Gene regulation modules in non-alcoholic fatty liver disease revealed by single-nucleus ATAC-seq

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, Tokyo, Japan.

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

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

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

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

* Correspondence: fumihiko@takeuchi.name

Running title: Gene regulation modules in NAFLD

Keywords: Gene regulation, non-alcoholic fatty liver disease, single-nucleus ATAC-seq, transcription factor

Summary blurb

This study uncovers cell type-specific global gene regulation in NAFLD progression, identifying core genes and shedding light on the disease's molecular mechanisms.

Abstract

We investigated the progression of non-alcoholic fatty liver disease (NAFLD) from fatty liver to steatohepatitis using single-nucleus and bulk ATAC-seq on the livers of rats fed a high-fat diet (HFD). Rats fed HFD for 4 weeks developed fatty liver, and those fed HFD for 8 weeks further progressed to steatohepatitis. We observed an increase in the proportion of inflammatory macrophages, consistent with the pathological progression. Utilizing machine learning, we divided global gene regulation into modules, wherein transcription factors within a module could regulate genes within the same module, reaffirming known regulatory relationships between transcription factors and biological processes. We identified core genes—central to co-expression and protein-protein interaction—for the biological processes discovered. Notably, a large part of the core genes overlapped with genes previously implicated in NAFLD. Single-nucleus ATAC-seq, combined with data-driven statistical analysis, offers insight into *in vivo* global gene regulation as a combination of modules and assists in identifying core genes of relevant biological processes.

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), with a global prevalence of ~25% (Friedman et al, 2018). NAFLD can gradually develop from fatty liver to steatohepatitis, cirrhosis, and eventually, hepatic cancer. There are currently no approved drugs for NAFLD, and existing biomarkers are inadequate. The liver is composed of multiple cell types, including hepatocytes, cholangiocytes, stellate cells, endothelial cells, and white blood cells, each with a distinct role in disease development. Recent single-cell RNA-seq studies have revealed cell type-specific gene expression changes in NAFLD (Ramachandran et al, 2019; Xiong et al, 2019; Chu et al, 2020).

A key upstream event in gene expression regulation is the binding of transcription factors (TFs) to DNA. ChIP-seq/Cut&Run/Cut&Tag assays can be performed on single cells but they can only measure the binding of one type of TF at a time (Salma et al, 2023). In contrast, single-nucleus ATAC-seq (snATAC-seq) detects open chromatin sites, where TFs with DNA binding motifs that match the site sequence are presumed to bind, thereby providing information on the binding of all types of TFs simultaneously. Using snATAC-seq data, recent algorithms have inferred gene regulation by TFs, forming a gene regulatory network (GRN) for individual cell types (Fleck et al, 2022; González-Blas et al, 2022; Jiang et al, 2022; Kamal et al, 2022; Kartha et al, 2022; Wang et al, 2022; Kamimoto et al, 2023). Unlike previous studies which focused on identifying the wiring of the GRN, our aim is to partition the entire GRN and extract major components that correspond to biological processes occurring in the relevant cells, such as lipid metabolism or inflammation in NAFLD. The binding of TFs to chromatin could not be analyzed in previous single-cell analyses of NAFLD, as they were all based on RNA-seq and not ATAC-seq.

77 The cell type composition of tissues can change during disease
78 development in various diseases and will impact the outcome. While cell type
79 counting of solid tissue has become feasible through single-cell assays, these
80 assays are still expensive. Instead, deconvolution, which is the statistical
81 inference of the cell type composition of tissue samples assayed in bulk, has
82 been utilized for transcriptome and DNA methylation assays (Teschendorff &
83 Zheng, 2017; Sturm et al, 2019). For ATAC-seq, Li and associates (Li et al,
84 2020) previously proposed an algorithm for deconvolution, but no
85 experimental validation with single-cell data has been conducted.
86 Aiming to understand the cell type composition and global TF
87 regulation in NAFLD development, we performed single-nucleus and bulk
88 ATAC-seq on the liver of rats fed or withdrawn from a high-fat diet (HFD). We
89 first analyzed the cell type composition in the liver and the gene expression
90 separately by cell type. Then, to investigate global TF regulation in each cell
91 type, we developed a novel statistical method that decomposes TF regulation
92 into major modules, each being characterized by biological processes. We
93 also developed a statistical method to detect core genes for the individual
94 biological processes. Taking snATAC-seq data from an *in vivo* experimental
95 model as input, our data-driven reductionist workflow proceeds from global
96 TF regulation to TF regulation modules, then to biological processes, and
97 finally to core genes. The core genes thus identified significantly overlap with
98 previously known NAFLD genes, thereby providing experimental validation for
99 our newly developed statistical methods.
100
101 **Results**
102 **Cell type composition of the liver**
103 Spontaneously Hypertensive Rats (SHR) fed a HFD for 4 weeks developed
104 fatty liver, and those maintained on HFD for 8 weeks further developed
105 inflammation and premature fibrosis (steatohepatitis) (Figure 1A; Table S1).

106 Under the washout condition, where 4 weeks of HFD is followed by 4 weeks
107 of a normal diet, fatty liver was partially ameliorated. Along with control rats
108 fed only a normal diet, five dietary conditions were examined.

109 In order to catalog cell types in the liver, we performed snATAC-seq
110 on one animal per dietary condition, totaling five animals. We obtained a total
111 of 14,486 nuclei (~3000 per sample), which were grouped into 16 clusters
112 based on the similarity of chromatin opening. Hepatocytes (seven clusters),
113 endothelial cells (three clusters), stellate cells (two clusters) and white blood
114 cells (four clusters) were clearly separated (Figures 1B and 1C). The clusters
115 for white blood cells corresponded to inflammatory macrophages, non-
116 inflammatory macrophages, T cells and NK cells combined, and B cells.
117 Cholangiocytes were included in the hepatocyte clusters, while neutrophils
118 and Kupffer cells were grouped into the macrophage clusters.

119 To measure the changes in cell type population through dietary
120 intervention, we performed bulk ATAC-seq on four animals per condition,
121 totaling 20 animals. Using snATAC-seq as reference, we inferred the cell type
122 composition in the bulk ATAC-seq samples by deconvolution. Beforehand,
123 using the five samples assayed in both single-nucleus and bulk ATAC-seq, the
124 accuracy of deconvolution for ATAC-seq was validated (Figure 1D). The
125 difference between the inferred proportion and the true proportion counted
126 by snATAC-seq had a mean of zero and a standard deviation of 0.035,
127 indicating high accuracy. Considering that we used four animals per condition,
128 the standard error was calculated as 0.018. Compared to the normal diet
129 condition, hepatocytes significantly decreased after 8 weeks of HFD (Figure
130 1E). Inflammatory macrophages largely increased after 4 weeks of HFD and
131 further increased after 8 weeks of HFD. Non-inflammatory macrophages also
132 increased after 8 weeks of HFD. Additionally, B cells increased under HFD
133 conditions. Consistent with our findings, previous studies on HFD-induced
134 NAFLD animal models reported an increase in liver-resident macrophages

135 (Kiki et al, 2007), recruitment of monocyte-derived macrophages (Zhong et
136 al, 2019), and an increase in intrahepatic B cells (Zhang et al, 2016; Barrow
137 et al, 2021). Under the washout condition, the cell type composition reverted
138 to the same as the normal diet condition. Having elucidated how the quantity
139 of cell types changed through dietary intervention, we next investigate the
140 qualitative change.
141
142 **Differential gene expression under HFD in each cell type**
143 We analyzed cell type-specific differential gene expression by comparing
144 nuclei of the same cell type across different dietary conditions (Table S2;
145 Dataset S1). Hereafter, clusters for inflammatory and non-inflammatory
146 macrophages are combined and referred to as the macrophage cell type. T
147 cells, NK cells, and B cells were not included in the analysis due to the limited
148 number of nuclei. The expression level of a gene in a nucleus was quantified
149 as a score based on chromatin opening in the gene body and its vicinity. The
150 biological processes characteristic of the differentially expressed genes were
151 cell type-dependent (Figure 2). Steroid metabolism and fatty acid metabolism
152 genes were expressed in hepatocytes differently under HFD and washout
153 compared with the normal diet. As for inflammation genes, their expression
154 scores altered in macrophages after 4 weeks of HFD and washout.
155 Inflammation genes and apoptosis genes were expressed in hepatocytes
156 differently after 8 weeks of HFD. Genes related to the actin filament-based
157 process were expressed in stellate cells differently after 8 weeks of HFD.
158 Conventional differential gene expression analysis using bulk tissue
159 samples has limitations, including the difficulty of detecting 1) changes in
160 minor cell types, 2) changes in a cell type with low basal gene expression
161 relative to other cell types, and 3) concurrent changes in gene expression
162 and cell population that mutually cancel out, for example, downregulated
163 genes in a cell type that is increasing in population. We cross-checked cell

164 type-specific differential gene expression with bulk tissue transcriptome
165 analysis for the comparison between the 4 weeks of HFD and the normal diet.
166 We found the results to replicate under the abovementioned limitations
167 (Figure S1). Differential gene expression was replicated in hepatocytes, but
168 to a lesser degree in minor cell types. For macrophages, whose proportion
169 increased after 4 weeks of HFD, the upregulated genes were replicated in
170 bulk experiments but the downregulated genes were not.
171
172 **Major modules of TF regulation and associated biological processes**
173 To elucidate the global picture of TF regulation, we inferred the regulatory
174 impact of all TFs on all genes and extracted major modules from these
175 regulatory relationships. The genome-wide binding of a TF in a nucleus was
176 quantified as a score based on the enrichment of the TF's motif in the open
177 chromatin. Utilizing snATAC-seq, we quantified in all nuclei of each cell type
178 (from rats under all dietary conditions) the genome-wide binding of TFs and
179 the expression of genes, and inferred the regulation of the latter by the
180 former by applying machine learning. The inferred impact of a TF on a gene
181 was represented as a fractional number, corresponding to the weight of a GRN
182 edge connecting the TF to the gene.
183 The regulatory relationships were subsequently divided into modules.
184 A module comprises a subset of TFs and a subset of genes, such that TFs
185 within a module regulate genes within the same module but do not regulate
186 genes belonging to other modules. In hepatocytes, we identified four modules
187 (Figure 3A). By themselves, modules are simply a grouping of TFs/genes
188 detected by machine learning and carry no inherent biological meaning on
189 their own. We applied gene set enrichment analysis to the genes within a
190 module, without using the information on TFs, to identify biological processes
191 that characterize each module (Figure 3B). In the following, we describe the
192 modules for hepatocytes, endothelial cells (Figure 4), stellate cells (Figure 5),

193 and macrophages (Figure 6). If a biological process was assigned to a module,
194 we sought literature that supports the regulation of the biological process by
195 the TFs within the module. The discovered linkages between the TFs and
196 biological processes found support in the literature for five of the seven
197 identified modules, suggesting the validity of our module discovery algorithm.
198 Module 1 of hepatocytes was characterized by the binding of STAT
199 family TFs (namely, STAT5B, STAT4 and STAT3) and by genes for steroid
200 metabolism.
201 Module 2 of hepatocytes, Module 1 of endothelial cells, and Module 1
202 of macrophages were characterized by the binding of AP-1 family TFs (namely,
203 FOS, JUND, BACH1 and BACH2), SMARCC1 and NFKB1, along with genes for
204 TNFα signaling via NF-κB. AP-1 TFs respond to cytokine stimuli (Hess et al,
205 2004). SMARCC1 is a member of the SWI/SNF family of proteins, which
206 function as a helicase and interacts with NF-κB (Zinzalla, 2016).
207 Module 3 of hepatocytes was characterized by the binding of TCF/LEF
208 family TFs (namely, LEF1 and TCF7L1) and by genes in the farnesoid X
209 receptor pathway for bile acid synthesis and genes involved in glutamine
210 family amino acid biosynthesis. The liver lobule, a unit constituting the liver,
211 houses a central vein in its core and a portal triad (portal vein, hepatic artery,
212 and bile duct) in its periphery, with hepatocyte metabolism varying between
213 pericentral and periportal zones. The glutamine family amino acid
214 biosynthesis genes include *Glul* and *Ass1*, which are highly expressed in the
215 pericentral and periportal zones, respectively (Halpern et al, 2017). Moreover,
216 LEF1 TF binds to β-catenin protein, activating the Wnt signaling pathway (Sun
217 & Weis, 2011), and liver-specific β-catenin KO mice exhibited defective
218 cholesterol and bile acid metabolism in the liver (Behari et al, 2010). Wnt-
219 activated genes are also highly expressed in the pericentral zone, suggesting
220 LEF1 TF's role in regulating bile acid metabolism and zonated gene expression
221 in the liver.

222 Module 2 of endothelial cells was characterized by the DNA binding of
223 AHCTF1 (aka ELYS) and ZNF740 and by genes for angiogenesis. AHCTF1 is a
224 DNA-binding protein necessary for mitosis (Rasala et al, 2006). ZNF740
225 activates angiogenesis in rat pulmonary artery endothelial cells (Yu et al,
226 2018).
227 Module 1 of stellate cells was characterized by SOX9 TF binding and
228 genes involved in the semaphorin-plexin signaling pathway.
229 Module 2 of macrophages was characterized by the binding of Maf
230 family TFs (namely, MAFA, MAFB and MAFG) and by genes for complement
231 system activation. In *Mafb*-deficient macrophages of mice, C1q production
232 decreased (Tran et al, 2017).
233 Module 3 of macrophages was characterized by the binding of IRF
234 family TFs (namely, IRF1, IRF2 and IRF7), which regulate immune responses,
235 and by genes associated with angiogenesis. IRF1 contributes to the
236 commitment of pro-inflammatory M1 macrophages, which produce
237 angiogenic stimulators (Chistiakov et al, 2018).
238
239 **Core genes in specific biological processes**
240 In the previous section, we discovered major modules of TF regulation and
241 searched databases, such as Gene Ontology, for biological processes
242 characterizing these modules. We defined a biological process by its
243 constituent genes, terming them as a gene set (GS). A GS characterizes a
244 module if it overlaps significantly with the genes within that module. Here,
245 we aim to find "core genes" of a biological process. In preparation, we
246 compute the activity of a GS in individual nuclei as the "GS activity score."
247 The score is obtained by averaging the binding/expression scores of the
248 TFs/genes in the set, or, more precisely, by computing the first principal
249 component. In hepatocytes, regarding the biological process of TNFα
250 signaling via NF-κB, the GS activity score was the highest in nuclei from 8

251 weeks of HFD (Figure 7A, top panel).

252 Core genes are those central to the GS in terms of two connections:
253 co-expression and protein-protein interaction. For co-expression, we first
254 compute the correlation coefficient between the GS activity score and the
255 expression score of a gene (or the binding score of a TF). A core gene is
256 required to have a high absolute value of this correlation coefficient.
257 Furthermore, the total amount of protein-protein interactions between the
258 gene and the GS should be high. Consequently, by plotting the correlation
259 coefficient on the horizontal axis and the total protein-protein interactions on
260 the vertical axis, core genes are expected to appear in either the top left or
261 top right corner. This approach can be used to prioritize known genes in the
262 target GS (Figure 7A, middle panel), and discover novel genes not included
263 in the GS (Figure 7A, bottom panel). In the following, we will explore some
264 of the biological processes identified in the previous section.

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

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

275 In macrophages, the GS activity score for TNFα signaling via NF-κB
276 (Figure 7C) was higher after 4 weeks of HFD. TFs that had positively
277 correlated binding scores and abundant protein-protein interactions included
278 JUN and KLF6. The positively correlated core genes were *Cd44*, *Fn1* and *Cxcr4*.
279 KLF6 regulates the switching of macrophages between pro- and anti-

280 inflammatory states (Date et al, 2014). In accordance, *Lyz2* and *Marco*,
281 marker genes for inflammatory and non-inflammatory macrophages,
282 respectively, showed a correlation of 0.39 and −0.49, with KLF6 binding in
283 our single-nucleus data.

284 In hepatocytes, the GS activity score for steroid metabolism (Figure
285 S2A) was the highest after 8 weeks of HFD and to a lesser degree after 4
286 weeks of HFD. The positively correlated core genes included *Abcg1*, *Abca1*
287 and *Scd*. The negatively correlated core genes included *Sqle*, *Hmgcr*, *Apoa1*,
288 *Sult2a1*, *Acox2* and *Alb*. Among the 15 cytochrome P450 genes showing a
289 strong negative correlation (correlation less than −0.3 and protein-protein
290 interactions within the top 5%), 13 were oxidoreductases of EC 1.14.14.1.
291 For reference, we labeled the binding of steroid sensing TFs, NR1H2/3 (aka
292 LXRα/β), SREBF1 and SREBF2, which were negatively correlated with the GS
293 expression.

294 To investigate lobular zonation genes in hepatocytes, the GS was
295 taken from the list of zonation genes in normal mice (q-value < 10^{-6} in Table
296 S3 in (Halpern et al, 2017)). The GS activity score showed a modest
297 difference between dietary conditions (Figure S2B), aligning with the concept
298 that zonation is a normal phenomenon unrelated to NAFLD. When classified
299 by the correlation with the GS activity score, the genes with a correlation
300 coefficient >0.3 were all pericentral and those with a correlation coefficient
301 <−0.3 were all periportal, showing that the GS activity score successfully
302 rediscovered the polarity of zonation.

303 In stellate cells, the GS activity score for semaphorin-plexin signaling
304 was higher at 8 weeks of HFD (Figure S2C). The positively correlated core
305 gene was *Nrp2*. The negatively correlated core genes were *Nrp1*, *Plxnd1*,
306 semaphorin genes (*Sema3d*, *Sema3e* and *Sema6d*) and *Ntn1*.
307

308 **Discussion**

309 By performing single-nucleus and bulk ATAC-seq on the liver of rats fed or
310 withdrawn from HFD, we studied the transition in cell type composition of
311 tissues and cell type-specific gene regulation in a rat model of NAFLD. The
312 accuracy of ATAC-seq-based deconvolution was validated and utilized to
313 analyze cell type composition. In accordance with the pathological
314 progression from fatty liver to steatohepatitis that occurred between 4 and 8
315 weeks of the HFD intervention in our rat model, the proportion of
316 inflammatory macrophages dramatically increased. In contrast, after 8 weeks
317 of HFD, the proportion of hepatocytes substantially decreased. In the washout
318 condition, where 4 weeks of HFD was followed by 4 weeks of a normal diet,
319 the cell type composition reverted to a state similar to that of rats solely fed
320 with the normal diet.

321 From the global analysis of TF binding and gene expression in single
322 nuclei, we identified major modules of TF regulation (Figure 8), which could
323 be shared among different cell types or be specific to certain cell types. One
324 module shared among hepatocytes, endothelial cells, and macrophages was
325 characterized by the binding of AP-1 TFs and a biological process of TNFα
326 signaling via NF-κB associated with core genes that appear to be cell type-
327 specific. For many of the modules identified, TFs within a given module were
328 known to regulate biological processes assignable to the genes in the same
329 module, thereby supporting the validity of our module discovery algorithm.
330 Lastly, for the biological processes that unveiled in the TF regulation analysis,
331 by incorporating known protein-protein interactions, we could identify core
332 genes, many of which overlap with genes previously implicated in NAFLD (see
333 below).

334 We hypothesized that the gene regulatory machinery of a cell is
335 modular and some of the modules would become active when perturbed by
336 HFD. For analogy, a car consists of many machinery parts, and on a cold

337 sunny day the engine and air conditioning are active but the wiper is not. To
338 discover gene regulatory modules responsive to an HFD stimulus, we
339 conducted an integrative analysis of TF binding and gene expression in nuclei
340 derived from rats subjected to varying dietary conditions. In each cell type,
341 we applied machine learning to decompose the global variation of TF
342 regulation into major modules, where the TFs within a given module appear
343 to regulate the genes within the same module. In the modules identified as
344 such, the TFs and the function of regulated genes agree with current
345 literature: e.g., the AP-1 TFs that regulate inflammation genes (Module 2 of
346 hepatocytes, Module 1 of endothelial cells and Module 1 of macrophages) are
347 known to respond to cytokine stimuli (Hess et al, 2004), and LEF1 that
348 regulates zonation (Module 3 of hepatocytes) is known to activate Wnt
349 signaling (Sun & Weis, 2011). Previous snATAC-seq studies have
350 characterized the difference between cell types (Cusanovich et al, 2018)
351 using TF binding. In contrast, we illustrated how TFs control the states of cells
352 belonging to one cell type (e.g., the hepatocytes with various degrees of
353 inflammation) possibly in a cell-type specific manner (e.g., a distinct set of
354 core genes and other target genes associated with AP-1 TFs in different cell
355 types).

356 A novel method developed in this article is the calculation of the GS
357 activity score in single cells to indicate the activity of a GS, as if it were the
358 expression of a representative gene. In experiments, gene expression can be
359 directly measured, but the activity of a pathway, although of interest, is not
360 directly measurable. One approach could be to take the average expression
361 of genes in that pathway, but this is suboptimal for the following reasons. A
362 simple average would attribute equal weight to both core and peripheral
363 genes in the pathway, thereby amplifying noise from the latter. Moreover,
364 activator and suppressor genes contribute to the pathway in opposing
365 directions. Rather than a simple average, we adopted the first principal

366 component to address these issues. The principal component approach was
367 adapted from the PLAGE program (Tomfohr et al, 2005), which computes
368 pathway activity scores in bulk tissue specimens. The GS activity score could
369 improve our understanding of biological processes. For example, *Abcg1* and
370 *Sqle* correlated positively and negatively, respectively, with the activity score
371 for the steroid metabolism pathway (Figure S2A). As this score increased
372 under the HFD, the GS activity represented by the score presumably functions
373 in response to the excess cholesterol contained in the diet. Thus, a positive
374 correlation of *Abcg1* (for lipid efflux) and a negative correlation of *Sqle* (for
375 sterol biosynthesis) are reasonable. Furthermore, the cell type-specific nature
376 of the GS activity score could reveal cell type-specific dynamics; the
377 inflammation process peaked under 4 weeks of HFD in macrophages and
378 under 8 weeks of HFD in hepatocytes (Figure 7). The GS activity score can
379 also be applied to single-cell and single-nucleus RNA-seq data.
380 Another novel method introduced in this article is the identification of
381 core genes of biological processes. Given a GS comprising known genes for
382 the process, we identified core genes as those exhibiting a strong correlation
383 with the GS activity score and possessing abundant protein-protein
384 interactions within the GS. Evidence for the former was derived from single-
385 nucleus experiment data, and for the latter from a public database. We
386 searched the literature for indications that core genes cause NAFLD or serve
387 as biomarkers for NAFLD, and found such evidence for a significant portion of
388 the core genes (see below): for the TNFα signaling via NF-κB process, 4 of 5
389 in hepatocytes, 4 of 8 in endothelial cells, and 1 of 3 in macrophages; for
390 steroid metabolism 3 of 9 in hepatocytes; and for semaphorin-plexin signaling
391 3 of 7 in stellate cells. The substantial overlap between our current findings
392 and known NAFLD genes serves as a proof-of-concept for our data-driven
393 approach, which utilizes snATAC-seq data from a rat model of dietary
394 intervention as input, then extracts modules, biological processes, and finally

395 core genes. Regarding TNFα signaling via NF-κB in hepatocytes, in animal
396 studies, *Tnf* KO mouse was protected from NAFLD (Kakino et al, 2017), *Nfkb1*
397 KO mouse exacerbated NAFLD (Locatelli et al, 2012), and *Il1r1* KO mouse
398 had less hepatic triacylglycerol (Roos et al, 2009). CXCL12 (aka SDF1)
399 induced migration of stellate cells (Sawitza et al, 2009), endothelial cells
400 (Salcedo et al, 1999) and $CD4^+$ T cells (Boujedidi et al, 2014). In human
401 biomarker studies, higher circulating TNFα level was associated with NAFLD
402 (Potoupni et al, 2021) and future NAFLD (Seo et al, 2013). In NAFLD patients,
403 soluble IL1R1 was higher (Ajmera et al., 2017). Regarding TNFα signaling via
404 NF-κB in endothelial cells, in animal studies, *Pecam1* KO mouse developed
405 NASH (Goel et al, 2007), whereas *Tlr4* KO mouse (Zhang et al, 2020a) and
406 *Il15* KO mouse (Cepero-Donates et al, 2016) were protected from NAFLD.
407 *Ccr5* antagonist alleviated NAFLD in mice (Pérez-Martínez et al, 2014).
408 Regarding TNFα signaling via NF-κB in macrophages, *Cd44* KO mouse had
409 less liver injury under NAFLD inducing diet, and NAFLD patients had elevated
410 serum soluble CD44 (Patouraux et al, 2017). Regarding steroid metabolism
411 in hepatocytes, *Scd1* KO mouse exacerbated NAFLD (Li et al, 2009), and
412 *Acox2* KO mouse developed NAFLD (Zhang et al, 2020b). NAFLD patients
413 exhibited lower serum APOA1 (Yang et al, 2016; Chen et al, 2022). Regarding
414 semaphorin-plexin signaling in stellate cells, in animal studies, *Nrp2*
415 conditional KO reduced liver fibrosis in mice (Rigotti et al, 2022). *Nrp1* KO
416 attenuated cell motility and overexpression increased motility (Cao et al,
417 2010). *Sema3e* KO alleviated liver fibrosis in mice (Yagai et al, 2014).
418 In the pursuit of potential drug targets or biomarkers for NAFLD, we
419 sought proteins whose subcellular localization is extracellular or plasma
420 membrane proteins (Binder et al, 2014) among the genes strongly
421 upregulated under dietary intervention (Table S2) or the core genes positively
422 correlated with NAFLD-related activity. Previous reports on NAFLD biomarkers
423 in humans are cited. For hepatocytes, we identified *Il1r1* (Ajmera et al, 2017),

424 *Il1r2* and *Tnf* (Potoupni et al, 2021). For endothelial cells, *Hgf* was found
425 (Koutsogiannis et al, 2010; Krawczyk et al, 2017; An et al, 2019). *Pecam1*
426 was listed for both endothelial and stellate cells. For stellate cells, we found
427 *Faslg* (Page et al, 2013) and *Nrp2*. For macrophages, we identified RT1-Db1,
428 *Cd44* (Patouraux et al, 2017), *Cxcr4* and *Fn1*. We also searched DGIdb (2022-
429 Feb) for existing drugs that could potentially be repurposed, finding
430 antagonist/antibody/inhibitor of upregulated or positively correlated genes,
431 *Bcl2l1*, *Cxcr4*, *Hgf*, *Il1r1* and *Tnf*, and agonist of downregulated or negatively
432 correlated genes, *Ccr5*, *Cxcr4*, *Tlr2* and *Tlr4* (Table S3). *Cxcr4* was positively
433 correlated with TNFα signaling via NF-κB in macrophages but negatively
434 correlated in endothelial cells.

435 This study has several limitations. Firstly, it was difficult to distinguish
436 TFs with very similar motifs. A specific TF was presumed to bind a specific
437 site if the chromatin was open near the site and if the DNA sequence of the
438 site matched the known TF-binding motif. However, instead, a different TF
439 with a similar motif could have bound at that location. Secondly, the
440 interpretation of gene expression requires caution. Gene expression can be
441 quantified by the amount of either A) mRNA transcription per unit of time, B)
442 mRNA in the nucleus, C) mRNA in the cell, D) protein translation per unit of
443 time, or E) protein. Bulk tissue measurements are conducted by ATAC-seq
444 (A), RNA-seq (C), Ribo-seq (D), or Western blotting (E), while single
445 nucleus/cell measurements are performed by single-nucleus ATAC-seq (A),
446 single-nucleus RNA-seq (B), or single-cell RNA-seq (C). While the gene
447 expression score in this study, which is akin to A, might reflect the state of
448 gene regulation, it could differ from E, which reflects the metabolic state, and
449 the fold-change value would differ among these types of gene expression.
450 Lastly, the TF regulation might differ between rat, mouse and human.

451 To discover TF modules, we utilized machine learning, the
452 performance of which can be influenced by the selection of hyperparameters.

453 If a hyperparameter allows the machine learning model to be excessively
454 complex and flexible relative to the available data, the model may
455 erroneously interpret noise as a signal, leading to overfitting. We chose
456 hyperparameters based on previous empirical validations (for the number of
457 candidate regulators at each node in GENIE3) or residual error (for the
458 number of factors in NMF) but we did not extensively explore different
459 hyperparameters. We evaluated the chosen hyperparameters post hoc.
460 Across a range of hyperparameters, we randomly divided the hepatocytes
461 into two halves, performed TF module discovery in each set, and assessed
462 consistency. We did not observe overfitting for the hyperparameters adopted
463 in this study (Figure S3). Nevertheless, an assessment using data from an
464 independent experiment would be beneficial.
465 By performing single-nucleus ATAC-seq in a rat model of NAFLD and
466 employing novel statistical methods, we elucidated a global picture of *in vivo*
467 TF regulation in each cell type as a set of modules and discovered core genes
468 for NAFLD-relevant biological processes. The core genes can serve as novel
469 candidates for NAFLD causation and our statistical methods can be applied to
470 elucidate TF regulation in biological systems beyond NAFLD.
471
472 **Materials and Methods**
473 **Animal experiment**
474 We used male rats of the SHR/Izm strain, which is an inbred genetic model
475 of essential hypertension. The strain was supplied by Japan SLC. Rats were
476 weaned at 4 weeks after birth and placed on a normal diet (SP diet, Funabashi
477 Farm, Japan). According to the study design, adult rats were either
478 maintained on the normal diet or switched to a high-fat atherogenic diet (24%
479 fat, 15% protein, 5% cholesterol, 2% cholic acid). Compared to models based
480 on nutritional deficiencies (such as the methionine and choline-deficient diet)
481 or chemically-induced models (such as carbon tetrachloride), HFD models

482 more closely resemble human NAFLD, both pathophysiologically and
483 phenotypically (Herck et al, 2017). The SHR strain can mimic NAFLD under a
484 hypertension background. Rats were euthanized under pentobarbital
485 anesthesia (200 mg/kg via intraperitoneal infusion), and the organs were
486 excised and immediately frozen at −70 C. We used the liver's largest lobe (left
487 lateral lobe) for pathology imaging and the right segment of the second-
488 largest lobe (left median lobe) for genomics analysis. The measured
489 outcomes were genomic and biochemical assays of the liver, pathology
490 imaging of the liver, biochemical assays of the serum, organ and body weight,
491 and blood pressure. The experimental unit was the individual animal. No
492 animal was excluded during the experiment. The allocation of animals to
493 dietary conditions was not randomized. While we controlled the dietary
494 conditions, other potential confounders were not accounted for. Group
495 allocation was known to all experimenters except for the pathologist, who
496 examined unlabeled preparations. All animal experiments conformed to the
497 Guidelines for Animal Experiments of the National Center for Global Health
498 and Medicine (NCGM), and were carried out in the animal facility of NCGM.
499 The animal experiments and the protocol were registered and approved by
500 the Animal Research Committee of NCGM (permission number 21009).
501
502 **Library preparation and sequencing for snATAC-seq**
503 Nuclei were isolated from flash-frozen liver tissue according to the 10x
504 Genomics protocol CG000212 for samples m154207 and m154211. For
505 samples m167108, m167203 and m168101, we used the Minute Single
506 Nucleus Isolation Kit for Tissues/Cells (Invent Biotechnologies, SN-047). For
507 each sample, the sequencing library was prepared with the Chromium Next
508 GEM Single Cell ATAC Reagent Kits v1.1 (10x Genomics, CG000209) and
509 sequenced on the NovaSeq 6000 sequencer (Illumina).
510

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

512 Bulk tissue ATAC-seq was performed according to the Omni-ATAC protocol
513 (Corces et al, 2017) using flash-frozen liver tissue. The library was
514 sequenced on the NextSeq sequencer (Illumina).

515

516 **Measuring triglycerides in liver tissue**

517 Triglycerides in the liver tissue were measured enzymatically using the
518 Adipogenesis Colorimetric/Fluorometric Assay Kit (BioVision, K610-100).
519 Tissue homogenate was obtained by dispersing 10 mg tissue in 1 ml lipid
520 extraction solution under 2200 rpm for 1 min, using a handheld disperser
521 (KINEMATICA, PT 1300 D). For fatty liver tissues, we prepared a 1/3 dilution
522 of the homogenate by adding more lipid extraction solution and examined
523 both the original homogenate and the diluted sample. Triglycerides were
524 measured in a colorimetric assay according to the manual. The triglyceride
525 quantity in mole units was converted to gram units by multiplying by 885.7.

526

527 **Measuring total cholesterol in liver tissue**

528 Total cholesterol in the liver tissue was measured enzymatically using the
529 Total Cholesterol and Cholesteryl Ester Colorimetric Assay Kit II (BioVision,
530 K623-100). Tissue homogenate was obtained by dispersing 15 mg tissue in
531 600 ul $CHCl_3$:IPA:NP-40 (7:11:0.1) at 2200 rpm for 1 min, using a handheld
532 disperser (KINEMATICA, PT 1300 D). The homogenate was then centrifuged
533 for 5 min at 15,000 *g*. Afterward, we transferred 200 ul of the liquid phase to
534 a new tube, dried it with air-flow-in at 40 C for 10 min using a centrifugal
535 concentrator (TOMY Digital Biology, CC-105), and dissolved the dried lipids
536 with 400 ul Cholesterol Assay Buffer. Total cholesterol was measured in a
537 colorimetric assay according to the manual.

538

539 **Data cleaning for snATAC-seq**

540 Data cleaning was conducted separately for each sample. ATAC-seq
541 fragments were assigned to individual nuclei, and only nuclei of good quality
542 were retained, as described below. In principle, one nucleus and one
543 barcoded bead are contained in one droplet of the Chromium assay, thus
544 ATAC-seq fragments originating from the same nucleus can be identified by
545 the shared unique barcode. Using Cell Ranger ATAC software (version 2.0.0),
546 we mapped the paired-end reads to the mRatBN7.2/rn7 rat reference genome
547 and identified ATAC-seq fragments, which were assigned to unique barcodes.

548 Occasionally in the Chromium assay, more than one barcoded bead
549 can be contained in a single droplet. In such cases, an excess of ATAC-seq
550 fragments would be commonly assigned to the barcodes. We detected these
551 instances and unified and renamed the barcodes using the bap2 software
552 (version 0.6.7) (Lareau et al, 2020). In subsequent analyses, we could
553 assume that one barcode corresponds to one droplet.

554 We managed the snATAC-seq data using the ArchR package (version
555 1.0.1) (Granja et al, 2021) of the R software (version 4.1.1). ATAC-seq
556 fragments from the autosomes and chromosome X were retained. We kept
557 nuclei that had abundant and enriched fragments around transcription start
558 sites (TSS) as follows: ≥1000 fragments within 1 kb of TSS, the proportion
559 of fragments within 1 kb of TSS being ≥7.5% (for samples m154207,
560 m154211 and m167203) or ≥15% (for samples m167108 and m168101),
561 and Tn5 insertions (which occur at both ends of a fragment) within 25 bp of
562 TSS occurring ≥4 times more frequently than insertions at 1900–2000 bp
563 from TSS.

564 For each nucleus, the proportion of fragment counts from a
565 chromosome (e.g., chromosome 1) is approximately constant, being
566 proportional to the chromosome length. However, this could deviate if a
567 droplet is contaminated by free DNA from broken nuclei (Orchard et al, 2020).

568 Therefore, we represented the proportion by a matrix indexed by nuclei
569 (rows) and chromosomes (columns). To standardize the matrix, we
570 subtracted column-wise averages and divided by the standard deviation. By
571 taking the sum of the absolute values of a row, we defined the deviation score
572 of a nucleus. To exclude contaminated droplets, we discarded nuclei with a
573 score of ≥40.

574 Since Tn5 inserts less efficiently into nucleosome-associated DNA,
575 ATAC-seq fragment size distribution exhibits periodicity (Cusanovich et al,
576 2018). Using the program provided by Cusanovich et al.
577 (https://atlas.gs.washington.edu/mouse-atac/docs/), we computed, for each
578 nucleus, the spectral densities for the 100–300 bp range, which corresponds
579 to the pitch of nucleosomes. To exclude nuclei with aberrant chromatin states,
580 we discarded those with scores below the 1st or above the 99th percentile.

581 When two or more nuclei are contained in one droplet, known as a
582 multiplet, those nuclei share the same barcode, violating the one-nucleus-to-
583 one-barcode assumption. We removed the barcodes for multiplets in two
584 ways. First, we inferred the number of homologous chromosomes from which
585 the ATAC-seq fragments assigned to a barcode were derived. This number is
586 analogous to ploidy, but differs because the container is a droplet, not a cell.
587 By counting overlaps among the fragments, the number of homologous
588 chromosomes was estimated to be either 2*n*, 4*n* or ≥6*n* using the R package
589 scPloidy (version 0.3.0). Since the majority of hepatocytes in rat liver are
590 either diploids or tetraploids (Katsuda et al, 2020) and other cell types in the
591 liver are diploids, droplets with ≥6*n* were regarded as potential multiplets and
592 discarded. Second, we detected doublets formed by two heterogeneous cell
593 types using ArchR. Droplets with a doublet enrichment score >10 were
594 discarded.

595

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

597 **snATAC-seq**

598 The Tn5 insertions observed in ATAC-seq occur at open chromatin regions,
599 such as promoters and enhancers, and the genome-wide profile of Tn5
600 insertions characterizes the distinct cell types. We performed nucleus
601 clustering based on the similarity of Tn5 insertions. Firstly, we represented
602 the Tn5 insertions in snATAC-seq as a "tile matrix": a binary matrix indexed
603 by tiles of 500 bp width across the genome (rows) and nuclei (columns),
604 coding for the presence (1) or absence (0) of Tn5 insertions. Using ArchR, we
605 generated a matrix with 5,230,329 rows and 14,615 columns. Subsequently,
606 we grouped nuclei based on the similarity of the corresponding columns (see
607 the following section). For each group, pseudo-bulk ATAC-seq datasets were
608 generated, and Tn5 insertion peaks were computed using ArchR and the
609 MACS2 program (version 2.1.4) (Zhang et al, 2008). We identified 599,035
610 chromatin accessibility peaks with a width of 501 bp. The fraction of reads in
611 peaks (FRiP) was 0.444 (sample m154207), 0.407 (m154211), 0.541
612 (m167108), 0.446 (m167203) and 0.557 (m168101). The proportion of
613 reads that aligned within 50bp of TSS was 0.0286 (sample m154207), 0.0224
614 (m154211), 0.0604 (m167108), 0.0340 (m167203) and 0.0646 (m168101).

615 Secondly, we represented the Tn5 insertions as a "peak matrix": a
616 binary matrix indexed by peaks (rows) and nuclei (columns), coding for the
617 presence or absence of Tn5 insertions. Again, we grouped nuclei based on
618 the similarity of the corresponding columns. The nucleus clustering based on
619 the peak matrix was used in downstream analysis, as it is likely more accurate
620 than the clustering based on the tile matrix.

621

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

623 The tile and peak matrices are large binary matrices indexed by genomic
624 positions (rows) and nuclei (columns). Since the column vectors are
625 extremely high-dimensional, we employed dimensionality reduction to

effectively analyze the similarity among nuclei. We used latent semantic indexing (Berry & Browne, 2005) and singular value decomposition (SVD) for this purpose. Dimensionality reduction was performed on all samples combined. As experimental conditions, including the Tn5:DNA ratio, may vary among samples, we introduced a step for sample batch correction. We applied the same algorithm to both 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

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

which ranges between zero and one. To limit noise, we retained rows with an IDF-weighted variance >0.4 for the tile matrix and >0.1 for 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 nucleus j belonging to sample s, the correction factor is

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

If the denominator equals zero, γ_{ij} is set as one. The above correction requires no knowledge of cell types. However, if cell type classification of nuclei is available, the correction can be refined. For nucleus j of cell type c

belonging to sample s, the correction factor becomes

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

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

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

1. Compute the sample batch correction factor γ_{ij} using equation (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 the truncated SVD, UDV^T, for the 30 largest singular values using the R package irlba (version 2.3.3) (Baglama & Reichel, 2005).
4. Take the truncated SVD for the two largest singular values, project the nuclei to the VD coordinates, and detect and discard outlier nuclei. Outliers were detected based on the sum of the squared distance to the five nearest neighbors.
5. Take the truncated SVD for the non-noise singular values, and project the nuclei to the VD coordinates. In our data, as the 11th and later singular values bottomed out and the corresponding columns of V appeared randomly distributed, 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 the 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

680 corresponded to hepatocytes, stellate cells, endothelial cells and white
681 blood cells.
682 7. Compute the sample batch correction factor γ_{ij} using equation (2),
683 incorporating the coarse clustering of nuclei.
684 8. Compute the matrix $(g_i \cdot d_j \cdot f_{ij} \cdot \gamma_{ij})$ and centralize each row by
685 subtracting the average.
686 9. Compute the truncated SVD, UDV^T, for the 30 largest singular values.
687 10. Take the truncated SVD for the non-noise singular values, and project the
688 nuclei to the VD coordinates. We took the eight largest singular values.
689 11. Identify clusters of nuclei through shared nearest-neighbor modularity
690 optimization.
691
692 **Cell type assignment for snATAC-seq**
693 For each nucleus cluster obtained above, we assigned a cell type based on
694 chromatin accessibility at known marker genes. For each cluster, chromatin
695 accessibility at the TSS and gene body of the genes was visually inspected
696 using the browser of ArchR. We used the following marker genes in the rat:
697 - *Alb*, *Apoc3*, *Cyp2e1* and *Cyp2f4* for hepatocyte
698 - *Epcam* and *Krt19* for cholangiocyte
699 - *Stab2* for endothelial cell
700 - *Dcn*, *Des*, *Lrat* and *Acta2* for stellate cell
701 - *Csf3r* for granulocyte
702 - *Elane*, *Ly6g6c*, *Ly6g6d*, *Ly6g6e* and *Mpo* for neutrophil
703 - *Csf1r* for macrophage (including Kupffer cell)
704 - *Lyz2* for inflammatory macrophage
705 - *Marco* for non-inflammatory macrophage
706 - *Cd3g* and *Gata3* for T cell
707 - *Gzma* and *Prf1* for cytotoxic T cell and NK cell
708 - *Ebf1* for B cell

Data cleaning for bulk ATAC-seq

In accordance with the ATAC-seq data processing workflow (Reske et al, 2020), we trimmed reads, mapped reads to the reference genome, eliminated mitochondrial DNA reads, subsampled to equalize the library complexity, and removed PCR duplicates.

Inferring cell type composition of bulk tissue samples

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

By summing the 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 number as

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

where c_j represents the number of nuclei included in the normalized ATAC-seq library for pure cell type j and d_k adjusts for the 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 denote C as the diagonal matrix with c_j on the diagonal. The values of c_j for hepatocytes were halved compared to the values for other cell types due to the higher ploidy of hepatocytes.

We used the DeconPeaker software (version 1.0) (Li et al, 2020) to select the peaks (rows) that distinguish cell types. DeconPeaker chooses these 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 forms the reference vector $\boldsymbol{w}_j$ for chromatin accessibility:

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

W represents 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, yielding 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 as

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

To avoid over-correction, we cap f_i to range between ${Q_1}^2/Q_2$ and ${Q_3}^2/Q_2$, where Q_1, Q_2, Q_3 are the quartiles of the original value. We represent by F the diagonal matrix with f_i in the diagonal. After correction, a bulk ATAC-seq experiment $\boldsymbol{y}_k$ becomes $F^{-1}\boldsymbol{y}_k$. Note 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 type 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 of bulk samples was compared

among dietary conditions. We performed multiple regression where the dependent variable was the imputed composition and the independent variables were indicators of dietary condition, with the normal diet serving as the reference. We applied Bonferroni's correction for multiple testing of seven cell types and three conditions (apart from the normal diet), setting the significance level 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 quantity 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 the library size normalization of the R package Linnorm (version 2.16.0) (Yip et al, 2017) and the imputation of the R package SAVER (version 1.1.2) (Huang et al, 2018). Prior to imputation, 47% of the matrix elements were zero, and post-imputation, 0.2% were set to the lowest value (0.001). We excluded low expression genes, where the lowest value occurred in ≥1% of nuclei. We took the log2 of the matrix and applied quantile normalization. The obtained matrix was used for inspecting cell type marker genes. We applied SVD and subtracted the first two components from the matrix, which captured the sample batch effect. Although the two components also captured major cell type differences, we anticipated that the removal would not harm downstream analysis, which compares nuclei of the same cell type between samples.

Quantifying TF binding in snATAC-seq

If the DNA sequence of a chromatin accessibility peak includes the motif of a

823 TF, Tn5 insertion in the peak suggests the binding of the corresponding TF.
824 For each nucleus, by aggregating over all peaks, we can infer the genome-
825 wide binding of the TF. For each TF, the enrichment of its motif in the peaks
826 was quantified by the chromVAR (Schep et al, 2017) deviation z-score in
827 ArchR. The CIS-BP database of TF binding motifs (Weirauch et al, 2014) was
828 used. In the same way as we did for the Gene Score Matrix, we applied
829 quantile normalization to the Motif Matrix and subtracted the first two
830 components of SVD.
831
832 **Gene set enrichment analysis of differentially expressed genes under**
833 **HFD**
834 To seek which biological processes and/or pathways are differentially
835 expressed under a particular dietary intervention, we performed gene set
836 enrichment analysis. The target GSs were obtained from the Molecular
837 Signatures Database (MSigDB) (version 7.4) (Subramanian et al, 2005) using
838 the R package msigdbr (version 7.4.1). Among the datasets in MSigDB, we
839 used Gene Ontology for Biological Process, Hallmark gene sets, KEGG
840 pathways, Reactome pathways, and WikiPathways. Using the sample batch-
841 corrected Gene Score Matrix, we compared nuclei from a pair of dietary
842 conditions, for example, hepatocytes from 4 weeks of HFD *vs.* the normal
843 diet. The statistical significance of gene set enrichment was tested using the
844 R package PADOG (version 1.34.0) (Tarca et al, 2012) and by performing
845 permutations for 10,000 times or 20 times the number of GSs in a dataset,
846 whichever was larger. Within each dataset, the GSs that attained a false
847 discovery rate <0.05 were considered statistically significant. Significant GSs
848 were visually inspected using the EnrichmentMap App (version 3.3.3) of the
849 Cytoscape software (version 3.8.2).
850
851 **Discovering major modules of TF regulation and assigning biological**

852 **processes**

853 To elucidate a global picture of TF regulation, we first inferred the regulatory
854 impacts of all TFs on all genes and then extracted major modules of the
855 regulatory relationships between TFs and genes. For the GRN identification in
856 the first step, the relationships between TFs and genes were uncovered by
857 incorporating a third piece of information, chromatin accessibility observed in
858 snATAC-seq, as in previous algorithms(Fleck et al, 2022; González-Blas et al,
859 2022; Jiang et al, 2022; Kamal et al, 2022; Kartha et al, 2022; Wang et al,
860 2022; Kamimoto et al, 2023) (Figure 9). Compared to other GRN
861 identification algorithms, our TF-gene linking criterion was relatively relaxed,
862 yielding more biologically relevant modules (Figure S4). For each gene, we
863 computed its regulatory TFs using the R package GENIE3 (version 1.14.0)
864 (Huynh-Thu et al, 2010), which implements a random forest algorithm, thus
865 enabling modeling of nonlinear regulation. The algorithm aims to predict the
866 expression level of the gene based on the binding of TFs and dietary
867 conditions. As for the number of candidate regulators randomly selected at
868 each node of the random forests, we chose the square root of the number of
869 predictors, as suggested in the GENIE3 publication. GENIE3 quantified the
870 impacts of regulators by the proportion of explained variance. By combining
871 the results from all genes, we obtained a regulator-regulatee matrix, in which
872 the rows correspond to TFs and dietary conditions, the columns correspond
873 to genes, and an element represents the importance of a TF to a gene.

874 Next, we applied nonnegative matrix factorization implemented in the
875 R package NMF (version 0.23.0) (Gaujoux & Seoighe, 2010) to the regulator-
876 regulatee matrix. We used the Brunet algorithm (Brunet et al, 2004) with 100
877 random starts. The number of factors was chosen based on the data as the
878 inflection point of residual error (Hutchins et al, 2008). In the output of NMF,
879 the membership of TFs (regulators) and genes (regulatees) to a factor was
880 defined by nonnegative weight. We termed these factors as “modules” of TF

881 regulation.

882 We discovered biological processes that characterize a module by
883 applying gene set enrichment analysis to the genes belonging to the module.
884 For each TF regulation module, we tested if heavily weighted genes were
885 enriched in certain GSs. For each GS (see above section), we computed the
886 P-value and *t*-statistics to test if genes in the given GS had a larger weight
887 than other genes. To account for multiple testing by the GSs, we performed
888 100 trials of permuting gene names and computed family-wise error rates.
889 The GSs that attained a family-wise error rate of <0.05 were considered
890 statistically significant. The permutation test does not require the assumption
891 of normality.

892

893 **GS activity score in single nuclei**

894 To infer the activity of a GS in single nuclei, we computed the "average" score
895 of TF binding and gene expression, with regard to the TFs/genes included in
896 the GS. Formally, we defined the weight of TFs/genes as one if it was included
897 in the GS and as zero otherwise. We then performed SVD for the TF binding
898 and gene expression matrix incorporating the weight. In the Motif Matrix (see
899 above), each TF (row) was standardized by subtracting the mean and being
900 divided by the standard deviation and then multiplied by the weight of the TF.
901 Similarly, in the Gene Score Matrix (see above), each gene (row) was
902 standardized by subtracting the mean and being divided by the standard
903 deviation and then multiplied by the weight of the gene. By stacking the two
904 matrices, we obtained a large matrix, in which the rows correspond to TFs
905 and genes and the columns correspond to nuclei. Finally, SVD was applied
906 and the first right-singular vector was defined as the GS activity score in
907 single nuclei.

908 Although the above-mentioned activity score can always be computed,
909 it would be meaningless if the TFs/genes in the GS were not co-expressed.

910 Therefore, we conducted a statistical test to exclude such meaningless GS.
911 We checked whether the TFs/genes in the GS were mutually correlated more
912 than by chance. We computed the null distribution of the first singular value
913 by performing 100 trials of the SVD procedure but with the weight of TFs
914 permutated and the weight of genes permutated and verified whether the
915 actual first singular value could attain an empirical P-value of <0.05. The
916 permutation test does not require the assumption of normality.
917 The principal component approach was adapted from the PLAGE
918 program (Tomfohr et al, 2005) that was developed to compute the pathway
919 activity score of a GS in bulk tissue specimens. Instead, we applied it to nuclei.
920 Compared to the number of specimens in a bulk experiment, the number of
921 nuclei is much larger, which enabled significance testing by permutation.
922

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

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Data Availability

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

Acknowledgments

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

Funding

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

Author contributions

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

Software, F.T.;

Formal Analysis, F.T.;

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

Writing – Original Draft, F.T.;

Writing – Review & Editing, all authors;

Supervision, N.K.

All authors have read and approved the manuscript.

Conflict of interest

The authors declare no conflict of interest.

963 **FIGURE LEGENDS**

964 **Figure 1**

965 Cell types observed in the liver and their composition

966 (A) Experimental design.

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

969 (C) Expression score of cell type marker genes in nuclei (vertical axis) is
970 compared between nuclei clusters (horizonal axis) in violin plots.

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

975 (E) Proportion of nuclei from a specific cell type (vertical axis) compared
976 across dietary conditions (horizontal axis). The proportion in bulk tissue
977 ATAC-seq samples was imputed by deconvolution (n=4 samples per
978 condition). For each cell type, the difference in proportion among dietary
979 conditions was tested by a linear regression where the dependent variable
980 was the imputed composition and the independent variables were indicators
981 of dietary condition. Asterisks (*) indicates statistically significant changes
982 from normal diet under the significance level of $0.05/[7 \times 3]$, adjusted for
983 multiple testing.

984 **Figure 2**

985 Biological processes that were differentially expressed under dietary
986 interventions

987 The heatmap illustrates whether a biological process (row) was differentially
988 expressed in a cell type under a dietary intervention compared with the
989 normal diet (column). The color indicates the statistical significance of the
990 gene set enrichment analysis: red if significant adjusting for multiple testing,
991 brown if nominally significant at $P < 0.01$, light brown if $P < 0.05$, and white

992 otherwise.

993 **Figure 3**

994 Major modules of TF regulation in hepatocytes

995 (A) The central heatmap visualizes the regulatory impact of TF binding
996 (columns) on genes (rows). While the impact was computed between all TFs
997 and all genes, only representative TFs and genes are plotted for visibility.
998 Subgroups of TFs/genes demonstrating strong within-group regulation were
999 extracted as modules, which are highlighted by red rectangles. The heatmap
1000 at the bottom shows the membership of TFs (columns) to modules (rows).
1001 The heatmap on the right displays the membership of genes (rows) to
1002 modules (columns). The heatmaps only include representative TFs/genes
1003 with membership rankings in the top five and membership weights exceeding
1004 half of the top.

1005 (B) Gene set enrichment analysis for genes belonging to the modules.
1006 Excessive overlap between genes belonging to a module and the GS for a
1007 biological process (horizontal axis) was tested, and the P-value was plotted
1008 (vertical axis). Bars are colored black if statistically significant, adjusting for
1009 multiple testing, and gray otherwise. Biological processes that were
1010 significant for at least one module are shown.

1011 **Figure 4**

1012 Major modules of TF regulation in endothelial cells

1013 See legend in Figure 3.

1014 **Figure 5**

1015 Major modules of TF regulation in stellate cells

1016 See legend in Figure 3.

1017 **Figure 6**

1018 Major modules of TF regulation in macrophages

1019 See legend in Figure 3.

1020 **Figure 7**

1021 Core genes for the TNFα signaling via NF-κB processes are identified based
1022 on correlation with GS activity score and protein-protein interaction

1023 (A) Analysis in hepatocytes using the GS for TNFα signaling via NF-κB. In the
1024 top panel, the GS activity score in single nuclei (vertical axis) is compared
1025 across dietary conditions (horizontal axis) in violin plots. Box plots within the
1026 violins indicate quartiles. The score distribution under a dietary intervention
1027 was compared to that under the normal diet condition using the area under
1028 the receiver operating characteristic curve: * and ** indicate larger than 0.75
1029 and 0.9, respectively. In the remaining panels, TFs and genes are plotted by
1030 their correlation with the GS activity score (horizontal axis) and the total
1031 amount of protein-protein interactions with the GS (vertical axis). Points
1032 representing TFs/genes having protein-protein interactions within the top 5%
1033 are colored red, while others are colored blue. The TFs/genes in the GS
1034 (middle panel) and not in the GS (bottom panel) are plotted separately; the
1035 former exhibit more abundant protein-protein interactions. TFs are labeled in
1036 all capital letters. The GS was taken from the MSigDB Hallmark database.
1037 Protein-protein interactions were taken from the STRING database (version
1038 11.0).

1039 (B) Analysis in endothelial cells.

1040 (C) Analysis in macrophages.

1041 **Figure 8**

1042 Modules of global TF regulation, biological processes, and core genes were
1043 detected by the data-driven approach

1044 See text for the full list of core genes.

1045 **Figure 9**

1046 Criteria for defining GRNs are compared across studies

1047 In all algorithms, the link between TFs and genes is established by
1048 incorporating the information from ATAC-seq peaks. The common

1049 requirement is matching the TF's motif with the DNA sequence at an ATAC-
1050 seq peak.
1051

1052 **REFERENCES**

1053

1054 Ajmera V, Perito ER, Bass NM, Terrault NA, Yates KP, Gill R, Loomba R, Diehl AM,
1055 Aouizerat BE, Network for the NCR (2017) Novel plasma biomarkers associated with
1056 liver disease severity in adults with nonalcoholic fatty liver disease. *Hepatology* 65:
1057 65–77. doi:10.1002/hep.28776.

1058 An X, Liu J, Zheng X, Dou Z, Li Y, Suo Y, Ma YN, Sun M, Tian ZY, Xu L (2019) Serum
1059 HGF ,PCIII and PLT are Noninvasive Markers for the diagnosis of nonalcoholic fatty
1060 liver disease. *Research Square*. doi:10.21203/rs.2.18035/v1.

1061 Baglama J, Reichel L (2005) Augmented Implicitly Restarted Lanczos
1062 Bidiagonalization Methods. *Siam J Sci Comput* 27: 19–42. doi:10.1137/04060593x.

1063 Barrow F, Khan S, Fredrickson G, Wang H, Dietsche K, Parthiban P, Robert S, Kaiser
1064 T, Winer S, Herman A *et al* (2021) Microbiota-Driven Activation of Intrahepatic B
1065 Cells Aggravates NASH Through Innate and Adaptive Signaling. *Hepatology Baltim*
1066 *Md* 74: 704–722. doi:10.1002/hep.31755.

1067 Behari J, Yeh T-H, Krauland L, Otruba W, Cieply B, Hauth B, Apte U, Wu T, Evans R,
1068 Monga SPS (2010) Liver-Specific β-Catenin Knockout Mice Exhibit Defective Bile
1069 Acid and Cholesterol Homeostasis and Increased Susceptibility to Diet-Induced
1070 Steatohepatitis. *Am J Pathology* 176: 744–753. doi:10.2353/ajpath.2010.090667.

1071 Berry MW, Browne M (2005) Understanding Search Engines (SIAM).

1072 Binder JX, Pletscher-Frankild S, Tsafou K, Stolte C, O'Donoghue SI, Schneider R,
1073 Jensen LJ (2014) COMPARTMENTS: unification and visualization of protein
1074 subcellular localization evidence. *Database J Biological Databases Curation* 2014:
1075 bau012. doi:10.1093/database/bau012.

1076 Boujedidi H, Robert O, Bignon A, Cassard-Doulcier A-M, Renoud M-L, Gary-Gouy H,
1077 Hemon P, Tharinger H, Prévot S, Bachelerie F *et al* (2014) CXCR4 dysfunction in
1078 non-alcoholic steatohepatitis in mice and patients. *Clin Sci* 128: 257–267.
1079 doi:10.1042/cs20130833.

1080 Brunet J-P, Tamayo P, Golub TR, Mesirov JP (2004) Metagenes and molecular
1081 pattern discovery using matrix factorization. *Proc National Acad Sci* 101: 4164-4169.
1082 doi:10.1073/pnas.0308531101.

1083 Cao S, Yaqoob U, Das A, Shergill U, Jagavelu K, Huebert RC, Routray C,
1084 Abdelmoneim S, Vasdev M, Leof E *et al* (2010) Neuropilin-1 promotes cirrhosis of
1085 the rodent and human liver by enhancing PDGF/TGF-β signaling in hepatic stellate
1086 cells. *J Clin Invest* 120: 2379-2394. doi:10.1172/jci41203.

1087 Cepero-Donates Y, Lacraz G, Ghobadi F, Rakotoarivelo V, Orkhis S, Mayhue M,
1088 Chen Y-G, Rola-Pleszczynski M, Menendez A, Ilangumaran S *et al* (2016)
1089 Interleukin-15-mediated inflammation promotes non-alcoholic fatty liver disease.
1090 *Cytokine* 82: 102-111. doi:10.1016/j.cyto.2016.01.020.

1091 Chen C, Li H, Song J, Zhang C, Li M, Mao Y, Liu A, Du J (2022) Role of
1092 Apolipoprotein A1 in PPAR Signaling Pathway for Nonalcoholic Fatty Liver Disease.
1093 *Ppar Res* 2022: 4709300. doi:10.1155/2022/4709300.

1094 Chistiakov DA, Myasoedova VA, Revin VV, Orekhov AN, Bobryshev YV (2018) The
1095 impact of interferon-regulatory factors to macrophage differentiation and
1096 polarization into M1 and M2. *Immunobiology* 223: 101-111.
1097 doi:10.1016/j.imbio.2017.10.005.

1098 Chu AL, Schilling JD, King KR, Feldstein AE (2020) The Power of Single Cell
1099 Analysis for the Study of Liver Pathobiology. *Hepatology*: hep.31485-23.
1100 doi:10.1002/hep.31485.

1101 Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna
1102 S, Satpathy AT, Rubin AJ, Montine KS, Wu B *et al* (2017) An improved ATAC-seq
1103 protocol reduces background and enables interrogation of frozen tissues. *Nature*
1104 *Methods* 14: 959-962. doi:10.1038/nmeth.4396.

1105 Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, Filippova
1106 GN, Huang X, Christiansen L, DeWitt WS *et al* (2018) A Single-Cell Atlas of
1107 In Vivo Mammalian Chromatin Accessibility. *Cell* 174: 1309-1324.e18.
1108 doi:10.1016/j.cell.2018.06.052.

1109 Date D, Das R, Narla G, Simon DI, Jain MK, Mahabeleshwar GH (2014) Kruppel-like
1110 Transcription Factor 6 Regulates Inflammatory Macrophage Polarization. *J Biol*
1111 *Chem* 289: 10318-10329. doi:10.1074/jbc.m113.526749.

1112 Fleck JS, Jansen SMJ, Wollny D, Zenk F, Seimiya M, Jain A, Okamoto R, Santel M,
1113 He Z, Camp JG *et al* (2022) Inferring and perturbing cell fate regulomes in human
1114 brain organoids. *Nature*: 1-8. doi:10.1038/s41586-022-05279-8.

1115 Friedman SL, Neuschwander-Tetri BA, Rinella M, Sanyal AJ (2018) Mechanisms of
1116 NAFLD development and therapeutic strategies. *Nat Med* 24: 908-922.
1117 doi:10.1038/s41591-018-0104-9.

1118 Gaujoux R, Seoighe C (2010) A flexible R package for nonnegative matrix
1119 factorization. *BMC Bioinformatics* 11: 367. doi:10.1186/1471-2105-11-367.

1120 Goel R, Boylan B, Gruman L, Newman PJ, North PE, Newman DK (2007) The
1121 proinflammatory phenotype of PECAM-1-deficient mice results in atherogenic diet-
1122 induced steatohepatitis. *Am J Physiol-gastr L* 293: G1205-G1214.
1123 doi:10.1152/ajpgi.00157.2007.

1124 González-Blas CB, Winter SD, Hulselmans G, Hecker N, Matetovici I, Christiaens V,
1125 Poovathingal S, Wouters J, Aibar S, Aerts S (2022) SCENIC+: single-cell multiomic
1126 inference of enhancers and gene regulatory networks. *Biorxiv*: 2022.08.19.504505.
1127 doi:10.1101/2022.08.19.504505.

1128 Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, Greenleaf
1129 WJ (2021) ArchR is a scalable software package for integrative single-cell chromatin
1130 accessibility analysis. *Nature Genetics* 53: 403 - 411. doi:10.1038/s41588-021-
1131 00790-6.

1132 Halpern KB, Shenhav R, Matcovitch-Natan O, Tóth B, Lemze D, Golan M, Massasa
1133 EE, Baydatch S, Landen S, Moor AE *et al* (2017) Single-cell spatial reconstruction
1134 reveals global division of labour in the mammalian liver. *Nature* 542: 352-356.
1135 doi:10.1038/nature21065.

1136 Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ,
1137 Darby C, Zager M *et al* (2021) Integrated analysis of multimodal single-cell data. *Cell*
1138 184: 3573-3587.e29. doi:10.1016/j.cell.2021.04.048.

1139 Herck MAV, Vonghia L, Francque SM (2017) Animal Models of Nonalcoholic Fatty
1140 Liver Disease—A Starter's Guide. *Nutrients* 9: 1072. doi:10.3390/nu9101072.

1141 Hess J, Angel P, Schorpp-Kistner M (2004) AP-1 subunits: quarrel and harmony
1142 among siblings. *J Cell Sci* 117: 5965-5973. doi:10.1242/jcs.01589.

1143 Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, Murray JI, Raj A, Li M,
1144 Zhang NR (2018) SAVER: gene expression recovery for single-cell RNA sequencing.
1145 *Nature Methods* 15: 539 - 542. doi:10.1038/s41592-018-0033-z.

1146 Hutchins LN, Murphy SM, Singh P, Graber JH (2008) Position-dependent motif
1147 characterization using non-negative matrix factorization. *Bioinformatics* 24: 2684-
1148 2690. doi:10.1093/bioinformatics/btn526.

1149 Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) Inferring Regulatory
1150 Networks from Expression Data Using Tree-Based Methods. *Plos One* 5: e12776.
1151 doi:10.1371/journal.pone.0012776.

1152 Jew B, Alvarez M, Rahmani E, Miao Z, Ko A, Garske KM, Sul JH, Pietiläinen KH,
1153 Pajukanta P, Halperin E (2020) Accurate estimation of cell composition in bulk
1154 expression through robust integration of single-cell information. *Nature*
1155 *Communications* 11: 1971. doi:10.1038/s41467-020-15816-6.

1156 Jiang J, Lyu P, Li J, Huang S, Tao J, Blackshaw S, Qian J, Wang J (2022) IReNA:
1157 Integrated regulatory network analysis of single-cell transcriptomes and chromatin
1158 accessibility profiles. *Iscience* 25: 105359. doi:10.1016/j.isci.2022.105359.

1159 Kakino S, Ohki T, Nakayama H, Yuan X, Otabe S, Hashinaga T, Wada N, Kurita Y,
1160 Tanaka K, Hara K *et al* (2017) Pivotal Role of TNF-a in the Development and
1161 Progression of Nonalcoholic Fatty Liver Disease in a Murine Model. *Horm Metab Res*
1162 50: 80-87. doi:10.1055/s-0043-118666.

1163 Kamal A, Arnold C, Claringbould A, Moussa R, Servaas NH, Kholmatov M, Daga N,
1164 Nogina D, Mueller-Dott S, Reyes-Palomares A *et al* (2022) GRaNIE and GRaNPA:
1165 Inference and evaluation of enhancer-mediated gene regulatory networks applied to
1166 study macrophages. *Biorxiv*. 2021.12.18.473290. doi:10.1101/2021.12.18.473290.

1167 Kamimoto K, Stringa B, Hoffmann CM, Jindal K, Solnica-Krezel L, Morris SA (2023)
1168 Dissecting cell identity via network inference and in silico gene perturbation. *Nature*
1169 614: 742-751. doi:10.1038/s41586-022-05688-9.

1170 Kartha VK, Duarte FM, Hu Y, Ma S, Chew JG, Lareau CA, Earl A, Burkett ZD,
1171 Kohlway AS, Lebofsky R *et al* (2022) Functional inference of gene regulation using
1172 single-cell multi-omics. *Cell Genom* 2: 100166. doi:10.1016/j.xgen.2022.100166.

1173 Katsuda T, Hosaka K, Matsuzaki J, Usuba W, Prieto-Vila M, Yamaguchi T, Tsuchiya
1174 A, Terai S, Ochiya T (2020) Transcriptomic Dissection of Hepatocyte Heterogeneity:
1175 Linking Ploidy, Zonation, and Stem/Progenitor Cell Characteristics. *Cellular and*
1176 *Molecular Gastroenterology and Hepatology* 9: 161-183.
1177 doi:10.1016/j.jcmgh.2019.08.011.

1178 Kiki I, Altunkaynak BZ, Altunkaynak ME, Vuraler O, Unal D, Kaplan S (2007) Effect of
1179 High Fat Diet on the Volume of Liver and Quantitative Feature of Kupffer Cells in
1180 the Female Rat: A Stereological and Ultrastructural Study. *Obes Surg* 17: 1381.
1181 doi:10.1007/s11695-007-9219-7.

1182 Koutsogiannis D, Summers K, George B, Adams P, Marotta P, Chakrabarti S (2010)
1183 Identification of Serum Biomarkers in End Stage Liver Disease. *Open Biomarkers J*
1184 3: 1-6. doi:10.2174/1875318301003010001.

1185 Krawczyk M, Zimmermann S, Hess G, Holz R, Dauer M, Raedle J, Lammert F,
1186 Grünhage F (2017) Panel of three novel serum markers predicts liver stiffness and
1187 fibrosis stages in patients with chronic liver disease. *Plos One* 12: e0173506.
1188 doi:10.1371/journal.pone.0173506.

1189 Lareau CA, Ma S, Duarte FM, Buenrostro JD (2020) Inference and effects of
1190 barcode multiplets in droplet-based single-cell assays. *Nature Communications* 11:
1191 866. doi:10.1038/s41467-020-14667-5.

1192 Li H, Sharma A, Luo K, Qin ZS, Sun X, Liu H (2020) DeconPeaker, a Deconvolution
1193 Model to Identify Cell Types Based on Chromatin Accessibility in ATAC-Seq Data of
1194 Mixture Samples. *Frontiers in Genetics* 11: 392. doi:10.3389/fgene.2020.00392.

1195 Li ZZ, Berk M, McIntyre TM, Feldstein AE (2009) Hepatic Lipid Partitioning and Liver
1196 Damage in Nonalcoholic Fatty Liver Disease ROLE OF STEAROYL-CoA
1197 DESATURASE. *J Biol Chem* 284: 5637-5644. doi:10.1074/jbc.m807616200.

1198 Locatelli I, Sutti S, Vacchiano M, Bozzola C, Albano E (2012) NF-κB1 deficiency
1199 stimulates the progression of non-alcoholic steatohepatitis (NASH) in mice by
1200 promoting NKT-cell-mediated responses. *Clin Sci* 124: 279-287.
1201 doi:10.1042/cs20120289.

1202 Orchard P, Kyono Y, Hensley J, Kitzman JO, Parker SCJ (2020) Quantification,
1203 Dynamic Visualization, and Validation of Bias in ATAC-Seq Data with ataqv. *Cell*
1204 *Systems* 10: 298-306.e4. doi:10.1016/j.cels.2020.02.009.

1205 Page S, Birerdinc A, Estep M, Stepanova M, Afendy A, Petricoin E, Younossi Z,
1206 Chandhoke V, Baranova A (2013) Knowledge-Based Identification of Soluble
1207 Biomarkers: Hepatic Fibrosis in NAFLD as an Example. *Plos One* 8: e56009.
1208 doi:10.1371/journal.pone.0056009.

1209 Patouraux S, Rousseau D, Bonnafous S, Lebeaupin C, Luci C, Canivet CM, Schneck
1210 A-S, Bertola A, Saint-Paul M-C, Iannelli A *et al* (2017) CD44 is a key player in non-
1211 alcoholic steatohepatitis. *J Hepatol* 67: 328-338. doi:10.1016/j.jhep.2017.03.003.

1212 Pérez-Martínez L, Pérez-Matute P, Aguilera-Lizarraga J, Rubio-Mediavilla S, Narro
1213 J, Recio E, Ochoa-Callejero L, Oteo J-A, Blanco J-R (2014) Maraviroc, a CCR5
1214 antagonist, ameliorates the development of hepatic steatosis in a mouse model of
1215 non-alcoholic fatty liver disease (NAFLD). *J Antimicrob Chemoth* 69: 1903-1910.
1216 doi:10.1093/jac/dku071.

1217 Potoupni V, Georgiadou M, Chatzigriva E, Polychronidou G, Markou E, Gakis CZ,
1218 Filimidou I, Karagianni M, Anastasilakis D, Evripidou K *et al* (2021) Circulating tumor
1219 necrosis factor-α levels in non-alcoholic fatty liver disease: A systematic review
1220 and a meta-analysis. *J Gastroen Hepatol* 36: 3002-3014. doi:10.1111/jgh.15631.

1221 Ramachandran P, Dobie R, Wilson-Kanamori JR, Dora EF, Henderson BEP, Luu NT,
1222 Portman JR, Matchett KP, Brice M, Marwick JA *et al* (2019) Resolving the fibrotic
1223 niche of human liver cirrhosis at single-cell level. *Nature* 575: 512-518.
1224 doi:10.1038/s41586-019-1631-3.

1225 Rasala BA, Orjalo AV, Shen Z, Briggs S, Forbes DJ (2006) ELYS is a dual
1226 nucleoporin/kinetochore protein required for nuclear pore assembly and proper cell
1227 division. *Proc National Acad Sci* 103: 17801–17806. doi:10.1073/pnas.0608484103.

1228 Reske JJ, Wilson MR, Chandler RL (2020) ATAC-seq normalization method can
1229 significantly affect differential accessibility analysis and interpretation. *Epigenetics &*
1230 *chromatin* 13: 22. doi:10.1186/s13072-020-00342-y.

1231 Rigotti FDA, Wiedmann L, Hubert MO, Vacca M, Hasan SS, Moll I, Carvajal S,
1232 Jiménez W, Starostecka M, Billeter AT *et al* (2022) Semaphorin 3C exacerbates liver
1233 fibrosis. *Biorxiv*. 2021.07.29.454292. doi:10.1101/2021.07.29.454292.

1234 Roos B de, Rungapamestry V, Ross K, Rucklidge G, Reid M, Duncan G, Horgan G,
1235 Toomey S, Browne J, Loscher CE *et al* (2009) Attenuation of inflammation and
1236 cellular stress-related pathways maintains insulin sensitivity in obese type I
1237 interleukin-1 receptor knockout mice on a high-fat diet. *Proteomics* 9: 3244–3256.
1238 doi:10.1002/pmic.200800761.

1239 Salcedo R, Wasserman K, Young HA, Grimm MC, Howard OMZ, Anver MR, Kleinman
1240 HK, Murphy WJ, Oppenheim JJ (1999) Vascular Endothelial Growth Factor and Basic
1241 Fibroblast Growth Factor Induce Expression of CXCR4 on Human Endothelial Cells
1242 In Vivo Neovascularization Induced by Stromal-Derived Factor-1α. *Am J Pathology*
1243 154: 1125–1135. doi:10.1016/s0002-9440(10)65365-5.

1244 Salma M, Andrieu-Soler C, Deleuze V, Soler E (2023) High-throughput methods for
1245 the analysis of transcription factors and chromatin modifications: Low input, single
1246 cell and spatial genomic technologies. *Blood Cells Mol Dis* 101: 102745.
1247 doi:10.1016/j.bcmd.2023.102745.

1248 Sawitza I, Kordes C, Reister S, Häussinger D (2009) The niche of stellate cells within
1249 rat liver. *Hepatology* 50: 1617–1624. doi:10.1002/hep.23184.

1250 Schep AN, Wu B, Buenrostro JD, Greenleaf WJ (2017) chromVAR: inferring
1251 transcription-factor-associated accessibility from single-cell epigenomic data. *Nat*
1252 *Methods* 14: 975–978. doi:10.1038/nmeth.4401.

1253 Seo YY, Cho YK, Bae J-C, Seo MH, Park SE, Rhee E-J, Park C-Y, Oh K-W, Park S-
1254 W, Lee W-Y (2013) Tumor Necrosis Factor-α as a Predictor for the Development

1255 of Nonalcoholic Fatty Liver Disease: A 4-Year Follow-Up Study. *Endocrinol Metab*
1256 28: 41-45. doi:10.3803/enm.2013.28.1.41.

1257 Sturm G, Finotello F, Petitprez F, Zhang JD, Baumbach J, Fridman WH, List M,
1258 Aneichyk T (2019) Comprehensive evaluation of transcriptome-based cell-type
1259 quantification methods for immuno-oncology. *Bioinformatics (Oxford, England)* 35:
1260 i436-i445. doi:10.1093/bioinformatics/btz363.

1261 Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA,
1262 Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al* (2005) Gene set enrichment
1263 analysis: A knowledge-based approach for interpreting genome-wide expression
1264 profiles. *Proceedings of the National Academy of Sciences of the United States of*
1265 *America* 102: 15545-15550. doi:10.1073/pnas.0506580102.

1266 Sun J, Weis WI (2011) Biochemical and Structural Characterization of β-Catenin
1267 Interactions with Nonphosphorylated and CK2-Phosphorylated Lef-1. *J Mol Biol* 405:
1268 519-530. doi:10.1016/j.jmb.2010.11.010.

1269 Tarca AL, Draghici S, Bhatti G, Romero R (2012) Down-weighting overlapping genes
1270 improves gene set analysis. *Bmc Bioinformatics* 13: 136. doi:10.1186/1471-2105-13-
1271 136.

1272 Teschendorff AE, Breeze CE, Zheng SC, Beck S (2017) A comparison of reference-
1273 based algorithms for correcting cell-type heterogeneity in Epigenome-Wide
1274 Association Studies. *BMC Bioinformatics* 18: 105. doi:10.1186/s12859-017-1511-5.

1275 Teschendorff AE, Zheng SC (2017) Cell-type deconvolution in epigenome-wide
1276 association studies: a review and recommendations. *Epigenomics* 9: 757-768.
1277 doi:10.2217/epi-2016-0153.

1278 Tomfohr J, Lu J, Kepler TB (2005) Pathway level analysis of gene expression using
1279 singular value decomposition. *Bmc Bioinformatics* 6: 225-225. doi:10.1186/1471-
1280 2105-6-225.

1281 Traag VA, Waltman L, Eck NJ van (2019) From Louvain to Leiden: guaranteeing well-
1282 connected communities. *Sci Rep-uk* 9: 5233. doi:10.1038/s41598-019-41695-z.

1283 Tran MTN, Hamada M, Jeon H, Shiraishi R, Asano K, Hattori M, Nakamura M,
1284 Imamura Y, Tsunakawa Y, Fujii R *et al* (2017) MafB is a critical regulator of
1285 complement component C1q. *Nat Commun* 8: 1700. doi:10.1038/s41467-017-01711-
1286 0.

1287 Wang L, Trasanidis N, Wu T, Dong G, Hu M, Bauer DE, Pinello L (2022) Dictys:
1288 dynamic gene regulatory network dissects developmental continuum with single-cell
1289 multi-omics. *Biorxiv*. 2022.09.14.508036. doi:10.1101/2022.09.14.508036.

1290 Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P,
1291 Najafabadi HS, Lambert SA, Mann I, Cook K *et al* (2014) Determination and Inference
1292 of Eukaryotic Transcription Factor Sequence Specificity. *Cell* 158: 1431-1443.
1293 doi:10.1016/j.cell.2014.08.009.

1294 Wilson DR, Jin C, Ibrahim JG, Sun W (2019) ICeD-T Provides Accurate Estimates of
1295 Immune Cell Abundance in Tumor Samples by Allowing for Aberrant Gene
1296 Expression Patterns. *J Am Stat Assoc* 115: 1-23.
1297 doi:10.1080/01621459.2019.1654874.

1298 Xiong X, Kuang H, Ansari S, Liu T, Gong J, Wang S, Zhao X-Y, Ji Y, Li C, Guo L *et al*
1299 (2019) Landscape of Intercellular Crosstalk in Healthy and NASH Liver Revealed by
1300 Single-Cell Secretome Gene Analysis. *Molecular cell* 75: 644-660.e5.
1301 doi:10.1016/j.molcel.2019.07.028.

1302 Yagai T, Miyajima A, Tanaka M (2014) Semaphorin 3E Secreted by Damaged
1303 Hepatocytes Regulates the Sinusoidal Regeneration and Liver Fibrosis during Liver
1304 Regeneration. *Am J Pathology* 184: 2250-2259. doi:10.1016/j.ajpath.2014.04.018.

1305 Yang MH, Sung J, Gwak G-Y (2016) The associations between apolipoprotein B, A1,
1306 and the B/A1 ratio and nonalcoholic fatty liver disease in both normal-weight and
1307 overweight Korean population. *J Clin Lipidol* 10: 289-298.
1308 doi:10.1016/j.jacl.2015.11.017.

1309 Yip SH, Wang P, Kocher J-PA, Sham PC, Wang J (2017) Linnorm: improved
1310 statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Research*
1311 45: e179. doi:10.1093/nar/gkx828.

1312 Yu X, Chen X, Zheng XD, Zhang J, Zhao X, Liu Y, Zhang H, Zhang L, Yu H, Zhang M
1313 *et al* (2018) Growth Differentiation Factor 11 Promotes Abnormal Proliferation and
1314 Angiogenesis of Pulmonary Artery Endothelial Cells. *Hypertension* 71: 729–741.
1315 doi:10.1161/hypertensionaha.117.10350.

1316 Zhang C, Jia Y, Liu B, Wang G, Zhang Y (2020a) TLR4 knockout upregulates the
1317 expression of Mfn2 and PGC-1α in a high-fat diet and ischemia-reperfusion mice
1318 model of liver injury. *Life Sci* 254: 117762. doi:10.1016/j.lfs.2020.117762.

1319 Zhang F, Jiang WW, Li X, Qiu XY, Wu Z, Chi YJ, Cong X, Liu YL (2016) Role of
1320 intrahepatic B cells in non-alcoholic fatty liver disease by secreting pro-
1321 inflammatory cytokines and regulating intrahepatic T cells. *J Digest Dis* 17: 464–474.
1322 doi:10.1111/1751-2980.12362.

1323 Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C,
1324 Myers RM, Brown M, Li W *et al* (2008) Model-based Analysis of ChIP-Seq (MACS).
1325 *Genome Biol* 9: R137–R137. doi:10.1186/gb-2008-9-9-r137.

1326 Zhang Y, Lu Z, Zeng W, Zhao J, Zhou X (2020b) Two sides of NNMT in alcoholic and
1327 nonalcoholic fatty liver development. *J Hepatol* 74: 1250–1253.
1328 doi:10.1016/j.jhep.2020.11.049.

1329 Zhong L, Huang L, Xue Q, Liu C, Xu K, Shen W, Deng L (2019) Cell-specific elevation
1330 of Runx2 promotes hepatic infiltration of macrophages by upregulating MCP-1 in
1331 high-fat diet-induced mice NAFLD. *J Cell Biochem* 120: 11761–11774.
1332 doi:10.1002/jcb.28456.

1333 Zinzalla G (2016) A New Way Forward in Cancer Drug Discovery: Inhibiting the
1334 SWI/SNF Chromatin Remodelling Complex. *Chembiochem* 17: 677–682.
1335 doi:10.1002/cbic.201500565.

1336

Figure 1

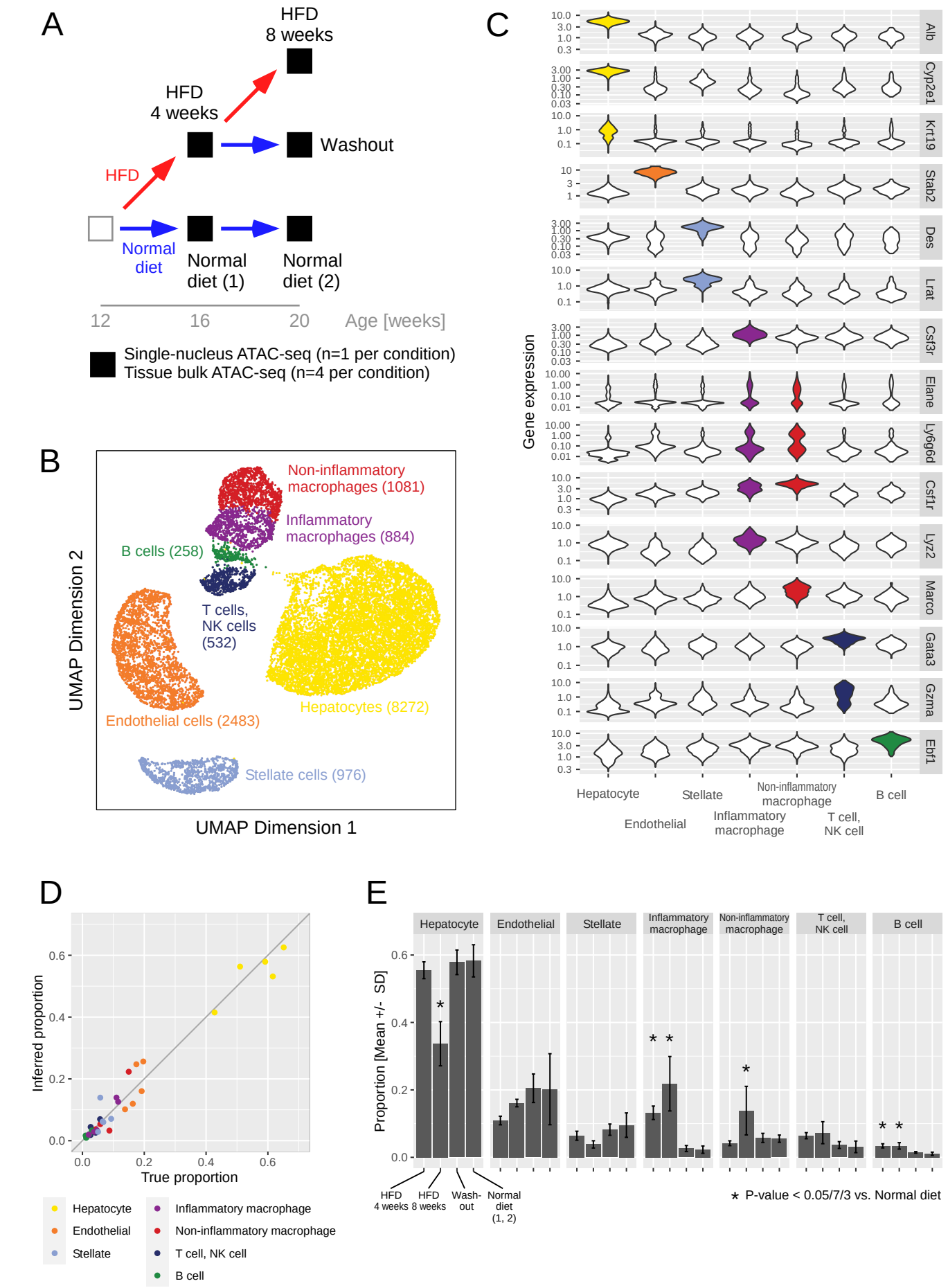

A
HFD
8 weeks
HFD
4 weeks
Washout
HFD
Normal
diet
Normal
diet (1)
Normal
diet (2)
12
16
20
Age [weeks]
Single-nucleus ATAC-seq (n=1 per condition)
Tissue bulk ATAC-seq (n=4 per condition)
B
Non-inflammatory
macrophages (1081)
Inflammatory
macrophages (884)
B cells (258)
T cells,
NK cells
(532)
Hepatocytes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

Figure 2

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

Compared to normal diet.

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

Figure 3. Hepatocyte

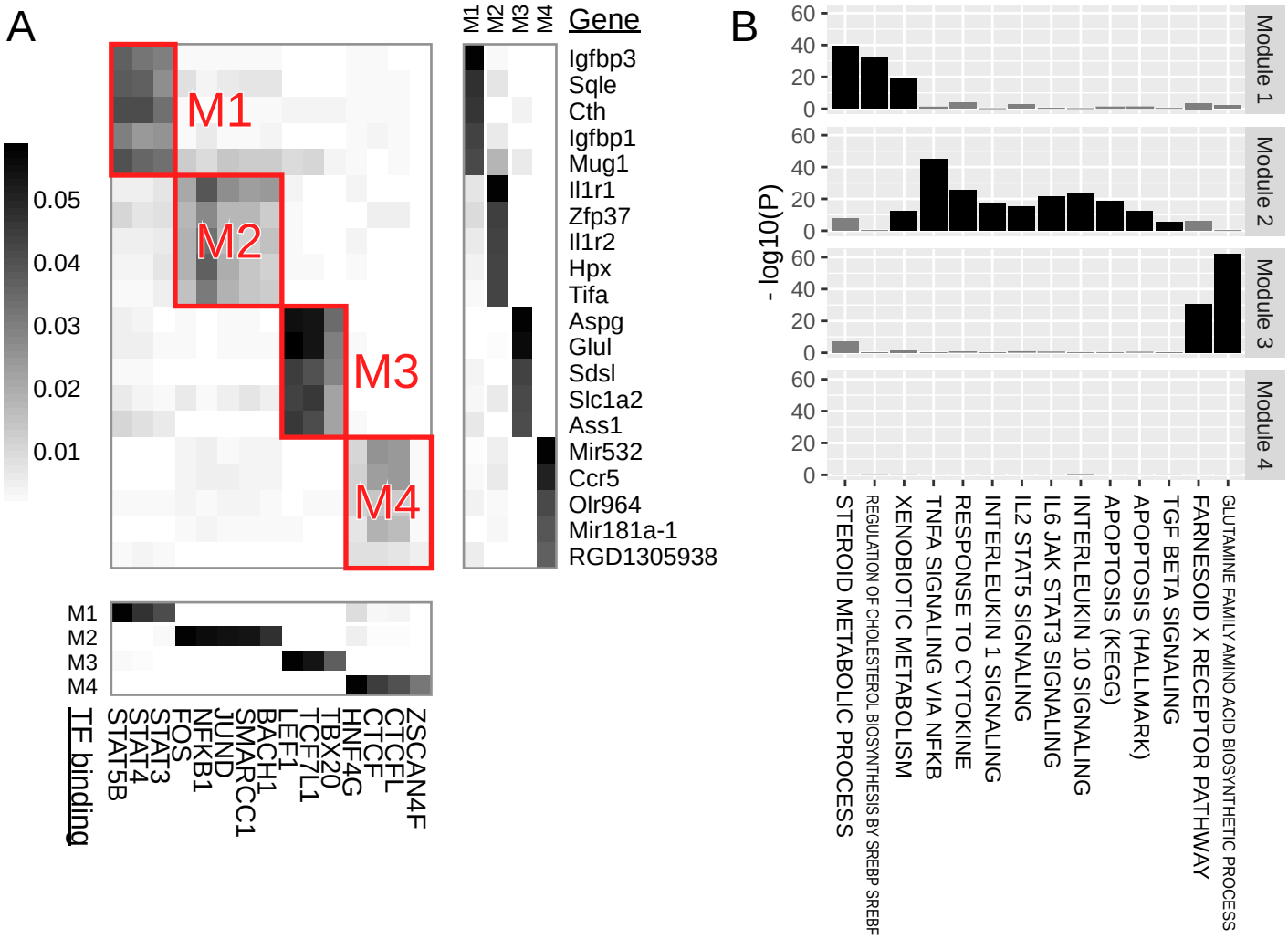


Figure 4. Endothelial

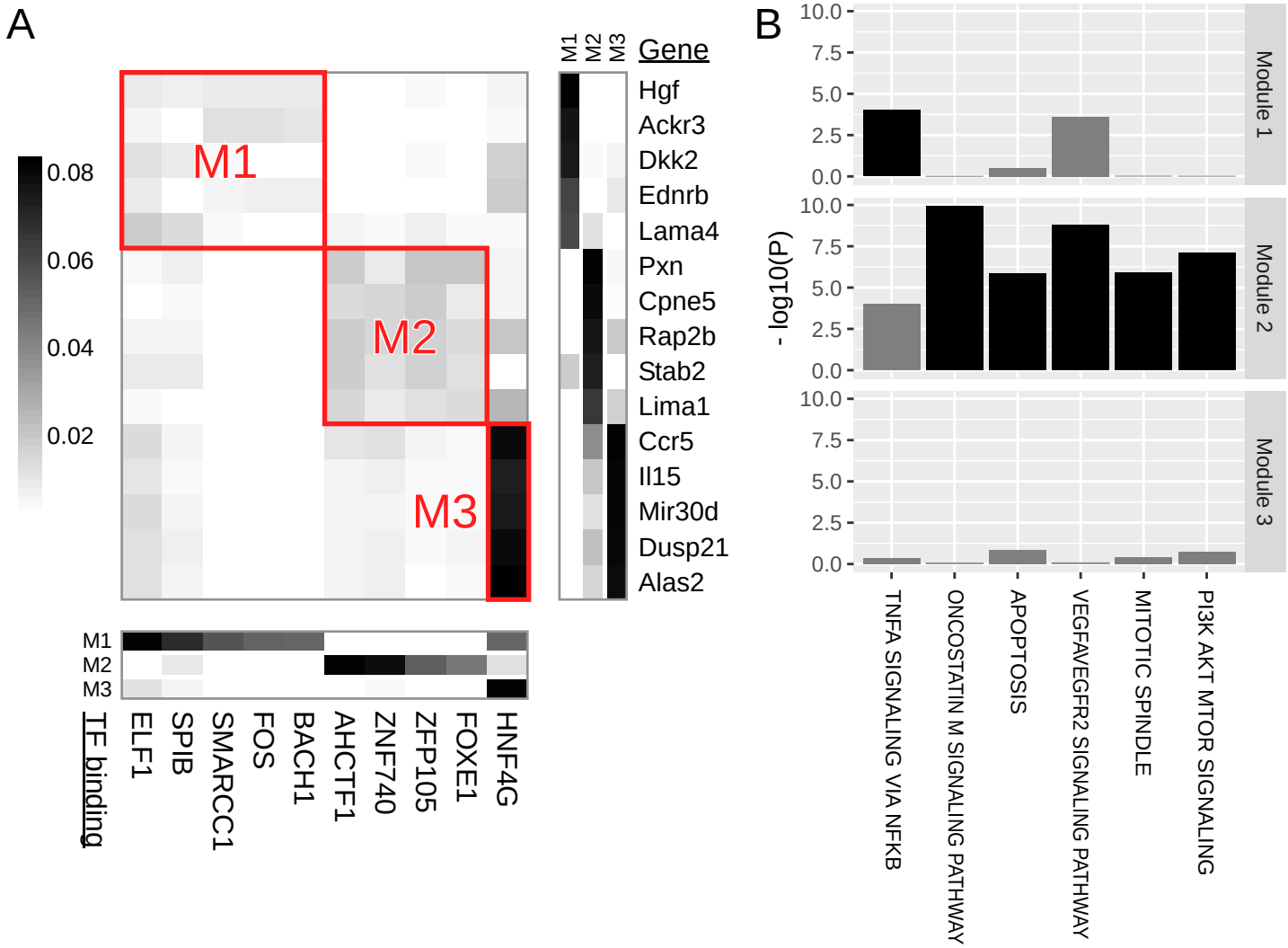


Figure 5. Stellate

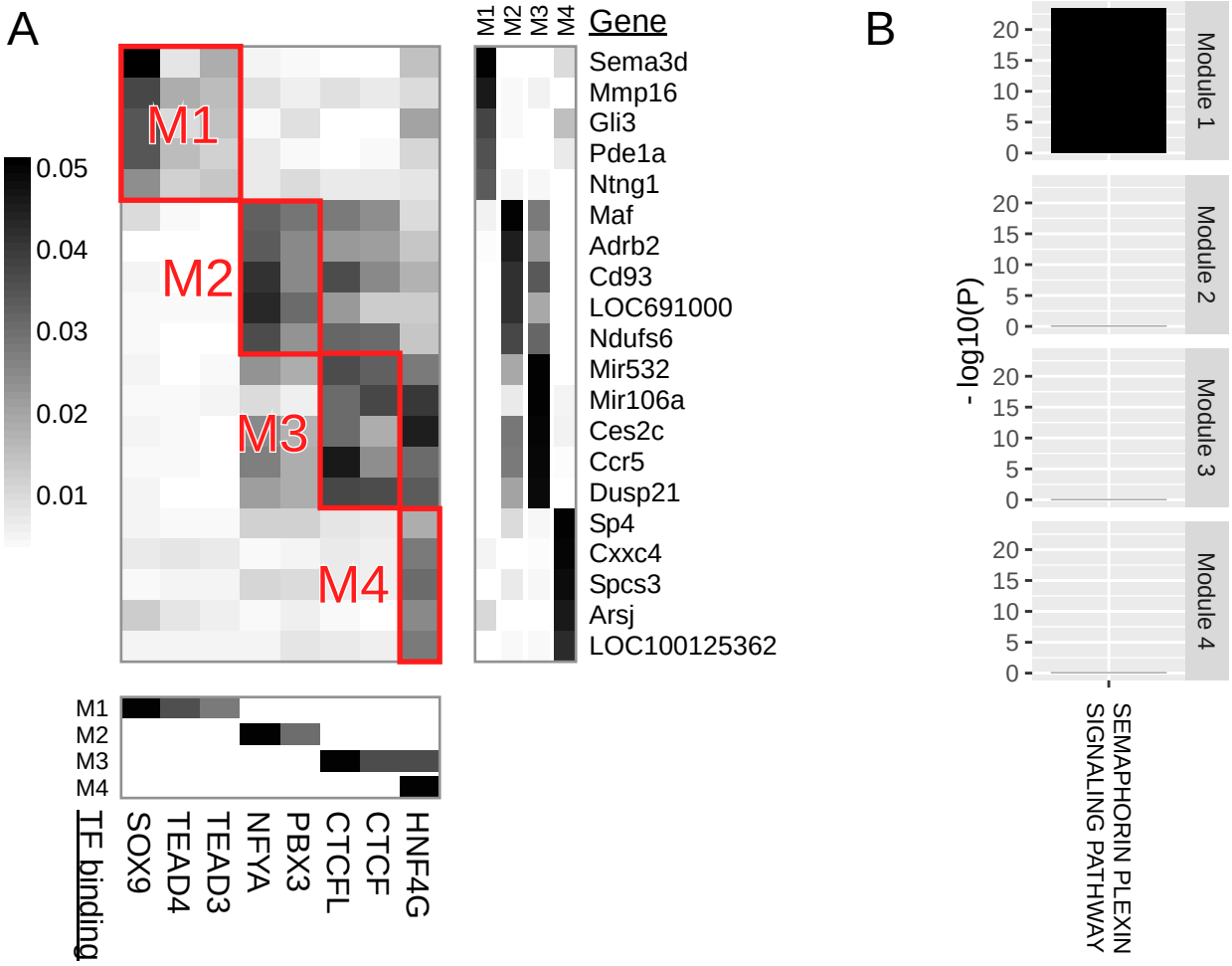


Figure 6. Macrophage

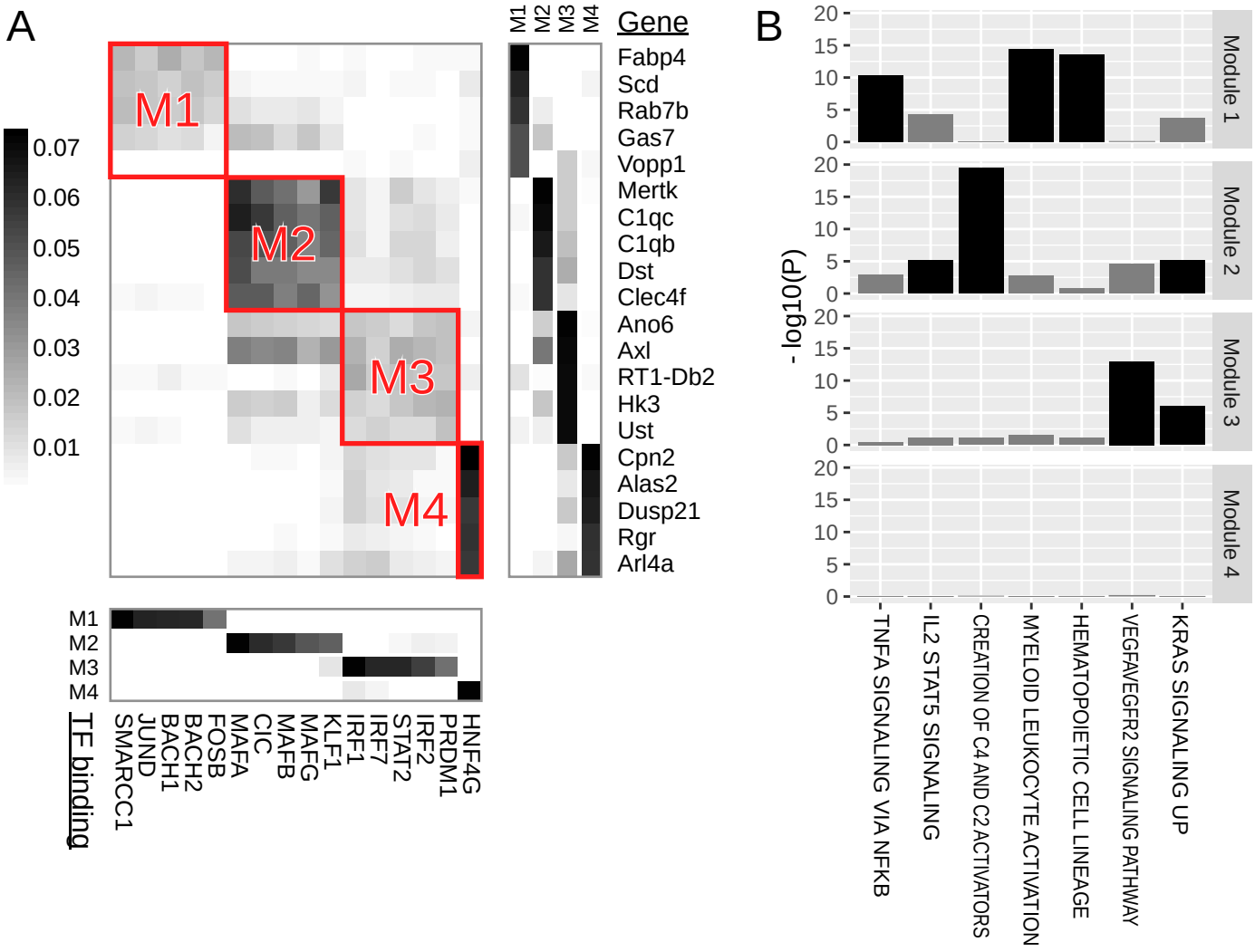


Figure 7

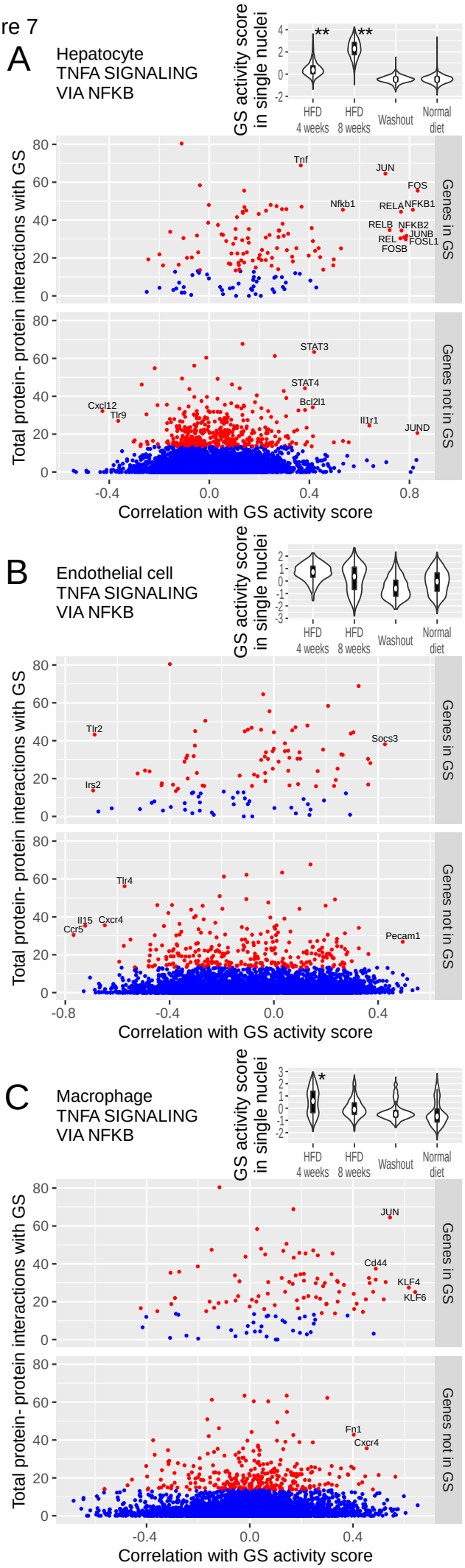
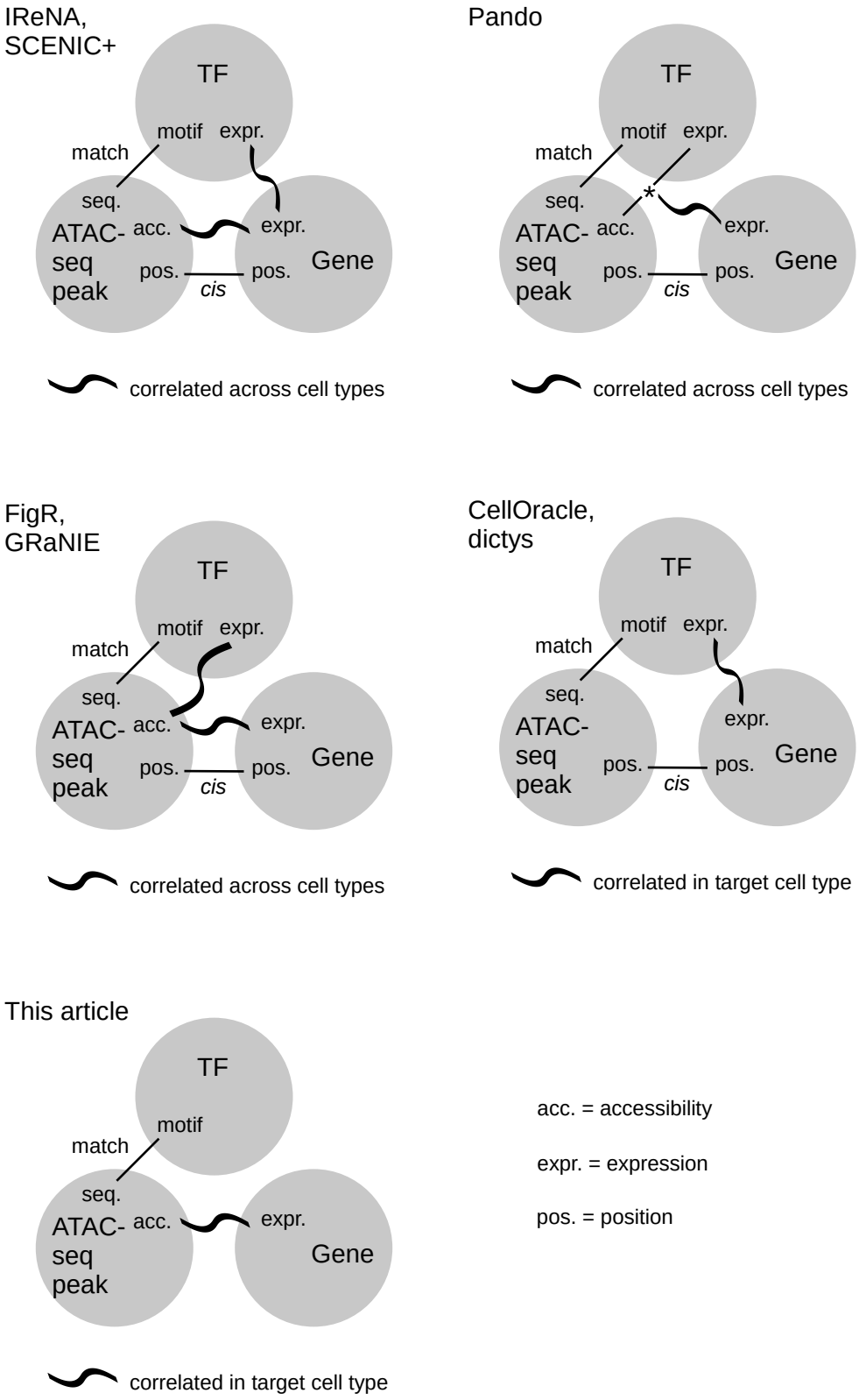

A
Hepatocyte
TNFA SIGNALING
VIA NFKB
GS activity score in single nuclei
HFD 4 weeks
HFD 8 weeks
Washout
Normal diet
Total protein- protein interactions with GS
Genes in GS
Tnf
JUN
FOS
Nfkb1
RELA
NFKB1
RELB
NFKB2
REL
JUNB
FOSL1
FOSB
Genes not in GS
STAT3
STAT4
Cxcl12
Tlr9
Bcl2l1
Il1r1
JUND
Correlation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

Figure 8

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

Figure 9


IReNA,
SCENIC+
TF
motif
expr.
match
seq.
ATAC-
seq
peak
acc.
expr.
pos.
pos.
Gene
cis
correlated across cell types
Pando
TF
motif
expr.
match
seq.
*
ATAC-
seq
peak
acc.
expr.
pos.
pos.
Gene
cis
correlated across cell types
FigR,
GRaNIE
TF
motif
expr.
match
seq.
ATAC-
seq
peak
acc.
expr.
pos.
pos.
Gene
cis
correlated across cell types
CellOracle,
dictys
TF
motif
expr.
match
seq.
ATAC-
seq
peak
expr.
pos.
pos.
Gene
cis
correlated in target cell type
This article
TF
motif
match
seq.
ATAC-
seq
peak
acc.
expr.
Gene
correlated in target cell type
acc. = accessibility
expr. = expression
pos. = position