noise from the latter. Moreover, there are activator and
333 suppressor genes that contribute to the pathway in opposite directions.

334 Instead of simple average, we used the first principal component to deal with
335 these issues. The principal component approach was adapted from the PLAGE
336 program (Tomfohr et al., 2005), which is used to compute pathway activity
337 scores in bulk tissue specimens. The GS activity score could improve the
338 understanding of biological processes. For example, *Abcg1* and *Sqle*
339 correlated positively and negatively, respectively, with the activity score for
340 the steroid metabolism pathway (Figure S2A). As this score became higher
341 under the HFD, the GS activity represented by the score presumably works
342 in response to the overdose of cholesterol contained in the diet. Thus, a
343 positive correlation of *Abcg1* (for lipid efflux) and a negative correlation of
344 *Sqle* (for sterol biosynthesis) are reasonable. Moreover, the cell type-specific
345 nature of the GS activity score could unravel cell type-specific dynamics; the
346 inflammation process peaked under 4 weeks of HFD in macrophages and
347 under 8 weeks of HFD in hepatocytes (Figure 7). The GS activity score can
348 also be applied to single-cell and single-nucleus RNA-seq data.
349 Another novel method in this article is finding the core genes of
350 biological processes. Given a GS comprising known genes for the process, we
351 identified core genes as those that show strong correlation with the GS
352 activity score as well as abundant protein-protein interactions with the GS.
353 We obtained evidence supporting the former from single-nucleus experiment
354 data and the latter from a public database. We sought literature indicating
355 that core genes cause NAFLD or serve as biomarkers for NAFLD and found
356 indications for a large part of the core genes (see below): for the TNFα
357 signaling via NF-κB process, 4 of 5 in hepatocytes, 4 of 8 in endothelial cells,
358 and 1 of 3 in macrophages; for steroid metabolism 3 of 9 in hepatocytes; for
359 semaphorin-plexin signaling 3 of 7 in stellate cells. The overlap with known
360 NAFLD genes suggests the biological validity of our data-driven approach,
361 using snATAC-seq data from a rat model of dietary intervention as input,
362 extracting modules, biological processes, and then core genes. The non-

363 overlapping genes can be candidates for future functional study of NAFLD.
364 Regarding TNFα signaling via NF-κB in hepatocytes, in animal studies, *Tnf* KO
365 mouse was protected from NAFLD (Kakino et al., 2017), *Nfkb1* KO mouse
366 exacerbated NAFLD (Locatelli et al., 2012), and *Il1r1* KO mouse had less
367 hepatic triacylglycerol (Roos et al., 2009). CXCL12 (aka SDF1) induced
368 migration of stellate cells (Sawitza et al., 2009), endothelial cells (Salcedo et
369 al., 1999) and $CD4^+$ T cell (Boujedidi et al., 2014). In human biomarker
370 studies, higher circulating TNFα level was associated with NAFLD (Potoupni
371 et al., 2021) and future NAFLD (Seo et al., 2013). In NAFLD patients, soluble
372 IL1R1 was higher (Ajmera et al., 2017). Regarding TNFα signaling via NF-κB
373 in endothelial cells, in animal studies, *Pecam1* KO mouse developed NASH
374 (Goel et al., 2007), whereas *Tlr4* KO mouse (Zhang et al., 2020a) and *Il15*
375 KO mouse (Cepero-Donates et al., 2016) were protected from NAFLD. *Ccr5*
376 antagonist eased NAFLD in mice (Pérez-Martínez et al., 2014). Regarding
377 TNFα signaling via NF-κB in macrophages, *Cd44* KO mouse had less liver
378 injury under NAFLD inducing diet, and NAFLD patients had elevated serum
379 soluble CD44 (Patouraux et al., 2017). Regarding steroid metabolism in
380 hepatocytes, *Scd1* KO mouse exacerbated NAFLD (Li et al., 2009), and *Acox2*
381 KO mouse developed NAFLD (Zhang et al., 2020b). NAFLD patients had lower
382 serum APOA1 (Chen et al., 2022; Yang et al., 2016). Regarding semaphorin-
383 plexin signaling in stellate cells, in animal studies, *Nrp2* conditional KO
384 reduced liver fibrosis in mice (Rigotti et al., 2022). *Nrp1* KO attenuated cell
385 motility and overexpression increased motility (Cao et al., 2010). *Sema3e* KO
386 eased liver fibrosis in mice (Yagai et al., 2014).
387 In search of drug targets or biomarkers of NAFLD, we looked up
388 proteins whose subcellular localization is extracellular or plasma membrane
389 proteins (Binder et al., 2014) among the genes strongly upregulated under a
390 dietary intervention (Table S2) or the core genes positively correlated with a
391 NAFLD-related activity. A number of genes have been reported as NAFLD

392 biomarkers in humans. For hepatocytes, we found *Il1r1* (Ajmera et al., 2017),
393 *Il1r2* and *Tnf* (Potoupni et al., 2021). For endothelial cells, there was *Hgf* (An
394 et al., 2019; Koutsogiannis et al., 2010; Krawczyk et al., 2017). *Pecam1* was
395 listed for both endothelial and stellate cells. For stellate cells, there were *Faslg*
396 (Page et al., 2013) and *Nrp2*. For macrophages, there were RT1-Db1, *Cd44*
397 (Patouraux et al., 2017), *Cxcr4* and *Fn1*.

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

413 By performing single-nucleus and bulk ATAC-seq, we analyzed the
414 transition of cell type composition and cell type-specific gene expression in a
415 rat model of NAFLD. Using novel statistical methods, we elucidated a global
416 picture of *in vivo* TF regulation in each cell type as a set of modules and
417 discovered core genes for NAFLD-relevant biological processes. The core
418 genes can serve as candidates for novel NAFLD genes and our statistical
419 methods can be applied to elucidate the TF regulation in other biological
420 systems.

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

Declaration of interests

The authors declare no competing interests.

431 **FIGURE LEGENDS**

432 **Figure 1**

433 Cell types observed in the liver and their composition

434 (A) Experimental design.

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

436 in parentheses.

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

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

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

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

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

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

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

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

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

446

447 **Figure 2**

448 Biological processes that were differentially expressed under dietary

449 interventions

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

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

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

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

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

455

456 **Figure 3**

457 Major modules of TF regulation in hepatocytes

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

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

460 heatmap at the bottom shows the membership of TFs (columns) to modules
461 (rows). The heatmap on the right shows the membership of genes (rows) to
462 modules (columns). The heatmaps only include the TFs/genes with
463 membership ranking in the top five and membership weight more than half
464 of the top.
465 (B) Gene set enrichment analysis for genes belonging to the modules.
466 Excessive overlap between genes belonging to a module and the GS for a
467 biological process (horizontal axis) was tested, and the P-value was plotted
468 (vertical axis). Bars are colored black if statistically significant, adjusting for
469 multiple testing, and gray otherwise. Biological processes that were
470 significant for at least one module are shown.
471
472 **Figure 4**
473 Major modules of TF regulation in endothelial cells
474 See legend in Figure 3.
475
476 **Figure 5**
477 Major modules of TF regulation in stellate cells
478 See legend in Figure 3.
479
480 **Figure 6**
481 Major modules of TF regulation in macrophages
482 See legend in Figure 3.
483
484 **Figure 7**
485 Core genes for the TNFα signaling via NF-κB processes are extracted based
486 on correlation with GS activity score and protein-protein interaction
487 (A) Analysis in hepatocytes using the GS for TNFα signaling via NF-κB. In the
488 top panel, the GS activity score in single nuclei (vertical axis) is compared

489 between dietary conditions (horizontal axis) in violin plots. Box plots within
490 the violins indicate quartiles. The score distribution under a diet intervention
491 was compared with that under normal diet condition by the area under the
492 receiver operating characteristic curve: * and ** indicate larger than 0.75
493 and 0.9, respectively. In the remaining panels, TFs and genes are plotted by
494 the correlation with the GS activity score (horizontal axis) and the total
495 amount of protein-protein interactions with the GS (vertical axis). The points
496 for TFs/genes having protein-protein interactions within the top 5% are
497 colored red, and others are colored blue. The TFs/genes in the GS (middle
498 panel) and not in the GS (bottom panel) are plotted separately. TFs are
499 labeled in all capital letters. The GS was taken from the MSigDB Hallmark
500 database. Protein-protein interactions were taken from the STRING database
501 (version 11.0).
502 (B) Analysis in endothelial cells.
503 (C) Analysis in macrophages.
504

505 **METHODS**

506 **Data and code availability**

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

512

513 **Animal experiment**

514 Male SHR/Izm strain rats were fed a normal diet (SP diet, Funabashi Farm,
515 Japan) or a high-fat atherogenic diet (24% fat, 15% protein, 5% cholesterol,
516 2% cholic acid). Compared to models based on nutritional deficiency (such
517 as methionine and choline-deficient diet) or chemical-induced models (such
518 as carbon tetrachloride), HFD models resemble more human NAFLD, both
519 pathophysiologically and phenotypically (Herck et al., 2017). SHR can mimic
520 NAFLD under a hypertension background. We used the liver's largest lobe
521 (left lateral lobe) for pathology imaging and the right segment of the second
522 largest lobe (left median lobe) for genomics analysis. All animal experiments
523 conformed to the Guidelines for Animal Experiments of the National Center
524 for Global Health and Medicine (NCGM) and were approved by the Animal
525 Research Committee of NCGM (permission number 21009).

526

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

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

534

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

536 Bulk tissue ATAC-seq was performed according to the Omni-ATAC protocol
537 (Corces et al., 2017) using flash-frozen liver tissue. The library was
538 sequenced using the NextSeq sequencer.

539

540 **Measuring triglycerides in liver tissue**

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

550

551 **Measuring total cholesterol in liver tissue**

552 Total cholesterol in liver tissue was measured enzymatically using the Total
553 Cholesterol and Cholesteryl Ester Colorimetric Assay Kit II (BioVision, K623-
554 100). Tissue homogenate was obtained by dispersing 15 mg tissue in 600 ul
555 $CHCl_3$:IPA:NP-40 (7:11:0.1) at 2200 rpm for 1 min, using a handheld
556 disperser (KINEMATICA, PT 1300 D). The homogenate was then centrifuged
557 for 5 min at 15,000 *g*. Thereafter, we transferred 200 ul of the liquid phase
558 to a new tube, dried with air-flow-in at 40 C for 10 min using a centrifugal
559 concentrator (TOMY Digital Biology, CC-105), and dissolved the dried lipids
560 with 400 ul Cholesterol Assay Buffer. Total cholesterol was measured in a
561 colorimetric assay according to the manual.

562

563 **Data cleaning for snATAC-seq**

564 Data cleaning was conducted separately by samples. ATAC-seq fragments
565 were assigned to single nuclei, and nuclei of good quality were retained as
566 described below. In principle, one nucleus and one barcoded bead are
567 contained in one droplet of the Chromium assay; thus, ATAC-seq fragments
568 from the same nucleus are identifiable by the shared unique barcode. Using
569 the Cell Ranger ATAC software (version 2.0.0), we mapped the paired-end
570 reads to the mRatBN7.2/rn7 rat reference genome and identified ATAC-seq
571 fragments, which were assigned to single barcodes.

572 Occasionally in the Chromium assay, more than one barcoded beads
573 are contained in one droplet, in which case, excessive ATAC-seq fragments
574 would be assigned in common to the barcodes. Such cases were detected,
575 and the barcodes were unified and renamed using the bap2 software (version
576 0.6.7) (Lareau et al., 2020). In the downstream analysis, we can assume that
577 one barcode corresponds to one droplet.

578 We managed the snATAC-seq data using the ArchR package (version
579 1.0.1) (Granja et al., 2021) of the R software (version 4.1.1). ATAC-seq
580 fragments from the autosomes and chromosome X were retained. We
581 retained nuclei that had abundant and enriched fragments around
582 transcription start sites (TSS) as follows: ≥1000 fragments within 1 kb to TSS,
583 the proportion of fragments within 1 kb to TSS being ≥7.5% (samples
584 m154207, m154211 and m167203) or ≥15% (samples m167108 and
585 m168101), and Tn5 insertions (which occur at both ends of a fragment)
586 within 25 bp to TSS occurring ≥4 times frequently than the insertions at
587 1900-2000 bp from TSS.

588 For each nucleus, the proportion of fragment counts from a
589 chromosome (e.g., chromosome 1) is approximately constant, being
590 proportional to the chromosome length, but could deviate if a droplet is
591 contaminated by free DNA from broken nuclei (Orchard et al., 2020).

592 Therefore, we represented the proportion by a matrix indexed by nuclei
593 (rows) and chromosomes (columns). To standardize the matrix, we
594 subtracted column-wise averages and divided by the standard deviation. By
595 taking the sum of the absolute values of a row, we defined the deviation score
596 of a nucleus. In order to omit contaminated droplets, we discarded nuclei with
597 a score ≥40.

598 Since Tn5 inserts less efficiently into nucleosome-associated DNA,
599 ATAC-seq fragment size distribution shows periodicity (Cusanovich et al.,
600 2018). Using the program provided by Cusanovich et al.
601 (https://atlas.gs.washington.edu/mouse-atac/docs/), for each nucleus, we
602 computed the spectral densities for the 100-300 bp range, which is the pitch
603 of nucleosomes. In order to omit nuclei with an aberrant chromatin state, we
604 discarded nuclei below the 1st or above the 99th percentile in this score.

605 When two or more nuclei are contained in one droplet, called a
606 multiplet, those nuclei share the same barcode, violating the one-nucleus-to-
607 one-barcode assumption. We removed the barcodes for multiplets in two
608 ways. First, for the set of ATAC-seq fragments assigned to a barcode, we
609 inferred the number of homologous chromosomes from which the fragments
610 were derived. The number is analogous to ploidy, but differs because the
611 container is not a cell, but a droplet. By counting overlaps among the
612 fragments, the number of homologous chromosomes was inferred either $2n$,
613 $4n$ or $\geq 6n$. Since the majority of hepatocytes in rat liver are diploids or
614 tetraploids (Katsuda et al., 2020) and other cell types in liver are diploids,
615 $\geq 6n$ droplets were regarded as possible multiplets and discarded. Second, we
616 detected doublets formed by two heterogeneous cell types using ArchR.
617 Droplets with doublet enrichment score >10 were discarded.

618

619 **Chromatin accessibility peak discovery and nucleus clustering for**

620 **snATAC-seq**

621 The Tn5 insertions observed in ATAC-seq occur at open chromatin regions,
622 such as promoters and enhancers, and the genome-wide profile of Tn5
623 insertions characterizes the types of cells. We clustered the nucleus based on
624 the similarity of Tn5 insertions. Firstly, we represented the Tn5 insertions in
625 snATAC-seq as a "tile matrix": a binary matrix indexed by tiles of 500 bp
626 width across the genome (rows) and nuclei (columns), coding the occurrence
627 (1) or non-occurrence (0) of Tn5 insertions. Using ArchR, we obtained a
628 matrix with 5,230,329 rows and 14,615 columns. We then grouped the nuclei
629 according to the similarity of corresponding columns (see next section). For
630 each group, pseudo-bulk ATAC-seq datasets were generated, and Tn5
631 insertion peaks were computed using ArchR and the MACS2 program (version
632 2.1.4) (Zhang et al., 2008). We obtained 599,035 chromatin accessibility
633 peaks of width 501 bp.

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

640

641 **Algorithm for dimensionality reduction and nucleus clustering**

642 The tile and peak matrices are large binary matrices indexed by genomic
643 positions (rows) and nuclei (columns). Since the column vectors are
644 extremely high dimensional, we need to reduce the dimensions in order to
645 properly analyze the similarity among nuclei. For dimensionality reduction,
646 we used latent semantic indexing (Berry and Browne, 2005) and singular
647 value decomposition (SVD). The dimensionality reduction was performed for
648 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 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) (Baglama and Reichel, 2005).
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) (Hao et al., 2021) and optimizing modularity with the Leiden algorithm (version 0.8.4) (Traag et al., 2019). 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 chromatin accessibility at known marker genes. For each cluster, chromatin accessibility at the TSS and gene body of the genes were visually inspected using the browser of ArchR. We used the following marker genes in the rat:

- *Alb*, *Apoc3*, *Cyp2e1* and *Cyp2f4* for hepatocyte
- *Epcam* and *Krt19* for cholangiocyte
- *Stab2* for endothelial cell
- *Dcn*, *Des*, *Lrat* and *Acta2* for stellate cell
- *Csf3r* for granulocyte
- *Elane*, *Ly6g6c*, *Ly6g6d*, *Ly6g6e* and *Mpo* for neutrophil
- *Csf1r* for macrophage (including Kupffer cell)
- *Lyz2* for inflammatory macrophage
- *Marco* for non-inflammatory macrophage
- *Cd3g* and *Gata3* for T cell
- *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 (Reske et al., 2020), we trimmed reads, mapped reads to the reference genome, removed

732 mitochondrial DNA reads, subsampled to equate the library complexity, and
733 removed PCR duplicates.
734
735 **Inferring cell type composition of bulk tissue samples**
736 We inferred the cell type composition of liver tissue samples by comparing
737 their bulk ATAC-seq data with the reference data generated by snATAC-seq.
738 As references, pseudo-bulk ATAC-seq data of pure cell types were computed
739 from snATAC-seq results. We combined several published deconvolution
740 methods. Let the indices be i for a chromatin accessibility peak, j for a cell
741 type, and k for a sample, either assayed in bulk, in single nuclei or both.
742 By adding Tn5 insertion counts of snATAC-seq over a_{jk} nuclei of cell
743 type j from sample k, we obtain Tn5 insertion count u_{ijk} at peak i. The
744 counts across all peaks are represented as a vector $\boldsymbol{u}_{jk}$ in bold font, in which
745 the index i is dropped. We quantile normalize the vectors $\boldsymbol{u}_{jk}$ and obtain
746 vectors $\boldsymbol{v}_{jk}$. Although quantile normalization is not a linear transformation,
747 we can regard the total number of nuclei in $\boldsymbol{v}_{jk}$ as
748 $$b_{jk} = \frac{\sum_i v_{ijk}}{\sum_i u_{ijk}} \cdot a_{jk}.$$
749 We factorize the count as
750 $$b_{jk} \approx c_j \cdot d_k,$$
751 where c_j represents how many nuclei are included in the normalized ATAC-
752 seq library for pure cell type j and d_k adjusts for experimental batch effect
753 between samples. Specifically, we solve a linear regression whose dependent
754 variable is the logarithm of b_{jk} and independent variables are the indicator
755 functions of j and k, and take the exponential of the obtained coefficients.
756 For the solution to be unique, we set $d_1 = 1$. We represent by C the diagonal
757 matrix with c_j in the diagonal. The values of c_j for hepatocytes were half the
758 values for other cell types due to higher ploidy of hepatocytes.
759 We used the DeconPeaker software (version 1.0) (Li et al., 2020) to
760 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 (Jew et al., 2020). 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 (Teschendorff et al., 2017), 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 (Wilson et al., 2019). 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

816 conditions (apart from normal diet), the significance level was set to
817 $0.05/[7 \times 3]$.
818
819 **Quantifying gene expression in snATAC-seq**
820 For each nucleus, the expression level of a gene was inferred from the number
821 of Tn5 insertions in the gene body and vicinity. The targeted biological
822 phenomenon is the chromatin opening of the gene rather than the amount of
823 mRNA in a nucleus or cell. Using ArchR, we computed the Gene Score Matrix,
824 which is indexed by genes (rows) and nuclei (columns) and represents the
825 insertion counts. As the matrix is analogous to the count matrix obtained in
826 single-cell RNA-seq, we applied library size normalization of the R package
827 Linnorm (version 2.16.0) (Yip et al., 2017) and imputation of the R package
828 SAVER (version 1.1.2) (Huang et al., 2018). Before imputation, 47% of the
829 matrix elements were zero, and after imputation, 0.2% were set to the lowest
830 value (0.001). We excluded low expression genes, where the lowest value
831 occurred in ≥1% of nuclei. We took log2 of the matrix and applied quantile
832 normalization. The obtained matrix was used for inspecting cell type marker
833 genes. We applied SVD and subtracted the first two components from the
834 matrix, which captured the sample batch effect. Although the two
835 components also captured major cell type differences, we anticipated that the
836 removal would not harm downstream analysis, which compares nuclei of the
837 same cell type between samples.
838
839 **Quantifying TF binding in snATAC-seq**
840 If the DNA sequence of a chromatin accessibility peak includes the motif of a
841 TF, Tn5 insertion in the peak suggests the binding of the corresponding TF.
842 For each nucleus, by aggregating over all peaks, we can infer the genome-
843 wide binding of the TF. This is obtained as the chromVAR (Schep et al., 2017)
844 deviation z-score in ArchR. The CIS-BP database of TF binding motifs

845 (Weirauch et al., 2014) was used. In the same way as we did for the Gene
846 Score Matrix, we applied quantile normalization to the Motif Matrix and
847 subtracted the first two components of SVD.
848

849 **Gene set enrichment analysis of differentially expressed genes under**
850 **HFD**

851 To seek which biological processes and/or pathways are differentially
852 expressed under a particular dietary intervention, we performed gene set
853 enrichment analysis. The target GS's were obtained from the Molecular
854 Signatures Database (MSigDB) (version 7.4) (Subramanian et al., 2005)
855 using the R package msigdbr (version 7.4.1). Among the datasets in MSigDB,
856 we used Gene Ontology for Biological Process, Hallmark gene sets, KEGG
857 pathways, Reactome pathways, and WikiPathways. Using the sample batch-
858 corrected Gene Score Matrix, we compared nuclei from a pair of dietary
859 conditions, for example, the hepatocytes from 4 weeks of HFD *vs.* the normal
860 diet. The statistical significance of gene set enrichment was tested using the
861 R package PADOG (version 1.34.0) (Tarca et al., 2012) and by performing
862 permutation for 10,000 times or 20 times the number of GS's in a dataset,
863 whichever was larger. Within each dataset, the GS's that attained a false
864 discover rate <0.05 were considered statistically significant. Significant GS's
865 were visually inspected using the EnrichmentMap App (version 3.3.3) of the
866 Cytoscape software (version 3.8.2).
867

868 **Discovering major modules of TF regulation and assigning biological**
869 **processes**

870 To elucidate a global picture of TF regulation, we inferred the regulatory
871 impacts of all TFs on all genes and then extracted major modules of the
872 regulatory relationship between TFs and genes. For each gene, we computed
873 its regulatory TFs using the R package GENIE3 (version 1.14.0) (Huynh-Thu

874 et al., 2010), which implements a random forest algorithm and thus can
875 model nonlinear regulation. The algorithm aimed to predict the expression
876 level of the gene based on the binding of TFs and dietary conditions. GENIE3
877 quantified the impacts of regulators by the proportion of explained variance.
878 By combining the results from all genes, we obtained a regulator-regulatee
879 matrix, in which the rows correspond to TFs and dietary conditions, the
880 columns correspond to genes, and an element represents the importance of
881 a TF to a gene.
882 Next, we applied the nonnegative matrix factorization implemented in
883 the R package NMF (version 0.23.0) (Gaujoux and Seoighe, 2010) to the
884 regulator-regulatee matrix. We used the Brunet algorithm (Brunet et al.,
885 2004) with 100 random starts. The number of factors was chosen based on
886 data as the inflection point of residual error (Hutchins et al., 2008). In the
887 output of NMF, the membership of TFs (regulators) and genes (regulatees) to
888 a factor was defined by nonnegative weight. We called the factors as
889 "modules" of TF regulation.
890 We discovered biological processes that characterize a module by
891 applying gene set enrichment analysis to genes belonging to the module. For
892 each module of TF regulation, we tested if heavily weighted genes were
893 enriched in some GS. For each GS (see above section), we computed the P-
894 value and *t*-statistics to test if genes in the given GS had a larger weight than
895 other genes. To account for multiple testing by the GS's, we performed 100
896 trials of shuffling gene names and computed family-wise error rates. The GS's
897 that attained a family-wise error rate of <0.05 were considered statistically
898 significant.
899
900 **GS activity score in single nuclei**
901 To infer the activity of a GS in single nuclei, we computed the "average" level
902 of TF binding and gene expression, with regards to the TFs/genes included in

903 the GS. Formally, we defined the weight of TFs/genes as one if it was included
904 in the GS and as zero otherwise. We then performed SVD for the TF binding
905 and gene expression matrix incorporating the weight. In the Motif Matrix (see
906 above), each TF (row) was normalized by subtracting the mean and being
907 divided by the standard deviation and then multiplied by the weight of the TF.
908 Similarly, in the Gene Score Matrix (see above), each gene (row) was
909 normalized by subtracting the mean and being divided by the standard
910 deviation and then multiplied by the weight of the gene. By stacking the two
911 matrices, we obtained a large matrix, in which the rows correspond to TFs
912 and genes and the columns correspond to nuclei. Finally, SVD was applied
913 and the first right-singular vector was defined as the GS activity score in
914 single nuclei.
915 Although the abovementioned activity score can always be computed,
916 it would be meaningless if the TFs/genes in the GS were not co-expressed.
917 Therefore, we conducted a statistical test to exclude such meaningless GS.
918 We checked whether the TFs/genes in the GS were mutually correlated more
919 than by chance. We computed the null distribution of the first singular value
920 by performing 100 trials of the SVD procedure but with the weight of TFs
921 permutated and the weight of genes permutated and verified whether the
922 actual first singular value could attain an empirical P-value of <0.05.
923 The principal component approach was adapted from the PLAGE
924 program (Tomfohr et al., 2005) that was developed to compute the pathway
925 activity score of a GS in bulk tissue specimens. Instead, we applied to nuclei.
926 Compared with the number of specimens, the large number of nuclei enabled
927 significance testing by permutation.
928

929 **REFERENCES**

930

931 Ajmera, V., Perito, E.R., Bass, N.M., Terrault, N.A., Yates, K.P., Gill, R., Loomba, R., Diehl, A.M.,
932 Aouizerat, B.E., and Network, for the N.C.R. (2017). Novel plasma biomarkers associated with
933 liver disease severity in adults with nonalcoholic fatty liver disease. Hepatology *65*, 65–77.
934 https://doi.org/10.1002/hep.28776.

935 An, X., Liu, J., Zheng, X., Dou, Z., Li, Y., Suo, Y., Ma, Y.N., Sun, M., Tian, Z.Y., and Xu, L.
936 (2019). Serum HGF ,PCIII and PLT are Noninvasive Markers for the diagnosis of
937 nonalcoholic fatty liver disease. https://doi.org/10.21203/rs.2.18035/v1.

938 Baglama, J., and Reichel, L. (2005). Augmented Implicitly Restarted Lanczos
939 Bidiagonalization Methods. Siam J Sci Comput *27*, 19–42.
940 https://doi.org/10.1137/04060593x.

941 Barrow, F., Khan, S., Fredrickson, G., Wang, H., Dietsche, K., Parthiban, P., Robert, S., Kaiser,
942 T., Winer, S., Herman, A., et al. (2021). Microbiota-Driven Activation of Intrahepatic B Cells
943 Aggravates NASH Through Innate and Adaptive Signaling. Hepatology Baltim Md *74*, 704–
944 722. https://doi.org/10.1002/hep.31755.

945 Behari, J., Yeh, T.-H., Krauland, L., Otruba, W., Cieply, B., Hauth, B., Apte, U., Wu, T., Evans,
946 R., and Monga, S.P.S. (2010). Liver-Specific β-Catenin Knockout Mice Exhibit Defective
947 Bile Acid and Cholesterol Homeostasis and Increased Susceptibility to Diet-Induced
948 Steatohepatitis. Am J Pathology *176*, 744–753. https://doi.org/10.2353/ajpath.2010.090667.

949 Berry, M.W., and Browne, M. (2005). Understanding Search Engines (SIAM).

950 Binder, J.X., Pletscher-Frankild, S., Tsafou, K., Stolte, C., O'Donoghue, S.I., Schneider, R.,
951 and Jensen, L.J. (2014). COMPARTMENTS: unification and visualization of protein
952 subcellular localization evidence. Database J Biological Databases Curation *2014*, bau012.
953 https://doi.org/10.1093/database/bau012.

954 Boujedidi, H., Robert, O., Bignon, A., Cassard-Doulcier, A.-M., Renoud, M.-L., Gary-Gouy, H.,
955 Hemon, P., Tharinger, H., Prévot, S., Bachelerie, F., et al. (2014). CXCR4 dysfunction in non-
956 alcoholic steatohepatitis in mice and patients. Clin Sci *128*, 257–267.
957 https://doi.org/10.1042/cs20130833.

958 Brunet, J.-P., Tamayo, P., Golub, T.R., and Mesirov, J.P. (2004). Metagenes and molecular
959 pattern discovery using matrix factorization. Proc National Acad Sci *101*, 4164-4169.
960 https://doi.org/10.1073/pnas.0308531101.

961 Cao, S., Yaqoob, U., Das, A., Shergill, U., Jagavelu, K., Huebert, R.C., Routray, C.,
962 Abdelmoneim, S., Vasdev, M., Leof, E., et al. (2010). Neuropilin-1 promotes cirrhosis of the
963 rodent and human liver by enhancing PDGF/TGF-β signaling in hepatic stellate cells. J Clin
964 Invest *120*, 2379-2394. https://doi.org/10.1172/jci41203.

965 Cepero-Donates, Y., Lacraz, G., Ghobadi, F., Rakotoarivelo, V., Orkhis, S., Mayhue, M., Chen,
966 Y.-G., Rola-Pleszczynski, M., Menendez, A., Ilangumaran, S., et al. (2016). Interleukin-15-
967 mediated inflammation promotes non-alcoholic fatty liver disease. Cytokine *82*, 102-111.
968 https://doi.org/10.1016/j.cyto.2016.01.020.

969 Chen, C., Li, H., Song, J., Zhang, C., Li, M., Mao, Y., Liu, A., and Du, J. (2022). Role of
970 Apolipoprotein A1 in PPAR Signaling Pathway for Nonalcoholic Fatty Liver Disease. Ppar
971 Res *2022*, 4709300. https://doi.org/10.1155/2022/4709300.

972 Chistiakov, D.A., Myasoedova, V.A., Revin, V.V., Orekhov, A.N., and Bobryshev, Y.V. (2018).
973 The impact of interferon-regulatory factors to macrophage differentiation and polarization
974 into M1 and M2. Immunobiology *223*, 101-111. https://doi.org/10.1016/j.imbio.2017.10.005.

975 Chu, A.L., Schilling, J.D., King, K.R., and Feldstein, A.E. (2020). The Power of Single Cell
976 Analysis for the Study of Liver Pathobiology. Hepatology hep.31485-23.
977 https://doi.org/10.1002/hep.31485.

978 Corces, M.R., Trevino, A.E., Hamilton, E.G., Greenside, P.G., Sinnott-Armstrong, N.A., Vesuna,
979 S., Satpathy, A.T., Rubin, A.J., Montine, K.S., Wu, B., et al. (2017). An improved ATAC-seq
980 protocol reduces background and enables interrogation of frozen tissues. Nature Methods
981 *14*, 959-962. https://doi.org/10.1038/nmeth.4396.

982 Cusanovich, D.A., Hill, A.J., Aghamirzaie, D., Daza, R.M., Pliner, H.A., Berletch, J.B., Filippova,
983 G.N., Huang, X., Christiansen, L., DeWitt, W.S., et al. (2018). A Single-Cell Atlas of
984 In Vivo Mammalian Chromatin Accessibility. Cell *174*, 1309-1324.e18.
985 https://doi.org/10.1016/j.cell.2018.06.052.

986 Date, D., Das, R., Narla, G., Simon, D.I., Jain, M.K., and Mahabeleshwar, G.H. (2014). Kruppel-
987 like Transcription Factor 6 Regulates Inflammatory Macrophage Polarization*. J Biol Chem
988 *289*, 10318-10329. https://doi.org/10.1074/jbc.m113.526749.

989 Friedman, S.L., Neuschwander-Tetri, B.A., Rinella, M., and Sanyal, A.J. (2018). Mechanisms of
990 NAFLD development and therapeutic strategies. Nat Med *24*, 908-922.
991 https://doi.org/10.1038/s41591-018-0104-9.

992 Gaujoux, R., and Seoighe, C. (2010). A flexible R package for nonnegative matrix
993 factorization. BMC Bioinformatics *11*, 367. https://doi.org/10.1186/1471-2105-11-367.

994 Goel, R., Boylan, B., Gruman, L., Newman, P.J., North, P.E., and Newman, D.K. (2007). The
995 proinflammatory phenotype of PECAM-1-deficient mice results in atherogenic diet-induced
996 steatohepatitis. Am J Physiol-Gastr L *293*, G1205-G1214.
997 https://doi.org/10.1152/ajpgi.00157.2007.

998 Granja, J.M., Corces, M.R., Pierce, S.E., Bagdatli, S.T., Choudhry, H., Chang, H.Y., and
999 Greenleaf, W.J. (2021). ArchR is a scalable software package for integrative single-cell
1000 chromatin accessibility analysis. Nature Genetics 1-31. https://doi.org/10.1038/s41588-
1001 021-00790-6.

1002 Halpern, K.B., Shenhav, R., Matcovitch-Natan, O., Tóth, B., Lemze, D., Golan, M., Massasa,
1003 E.E., Baydatch, S., Landen, S., Moor, A.E., et al. (2017). Single-cell spatial reconstruction
1004 reveals global division of labour in the mammalian liver. Nature *542*, 352-356.
1005 https://doi.org/10.1038/nature21065.

1006 Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J.,
1007 Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*,
1008 3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

1009 Herck, M.A.V., Vonghia, L., and Francque, S.M. (2017). Animal Models of Nonalcoholic Fatty
1010 Liver Disease—A Starter's Guide. Nutrients *9*, 1072. https://doi.org/10.3390/nu9101072.

1011 Hess, J., Angel, P., and Schorpp-Kistner, M. (2004). AP-1 subunits: quarrel and harmony
1012 among siblings. J Cell Sci *117*, 5965-5973. https://doi.org/10.1242/jcs.01589.

1013 Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J.I., Raj, A., Li, M.,
1014 and Zhang, N.R. (2018). SAVER: gene expression recovery for single-cell RNA sequencing.
1015 Nature Methods 1–11. https://doi.org/10.1038/s41592-018-0033-z.

1016 Hutchins, L.N., Murphy, S.M., Singh, P., and Graber, J.H. (2008). Position-dependent motif
1017 characterization using non-negative matrix factorization. Bioinformatics *24*, 2684–2690.
1018 https://doi.org/10.1093/bioinformatics/btn526.

1019 Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring Regulatory
1020 Networks from Expression Data Using Tree-Based Methods. Plos One *5*, e12776.
1021 https://doi.org/10.1371/journal.pone.0012776.

1022 Jew, B., Alvarez, M., Rahmani, E., Miao, Z., Ko, A., Garske, K.M., Sul, J.H., Pietiläinen, K.H.,
1023 Pajukanta, P., and Halperin, E. (2020). Accurate estimation of cell composition in bulk
1024 expression through robust integration of single-cell information. Nature Communications 1–
1025 11. https://doi.org/10.1038/s41467-020-15816-6.

1026 Kakino, S., Ohki, T., Nakayama, H., Yuan, X., Otabe, S., Hashinaga, T., Wada, N., Kurita, Y.,
1027 Tanaka, K., Hara, K., et al. (2017). Pivotal Role of TNF-? in the Development and Progression
1028 of Nonalcoholic Fatty Liver Disease in a Murine Model. Horm Metab Res *50*, 80–87.
1029 https://doi.org/10.1055/s-0043-118666.

1030 Katsuda, T., Hosaka, K., Matsuzaki, J., Usuba, W., Prieto-Vila, M., Yamaguchi, T., Tsuchiya, A.,
1031 Terai, S., and Ochiya, T. (2020). Transcriptomic Dissection of Hepatocyte Heterogeneity:
1032 Linking Ploidy, Zonation, and Stem/Progenitor Cell Characteristics. Cellular and Molecular
1033 Gastroenterology and Hepatology *9*, 161–183. https://doi.org/10.1016/j.jcmgh.2019.08.011.

1034 Kiki, I., Altunkaynak, B.Z., Altunkaynak, M.E., Vuraler, O., Unal, D., and Kaplan, S. (2007).
1035 Effect of High Fat Diet on the Volume of Liver and Quantitative Feature of Kupffer Cells in
1036 the Female Rat: A Stereological and Ultrastructural Study. Obes Surg *17*, 1381.
1037 https://doi.org/10.1007/s11695-007-9219-7.

1038 Koutsogiannis, D., Summers, K., George, B., Adams, P., Marotta, P., and Chakrabarti, S.
1039 (2010). Identification of Serum Biomarkers in End Stage Liver Disease. Open Biomarkers J
1040 *3*, 1–6. https://doi.org/10.2174/1875318301003010001.

1041 Krawczyk, M., Zimmermann, S., Hess, G., Holz, R., Dauer, M., Raedle, J., Lammert, F., and
1042 Grünhage, F. (2017). Panel of three novel serum markers predicts liver stiffness and fibrosis

1043 stages in patients with chronic liver disease. Plos One *12*, e0173506.
1044 https://doi.org/10.1371/journal.pone.0173506.

1045 Lareau, C.A., Ma, S., Duarte, F.M., and Buenrostro, J.D. (2020). Inference and effects of
1046 barcode multiplets in droplet-based single-cell assays. Nature Communications 1-9.
1047 https://doi.org/10.1038/s41467-020-14667-5.

1048 Li, H., Sharma, A., Luo, K., Qin, Z.S., Sun, X., and Liu, H. (2020). DeconPeaker, a
1049 Deconvolution Model to Identify Cell Types Based on Chromatin Accessibility in ATAC-Seq
1050 Data of Mixture Samples. Frontiers in Genetics *11*, 392.
1051 https://doi.org/10.3389/fgene.2020.00392.

1052 Li, Z.Z., Berk, M., McIntyre, T.M., and Feldstein, A.E. (2009). Hepatic Lipid Partitioning and
1053 Liver Damage in Nonalcoholic Fatty Liver Disease ROLE OF STEAROYL-CoA
1054 DESATURASE*. J Biol Chem *284*, 5637-5644. https://doi.org/10.1074/jbc.m807616200.

1055 Locatelli, I., Sutti, S., Vacchiano, M., Bozzola, C., and Albano, E. (2012). NF-κB1 deficiency
1056 stimulates the progression of non-alcoholic steatohepatitis (NASH) in mice by promoting
1057 NKT-cell-mediated responses. Clin Sci *124*, 279-287. https://doi.org/10.1042/cs20120289.

1058 Orchard, P., Kyono, Y., Hensley, J., Kitzman, J.O., and Parker, S.C.J. (2020). Quantification,
1059 Dynamic Visualization, and Validation of Bias in ATAC-Seq Data with ataqv. Cell Systems
1060 *10*, 298-306.e4. https://doi.org/10.1016/j.cels.2020.02.009.

1061 Page, S., Birerdinc, A., Estep, M., Stepanova, M., Afendy, A., Petricoin, E., Younossi, Z.,
1062 Chandhoke, V., and Baranova, A. (2013). Knowledge-Based Identification of Soluble
1063 Biomarkers: Hepatic Fibrosis in NAFLD as an Example. Plos One *8*, e56009.
1064 https://doi.org/10.1371/journal.pone.0056009.

1065 Patouraux, S., Rousseau, D., Bonnafous, S., Lebeaupin, C., Luci, C., Canivet, C.M., Schneck,
1066 A.-S., Bertola, A., Saint-Paul, M.-C., Iannelli, A., et al. (2017). CD44 is a key player in non-
1067 alcoholic steatohepatitis. J Hepatol *67*, 328-338. https://doi.org/10.1016/j.jhep.2017.03.003.

1068 Pérez-Martínez, L., Pérez-Matute, P., Aguilera-Lizarraga, J., Rubio-Mediavilla, S., Narro, J.,
1069 Recio, E., Ochoa-Callejero, L., Oteo, J.-A., and Blanco, J.-R. (2014). Maraviroc, a CCR5
1070 antagonist, ameliorates the development of hepatic steatosis in a mouse model of non-
1071 alcoholic fatty liver disease (NAFLD). J Antimicrob Chemoth *69*, 1903-1910.
1072 https://doi.org/10.1093/jac/dku071.

1073 Potoupni, V., Georgiadou, M., Chatzigriva, E., Polychronidou, G., Markou, E., Gakis, C.Z.,
1074 Filimidou, I., Karagianni, M., Anastasilakis, D., Evripidou, K., et al. (2021). Circulating tumor
1075 necrosis factor-α levels in non-alcoholic fatty liver disease: A systematic review and a
1076 meta-analysis. J Gastroen Hepatol *36*, 3002–3014. https://doi.org/10.1111/jgh.15631.

1077 Rasala, B.A., Orjalo, A.V., Shen, Z., Briggs, S., and Forbes, D.J. (2006). ELYS is a dual
1078 nucleoporin/kinetochore protein required for nuclear pore assembly and proper cell division.
1079 Proc National Acad Sci *103*, 17801–17806. https://doi.org/10.1073/pnas.0608484103.

1080 Reske, J.J., Wilson, M.R., and Chandler, R.L. (2020). ATAC-seq normalization method can
1081 significantly affect differential accessibility analysis and interpretation. Epigenetics &
1082 Chromatin 1–17. https://doi.org/10.1186/s13072-020-00342-y.

1083 Rigotti, F.D.A., Wiedmann, L., Hubert, M.O., Vacca, M., Hasan, S.S., Moll, I., Carvajal, S.,
1084 Jiménez, W., Starostecka, M., Billeter, A.T., et al. (2022). Semaphorin 3C exacerbates liver
1085 fibrosis. Biorxiv 2021.07.29.454292. https://doi.org/10.1101/2021.07.29.454292.

1086 Roos, B. de, Rungapamestry, V., Ross, K., Rucklidge, G., Reid, M., Duncan, G., Horgan, G.,
1087 Toomey, S., Browne, J., Loscher, C.E., et al. (2009). Attenuation of inflammation and cellular
1088 stress-related pathways maintains insulin sensitivity in obese type I interleukin-1 receptor
1089 knockout mice on a high-fat diet. Proteomics *9*, 3244–3256.
1090 https://doi.org/10.1002/pmic.200800761.

1091 Salcedo, R., Wasserman, K., Young, H.A., Grimm, M.C., Howard, O.M.Z., Anver, M.R., Kleinman,
1092 H.K., Murphy, W.J., and Oppenheim, J.J. (1999). Vascular Endothelial Growth Factor and
1093 Basic Fibroblast Growth Factor Induce Expression of CXCR4 on Human Endothelial Cells In
1094 Vivo Neovascularization Induced by Stromal-Derived Factor-1α. Am J Pathology *154*,
1095 1125–1135. https://doi.org/10.1016/s0002-9440(10)65365-5.

1096 Sawitza, I., Kordes, C., Reister, S., and Häussinger, D. (2009). The niche of stellate cells
1097 within rat liver. Hepatology *50*, 1617–1624. https://doi.org/10.1002/hep.23184.

1098 Schep, A.N., Wu, B., Buenrostro, J.D., and Greenleaf, W.J. (2017). chromVAR: inferring
1099 transcription-factor-associated accessibility from single-cell epigenomic data. Nat Methods
1100 *14*, 975–978. https://doi.org/10.1038/nmeth.4401.

1101 Seo, Y.Y., Cho, Y.K., Bae, J.-C., Seo, M.H., Park, S.E., Rhee, E.-J., Park, C.-Y., Oh, K.-W.,
1102 Park, S.-W., and Lee, W.-Y. (2013). Tumor Necrosis Factor-α as a Predictor for the

1103 Development of Nonalcoholic Fatty Liver Disease: A 4-Year Follow-Up Study. Endocrinol
1104 Metab *28*, 41-45. https://doi.org/10.3803/enm.2013.28.1.41.

1105 Sturm, G., Finotello, F., Petitprez, F., Zhang, J.D., Baumbach, J., Fridman, W.H., List, M., and
1106 Aneichyk, T. (2019). Comprehensive evaluation of transcriptome-based cell-type
1107 quantification methods for immuno-oncology. Bioinformatics (Oxford, England) *35*, i436-i445.
1108 https://doi.org/10.1093/bioinformatics/btz363.

1109 Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A.,
1110 Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment
1111 analysis: A knowledge-based approach for interpreting genome-wide expression profiles.
1112 Proceedings of the National Academy of Sciences of the United States of America *102*,
1113 15545-15550. https://doi.org/10.1073/pnas.0506580102.

1114 Sun, J., and Weis, W.I. (2011). Biochemical and Structural Characterization of β-Catenin
1115 Interactions with Nonphosphorylated and CK2-Phosphorylated Lef-1. J Mol Biol *405*, 519-
1116 530. https://doi.org/10.1016/j.jmb.2010.11.010.

1117 Tarca, A.L., Draghici, S., Bhatti, G., and Romero, R. (2012). Down-weighting overlapping
1118 genes improves gene set analysis. Bmc Bioinformatics *13*, 136.
1119 https://doi.org/10.1186/1471-2105-13-136.

1120 Teschendorff, A.E., and Zheng, S.C. (2017). Cell-type deconvolution in epigenome-wide
1121 association studies: a review and recommendations. Epigenomics *9*, 757-768.
1122 https://doi.org/10.2217/epi-2016-0153.

1123 Teschendorff, A.E., Breeze, C.E., Zheng, S.C., and Beck, S. (2017). A comparison of
1124 reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide
1125 Association Studies. BMC Bioinformatics *18*, 105. https://doi.org/10.1186/s12859-017-
1126 1511-5.

1127 Tomfohr, J., Lu, J., and Kepler, T.B. (2005). Pathway level analysis of gene expression using
1128 singular value decomposition. Bmc Bioinformatics *6*, 225-225.
1129 https://doi.org/10.1186/1471-2105-6-225.

1130 Traag, V.A., Waltman, L., and Eck, N.J. van (2019). From Louvain to Leiden: guaranteeing
1131 well-connected communities. Sci Rep-Uk *9*, 5233. https://doi.org/10.1038/s41598-019-
1132 41695-z.

1133 Tran, M.T.N., Hamada, M., Jeon, H., Shiraishi, R., Asano, K., Hattori, M., Nakamura, M.,
1134 Imamura, Y., Tsunakawa, Y., Fujii, R., et al. (2017). MafB is a critical regulator of complement
1135 component C1q. Nat Commun *8*, 1700. https://doi.org/10.1038/s41467-017-01711-0.

1136 Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P.,
1137 Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and Inference
1138 of Eukaryotic Transcription Factor Sequence Specificity. Cell *158*, 1431-1443.
1139 https://doi.org/10.1016/j.cell.2014.08.009.

1140 Wilson, D.R., Jin, C., Ibrahim, J.G., and Sun, W. (2019). ICeD-T Provides Accurate Estimates
1141 of Immune Cell Abundance in Tumor Samples by Allowing for Aberrant Gene Expression
1142 Patterns. J Am Stat Assoc *115*, 1-23. https://doi.org/10.1080/01621459.2019.1654874.

1143 Yagai, T., Miyajima, A., and Tanaka, M. (2014). Semaphorin 3E Secreted by Damaged
1144 Hepatocytes Regulates the Sinusoidal Regeneration and Liver Fibrosis during Liver
1145 Regeneration. Am J Pathology *184*, 2250-2259. https://doi.org/10.1016/j.ajpath.2014.04.018.

1146 Yang, M.H., Sung, J., and Gwak, G.-Y. (2016). The associations between apolipoprotein B, A1,
1147 and the B/A1 ratio and nonalcoholic fatty liver disease in both normal-weight and
1148 overweight Korean population. J Clin Lipidol *10*, 289-298.
1149 https://doi.org/10.1016/j.jacl.2015.11.017.

1150 Yip, S.H., Wang, P., Kocher, J.-P.A., Sham, P.C., and Wang, J. (2017). Linnorm: improved
1151 statistical analysis for single cell RNA-seq expression data. Nucleic Acids Research *45*,
1152 e179. https://doi.org/10.1093/nar/gkx828.

1153 Yu, X., Chen, X., Zheng, X.D., Zhang, J., Zhao, X., Liu, Y., Zhang, H., Zhang, L., Yu, H., Zhang,
1154 M., et al. (2018). Growth Differentiation Factor 11 Promotes Abnormal Proliferation and
1155 Angiogenesis of Pulmonary Artery Endothelial Cells. Hypertension *71*, 729-741.
1156 https://doi.org/10.1161/hypertensionaha.117.10350.

1157 Zhang, C., Jia, Y., Liu, B., Wang, G., and Zhang, Y. (2020a). TLR4 knockout upregulates the
1158 expression of Mfn2 and PGC-1α in a high-fat diet and ischemia-reperfusion mice model of
1159 liver injury. Life Sci *254*, 117762. https://doi.org/10.1016/j.lfs.2020.117762.

1160 Zhang, F., Jiang, W.W., Li, X., Qiu, X.Y., Wu, Z., Chi, Y.J., Cong, X., and Liu, Y.L. (2016). Role of
1161 intrahepatic B cells in non-alcoholic fatty liver disease by secreting pro-inflammatory

1162 cytokines and regulating intrahepatic T cells. J Digest Dis *17*, 464–474.
1163 https://doi.org/10.1111/1751-2980.12362.

1164 Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C.,
1165 Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS).
1166 Genome Biol *9*, R137–R137. https://doi.org/10.1186/gb-2008-9-9-r137.

1167 Zhang, Y., Lu, Z., Zeng, W., Zhao, J., and Zhou, X. (2020b). Two sides of NNMT in alcoholic
1168 and nonalcoholic fatty liver development. J Hepatol *74*, 1250–1253.
1169 https://doi.org/10.1016/j.jhep.2020.11.049.

1170 Zhong, L., Huang, L., Xue, Q., Liu, C., Xu, K., Shen, W., and Deng, L. (2019). Cell-specific
1171 elevation of Runx2 promotes hepatic infiltration of macrophages by upregulating MCP-1 in
1172 high-fat diet-induced mice NAFLD. J Cell Biochem *120*, 11761–11774.
1173 https://doi.org/10.1002/jcb.28456.

1174 Zinzalla, G. (2016). A New Way Forward in Cancer Drug Discovery: Inhibiting the SWI/SNF
1175 Chromatin Remodelling Complex. Chembiochem *17*, 677–682.
1176 https://doi.org/10.1002/cbic.201500565.

1177

Figure 1

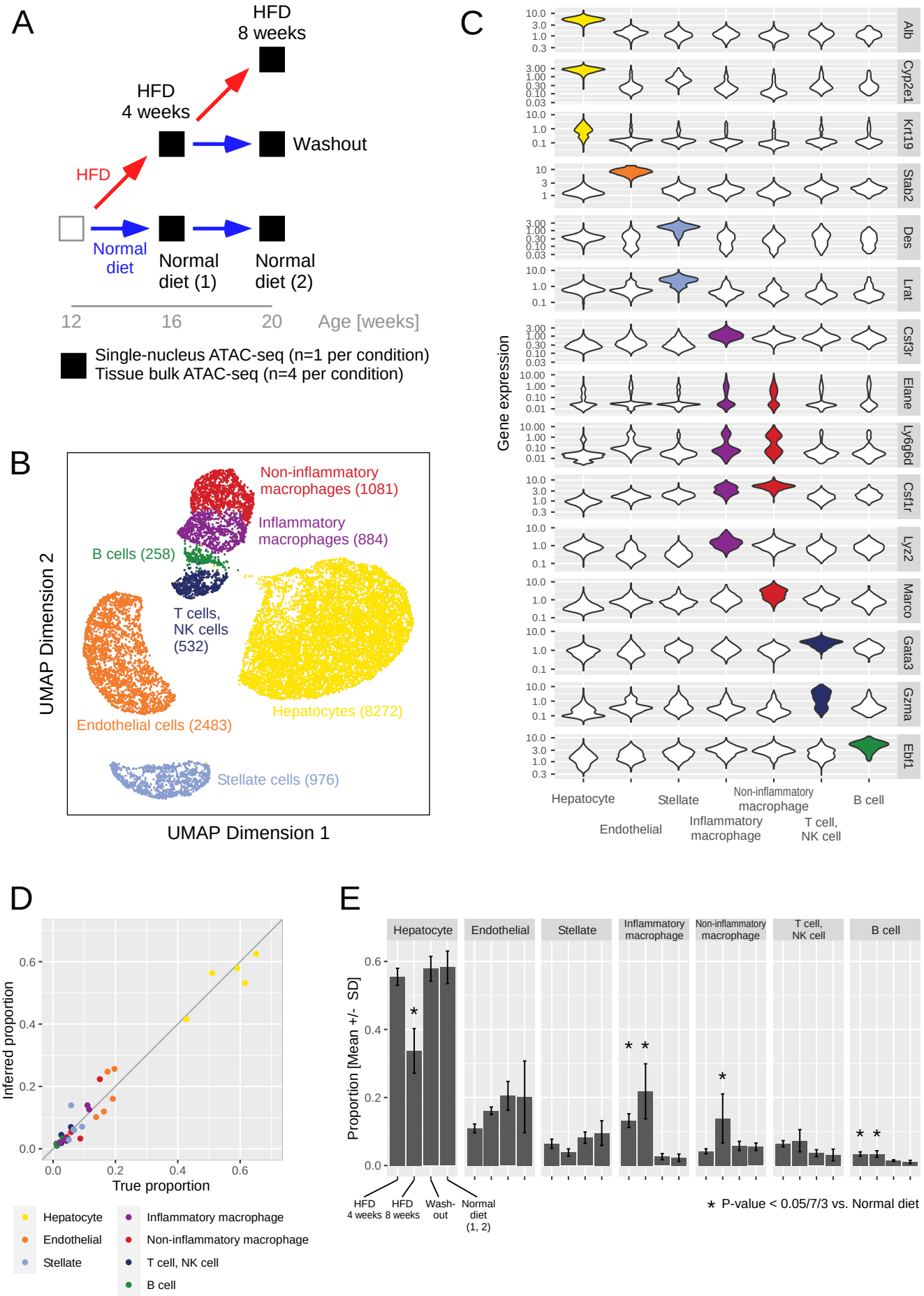

A
HFD
8 weeks
HFD
4 weeks
Washout
HFD
Normal
diet
Normal
diet (1)
Normal
diet (2)
12
16
20
Age [weeks]
Single-nucleus ATAC-seq (n=1 per condition)
Tissue bulk ATAC-seq (n=4 per condition)
B
Non-inflammatory
macrophages (1081)
Inflammatory
macrophages (884)
B cells (258)
T cells,
NK cells
(532)
Hepatocytes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

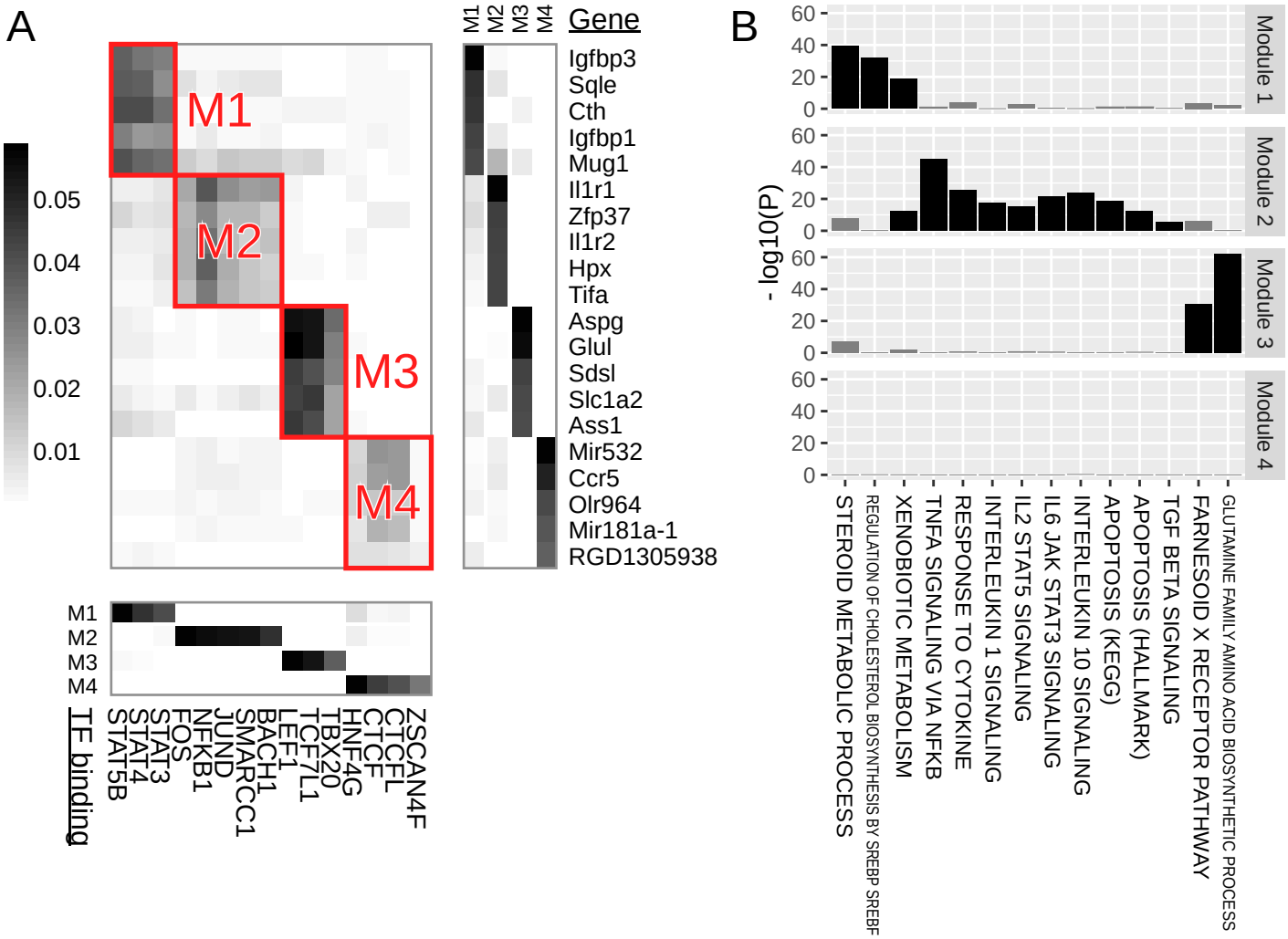


Figure 4. Endothelial

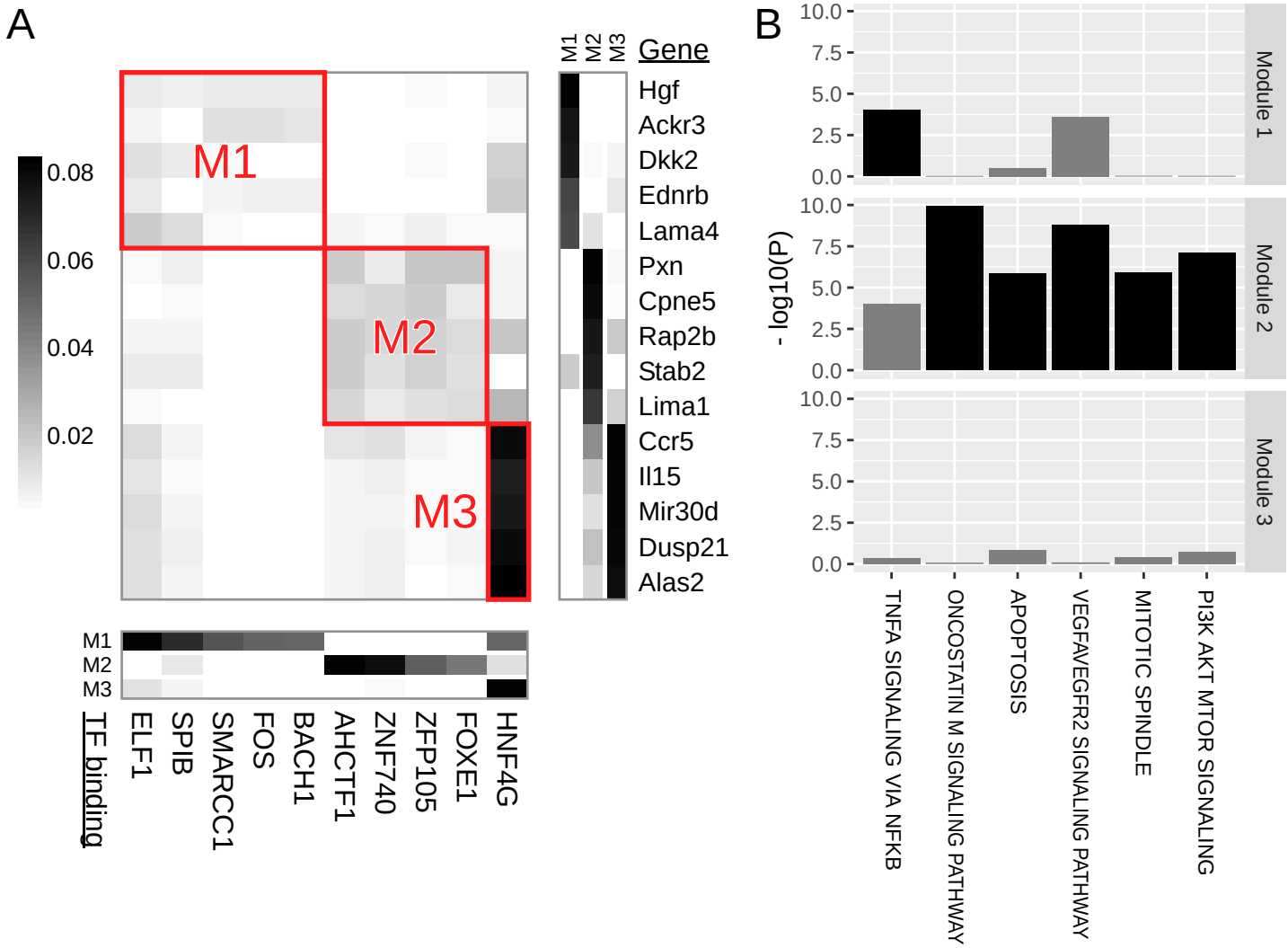


Figure 5. Stellate

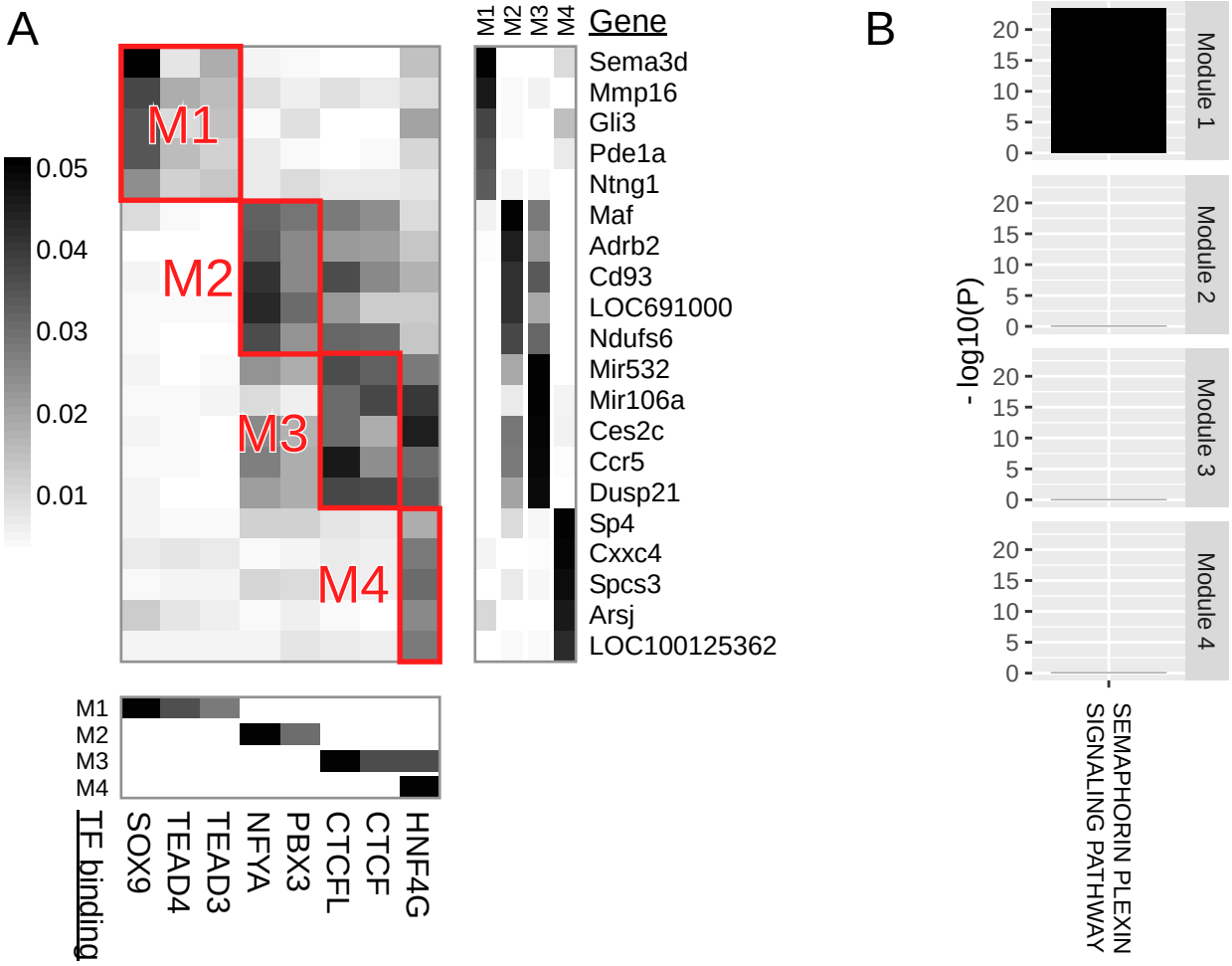


Figure 6. Macrophage

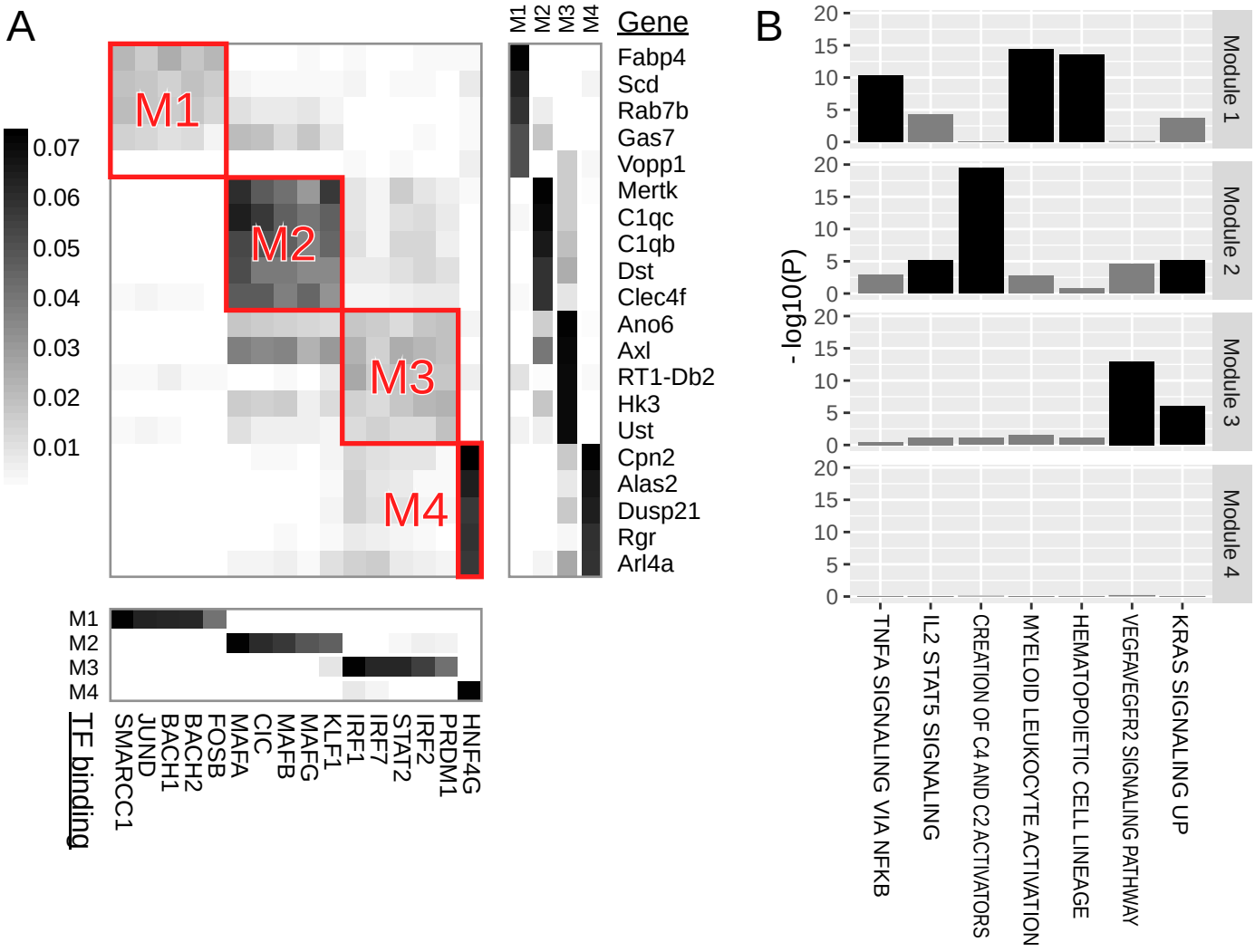
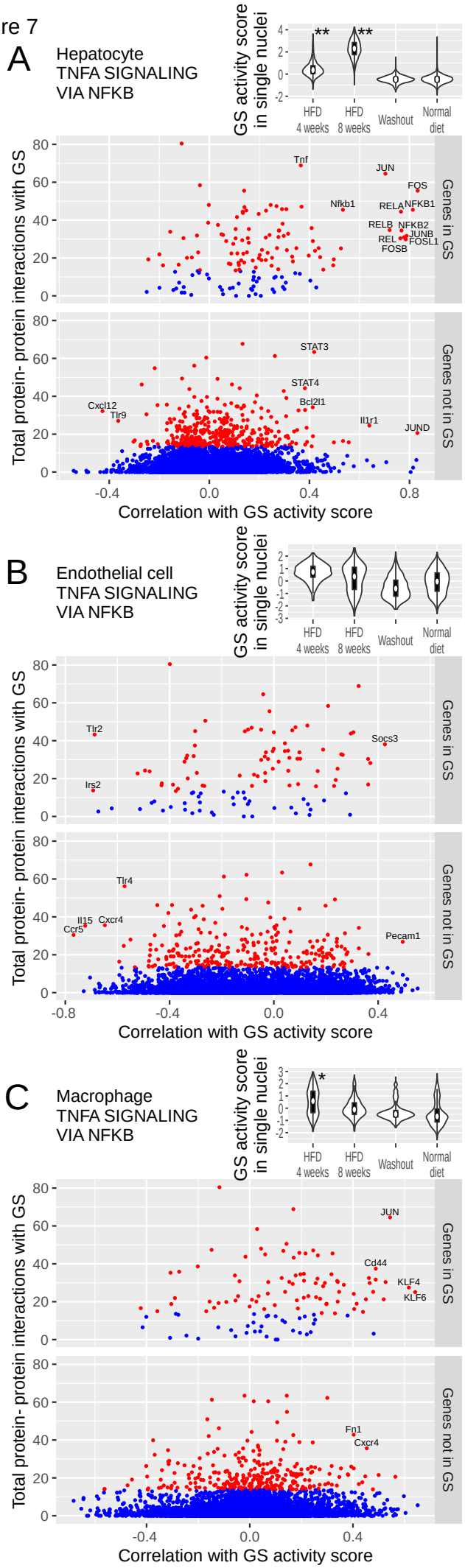


Figure 7


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