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

Fumihiko Takeuchi¹² *, Yi-Qiang Liang¹, Hana Shimizu-Furusawa³, Masato Isono¹, Mia Yang Ang¹⁴, Kotaro Mori², Taizo Mori⁵, Eiji Kakazu⁵, Sachiyo Yoshio⁵, Norihiro Kato¹²⁴

¹ Department of Gene Diagnostics and Therapeutics, Research Institute, National Center for Global Health and Medicine, Japan.
² Medical Genomics Center, Research Institute, National Center for Global Health and Medicine, Japan.
³ Department of Hygiene and Public Health, School of Medicine, Teikyo University, Japan.
⁴ Department of Clinical Genome Informatics, Graduate School of Medicine, The University of Tokyo, Japan.
⁵ Department of Liver Diseases, The Research Center for Hepatitis and Immunology, National Center for Global Health and Medicine, Japan.

* Correspondence: fumihiko@takeuchi.name
SUMMARY

Non-alcoholic fatty liver disease (NAFLD) develops from fatty liver to steatohepatitis during which multiple cell types may play different roles. Aiming to understand tissue composition of cell types, their gene expression and global gene regulation in the development of NAFLD, we performed single-nucleus and bulk ATAC-seq on the liver of rats fed with a high-fat diet. By machine learning, we divided global gene expression into modules, such that transcription factors in a module regulate a set of genes in the same module. Consequently, many of the modules rediscovered known regulatory relationship between the transcription factors and biological processes. For the discovered biological processes, we searched core genes, which were defined as genes central regarding co-expression and protein-protein interaction. A large part of the core genes overlapped with previously implicated NAFLD genes. Our statistical methods help elucidate the global gene regulation in vivo as a combination of modules and discover core genes of the relevant biological processes.

Abbreviations

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

Non-alcoholic fatty liver disease (NAFLD) is a continuum of liver abnormalities ranging from non-alcoholic fatty liver (NAFL) to non-alcoholic steatohepatitis (NASH). It has a global prevalence of ~25% (Friedman et al., 2018). NAFLD can gradually develop from fatty liver, steatohepatitis and cirrhosis to hepatic cancer. No approved drugs are available for NAFLD, and biomarkers are currently insufficient. The liver is composed of multiple cell types, including hepatocytes, cholangiocytes, stellate cells, endothelial cells, and white blood cells, each of which plays a different role in disease development. Recent single-cell RNA-seq studies have uncovered cell type-specific gene expression changes in NAFLD (Chu et al., 2020).

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

Tissue composition of cell types changes during disease development in various diseases and affects the outcome. Cell type counting of solid tissue became feasible through the single-cell assay, which is still expensive. Instead, deconvolution, the statistical inference of cell type composition of tissue samples assayed in bulk, has been used for transcriptome and DNA methylation assays (Sturm et al., 2019; Teschendorff and Zheng, 2017). For ATAC-seq, Li and associates (Li et al., 2020) previously proposed an algorithm for deconvolution but no experimental validation has been done with single-cell data.
Aiming to understand the cell type composition and global TF regulation in NAFLD development, we performed single-nucleus and bulk ATAC-seq on the liver of rats fed or withdrawn from a high-fat diet (HFD). We first analyzed the composition of cell types in the liver and the gene expression separately by cell type, followed by experimental validation of ATAC-seq-based deconvolution. Then, to explore global TF regulation in each cell type, we developed a new statistical method, identifying major modules with each characterized by a set of biological processes. We also developed a statistical method to detect core genes for the individual biological processes. Taking snATAC-seq data of an in vivo experimental model as the input, our data-driven workflow proceeds top-down from global TF regulation to gene, i.e., step-wise search of major TF regulation modules, biological processes and core genes. The core genes thus identified overlapped well with the previously known NAFLD genes, providing experimental validation to our newly developed statistical methods.

RESULTS

Cell type composition of the liver

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

In order to catalog cell types in the liver, we performed snATAC-seq on one animal per dietary condition. We obtained a total of 14,486 nuclei (~3000 per sample), which were grouped into 16 clusters by the similarity of chromatin opening. Hepatocytes (seven clusters), endothelial cells (three clusters), stellate cells (two clusters) and white blood cells (four clusters)
The clusters for white blood cells corresponded to inflammatory macrophages, non-inflammatory macrophages, T cells and NK cells combined, and B cells. Cholangiocytes were included in the hepatocyte clusters, and neutrophils and Kupffer cells were included in the macrophage clusters.

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

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

We analyzed cell type-specific differential gene expression by comparing the nuclei of the same cell type between dietary conditions (Table S2). Hereafter,
clusters for inflammatory and non-inflammatory macrophages are combined
and regarded as the macrophage cell type. T cells, NK cells, and B cells were
not analyzed because the number of nuclei was limited. The biological
processes characteristic of the differentially expressed genes were cell type-
dependent (Figure 2). Steroid metabolism and fatty acid metabolism genes
were expressed in hepatocytes differently under HFD and washout compared
with the normal diet. As of inflammation genes, the expression altered in
macrophages after 4 weeks of HFD and washout. The inflammation genes
and apoptosis genes were expressed in hepatocytes differently after 8 weeks
of HFD. Genes for the actin filament-based process were expressed in stellate
cells differently after 8 weeks of HFD.

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

Major modules of TF regulation and associated biological processes
To elucidate the global picture of TF regulation, we inferred the regulatory
impact of all TFs on all genes and extracted major modules of the regulatory
relation. By snATAC-seq, we measured in all nuclei of each cell type the
genome-wide binding of TFs and the expression of genes, and inferred the regulation of the latter by the former by applying machine learning. The regulatory relation was divided into modules. A module comprises a subset of TFs and a subset of genes, such that TFs belonging to a module regulate genes belonging to the same module but do not regulate genes belonging to other modules. In hepatocytes we obtained four modules (Figure 3A). Modules are simply a grouping of TFs/genes detected by machine learning and have no biological meaning by themselves. By applying gene set enrichment analysis to the genes belonging to a module, but not using the information on TFs, we searched for biological processes that characterize the module (Figure 3B). In the following, we describe modules for hepatocytes, endothelial cells (Figure 4), stellate cells (Figure 5), and macrophages (Figure 6). If a biological process was assigned to a module, we sought the literature that supports the regulation of the biological process by the TFs belonging to the module. Among five of the seven modules, there was support for the discovered linkage between the TFs and biological processes, suggesting the validity of the module discovery algorithm.

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

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

Module 3 of hepatocytes was characterized by the binding of TCF/LEF family TFs (namely, LEF1 and TCF7L1) and by genes in the farnesoid X receptor pathway for bile acid synthesis and genes for the glutamine family...
amino acid biosynthesis. Within the liver lobule, a unit that assembles to form the liver, a central vein is located in the center, the portal triad (portal vein, hepatic artery, and bile duct) is located in the periphery, and the metabolism of hepatocyte in between differs between the pericentral and periportal zones. The glutamine family amino acid biosynthesis genes include Glul and Ass1, which are highly expressed in the pericentral and periportal zones, respectively (Halpern et al., 2017). In addition, the LEF1 TF binds to β-catenin protein and activates Wnt signaling pathway (Sun and Weis, 2011), and liver-specific β-catenin KO mice exhibited defective cholesterol and bile acid metabolism in the liver (Behari et al., 2010). Moreover, Wnt-activated genes are highly expressed in the pericentral zone. Thus, LEF1 TF presumably regulates bile acid metabolism and zonated gene expression in the liver.

Module 2 of endothelial cells was characterized by the DNA binding of AHCTF1 (aka ELYS) and ZNF740 and by genes for angiogenesis. AHCTF1 is a DNA-binding protein that is required in mitosis (Rasala et al., 2006). ZNF740 activates angiogenesis in pulmonary artery endothelial cells of rats (Yu et al., 2018).

Module 1 of stellate cells was characterized by SOX9 TF binding and genes in the semaphorin-plexin signaling pathway.

Module 2 of macrophages was characterized by the binding of Maf family TFs (namely, MAFA, MAFB and MAFG) and by genes for complement system activation. In Mafb-deficient macrophages of mice, C1q production decreased (Tran et al., 2017).

Module 3 of macrophages was characterized by the binding of IRF family TFs (namely, IRF1, IRF2 and IRF7), which regulate immune responses, and by genes for angiogenesis. IRF1 contributes to the commitment of pro-inflammatory M1 macrophages, which produce angiogenic stimulators (Chistiakov et al., 2018).
Core genes in specific biological processes

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

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

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

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

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

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

To investigate lobular zonation genes in hepatocytes, the GS was taken from the list of zonation genes in normal mice (q-value < $10^{-6}$ in Table S3 in (Halpern et al., 2017)). The GS activity score showed a modest difference between dietary conditions (Figure S2B), which agrees with zonation being a normal phenomenon unrelated to NAFLD. When classified
by the correlation with the GS activity score, the genes with correlation coefficient > 0.3 were all pericentral and those with correlation coefficient < -0.3 were all periportal, showing that the GS activity score rediscovered zonation.

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

**DISCUSSION**

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

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

In each cell type, we can decompose the global variation of TF regulation into major modules, where the TFs belonging to a given module appear to regulate a set of genes that belong to the same module. Using the data for TF binding and gene expression measured over nuclei, we extract modules by applying the machine learning technique. In the discovered modules, the TFs and the function of regulated genes agree with the literature: e.g., the AP-1 TFs that regulate inflammation genes (Module 2 of hepatocytes, Module 1 of endothelial cells and Module 1 of macrophages) are known to respond to cytokine stimuli (Hess et al., 2004), and LEF1 that regulates zonation (Module 3 of hepatocytes) is also known to activate Wnt signaling (Sun and Weis, 2011). Previous snATAC-seq studies have characterized the difference between cell types (Cusanovich et al., 2018) using TF binding. In contrast, we illustrated how TFs control the variation of cells within one cell type, for example, the hepatocytes with various degrees of inflammation.

One novel method developed in this article is the computation of the GS activity score in single cells to indicate the activity of GS, as if it was the expression of a specific gene. In experiments, gene expressions can be directly measured. The activity of a pathway is of interest but not directly measurable. One way would be to take the average expression of genes in that pathway, but this is not optimal for the following reasons. A simple average would give equal weight to both core and periphery genes in the pathway, increasing noise from the latter. Moreover, there are activator and suppressor genes that contribute to the pathway in opposite directions.
Instead of simple average, we used the first principal component to deal with these issues. The principal component approach was adapted from the PLAGE program (Tomfohr et al., 2005), which is used to compute pathway activity scores in bulk tissue specimens. The GS activity score could improve the understanding of biological processes. For example, Abcg1 and Sqle correlated positively and negatively, respectively, with the activity score for the steroid metabolism pathway (Figure S2A). As this score became higher under the HFD, the GS activity represented by the score presumably works in response to the overdose of cholesterol contained in the diet. Thus, a positive correlation of Abcg1 (for lipid efflux) and a negative correlation of Sqle (for sterol biosynthesis) are reasonable. Moreover, the cell type-specific nature of the GS activity score could unravel cell type-specific dynamics; the inflammation process peaked under 4 weeks of HFD in macrophages and under 8 weeks of HFD in hepatocytes (Figure 7). The GS activity score can also be applied to single-cell and single-nucleus RNA-seq data.

Another novel method in this article is finding the core genes of biological processes. Given a GS comprising known genes for the process, we identified core genes as those that show strong correlation with the GS activity score as well as abundant protein-protein interactions with the GS. We obtained evidence supporting the former from single-nucleus experiment data and the latter from a public database. We sought literature indicating that core genes cause NAFLD or serve as biomarkers for NAFLD and found indications for a large part of the core genes (see below): for the TNFα signaling via NF-κB process, 4 of 5 in hepatocytes, 4 of 8 in endothelial cells, and 1 of 3 in macrophages; for steroid metabolism 3 of 9 in hepatocytes; for semaphorin-plexin signaling 3 of 7 in stellate cells. The overlap with known NAFLD genes suggests the biological validity of our data-driven approach, using snATAC-seq data from a rat model of dietary intervention as input, extracting modules, biological processes, and then core genes. The non-
overlapping genes can be candidates for future functional study of NAFLD. Regarding TNFα signaling via NF-κB in hepatocytes, in animal studies, Tnf KO mouse was protected from NAFLD (Kakino et al., 2017), Nfkb1 KO mouse exacerbated NAFLD (Locatelli et al., 2012), and Il1r1 KO mouse had less hepatic triacylglycerol (Roos et al., 2009). CXCL12 (aka SDF1) induced migration of stellate cells (Sawitza et al., 2009), endothelial cells (Salcedo et al., 1999) and CD4+ T cell (Boujedidi et al., 2014). In human biomarker studies, higher circulating TNFα level was associated with NAFLD (Potoupni et al., 2021) and future NAFLD (Seo et al., 2013). In NAFLD patients, soluble IL1R1 was higher (Ajmera et al., 2017). Regarding TNFα signaling via NF-κB in endothelial cells, in animal studies, Pecam1 KO mouse developed NASH (Goel et al., 2007), whereas Tlr4 KO mouse (Zhang et al., 2020a) and Il15 KO mouse (Cepero-Donates et al., 2016) were protected from NAFLD. Ccr5 antagonist eased NAFLD in mice (Pérez-Martínez et al., 2014). Regarding TNFα signaling via NF-κB in macrophages, Cd44 KO mouse had less liver injury under NAFLD inducing diet, and NAFLD patients had elevated serum soluble CD44 (Patouraux et al., 2017). Regarding steroid metabolism in hepatocytes, Scd1 KO mouse exacerbated NAFLD (Li et al., 2009), and Acox2 KO mouse developed NAFLD (Zhang et al., 2020b). NAFLD patients had lower serum APOA1 (Chen et al., 2022; Yang et al., 2016). Regarding semaphorin-plexin signaling in stellate cells, in animal studies, Nrp2 conditional KO reduced liver fibrosis in mice (Rigotti et al., 2022). Nrp1 KO attenuated cell motility and overexpression increased motility (Cao et al., 2010). Sema3e KO eased liver fibrosis in mice (Yagai et al., 2014).

In search of drug targets or biomarkers of NAFLD, we looked up proteins whose subcellular localization is extracellular or plasma membrane proteins (Binder et al., 2014) among the genes strongly upregulated under a dietary intervention (Table S2) or the core genes positively correlated with a NAFLD-related activity. A number of genes have been reported as NAFLD
biomarkers in humans. For hepatocytes, we found Il1r1 (Ajmera et al., 2017), Il1r2 and Tnf (Potoupni et al., 2021). For endothelial cells, there was Hgf (An et al., 2019; Koutsogiannis et al., 2010; Krawczyk et al., 2017). Pecam1 was listed for both endothelial and stellate cells. For stellate cells, there were Faslg (Page et al., 2013) and Nrp2. For macrophages, there were RT1-Db1, Cd44 (Patouraux et al., 2017), Cxcr4 and Fn1.

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

By performing single-nucleus and bulk ATAC-seq, we analyzed the transition of cell type composition and cell type-specific gene expression in a rat model of NAFLD. Using novel statistical methods, we elucidated a global picture of in vivo TF regulation in each cell type as a set of modules and discovered core genes for NAFLD-relevant biological processes. The core genes can serve as candidates for novel NAFLD genes and our statistical methods can be applied to elucidate the TF regulation in other biological systems.
Acknowledgments
The authors thank the Center for Omics and Bioinformatics, Graduate School of Frontier Sciences, University of Tokyo and Azenta for their assistance in snATAC-seq. The authors thank Enago (www.enago.jp) for the English language review. This work was supported by the NCGM Intramural Research Fund (20A1013) and by JSPS KAKENHI [grant number JP16K07218].

Declaration of interests
The authors declare no competing interests.
FIGURE LEGENDS

Figure 1
Cell types observed in the liver and their composition
(A) Experimental design.
(B) UMAP plot of the observed nuclei. The number of observed nuclei is shown in parentheses.
(C) Expression level of cell type marker genes in nuclei (vertical axis) is compared between nuclei clusters (horizontal axis) in violin plots.
(D) Validation of deconvolution based on ATAC-seq using the samples assayed in both single-nucleus and bulk ATAC-seq. The true cell type proportion counted in snATAC-seq (horizontal axis) is compared with the proportion inferred based on bulk ATAC-seq (vertical axis).
(E) Proportion of nuclei from a specific cell type (vertical axis) is compared between dietary conditions (horizontal axis). The proportion was measured by bulk tissue ATAC-seq (n=4 samples per condition).

Figure 2
Biological processes that were differentially expressed under dietary interventions
The heatmap shows whether a biological process (row) was differentially expressed in a cell type under a dietary intervention compared to the normal diet (column). The coloring indicates statistical significance of gene set enrichment analysis: red if significant adjusting for multiple testing, brown if nominally significant at P<0.01, light brown if P<0.05, and white otherwise.

Figure 3
Major modules of TF regulation in hepatocytes
(A) The central heatmap shows the regulatory impact of TF binding (columns) on genes (rows). Major modules are highlighted by red rectangles. The
heatmap at the bottom shows the membership of TFs (columns) to modules (rows). The heatmap on the right shows the membership of genes (rows) to modules (columns). The heatmaps only include the TFs GENES with membership ranking in the top five and membership weight more than half of the top.

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

**Figure 4**
Major modules of TF regulation in endothelial cells
See legend in Figure 3.

**Figure 5**
Major modules of TF regulation in stellate cells
See legend in Figure 3.

**Figure 6**
Major modules of TF regulation in macrophages
See legend in Figure 3.

**Figure 7**
Core genes for the TNFα signaling via NF-κB processes are extracted based on correlation with GS activity score and protein-protein interaction (A) Analysis in hepatocytes using the GS for TNFα signaling via NF-κB. In the top panel, the GS activity score in single nuclei (vertical axis) is compared
between dietary conditions (horizontal axis) in violin plots. Box plots within
the violins indicate quartiles. The score distribution under a diet intervention
was compared with that under normal diet condition by the area under the
receiver operating characteristic curve: * and ** indicate larger than 0.75
and 0.9, respectively. In the remaining panels, TFs and genes are plotted by
the correlation with the GS activity score (horizontal axis) and the total
amount of protein-protein interactions with the GS (vertical axis). The points
for TFs/genes having protein-protein interactions within the top 5% are
colored red, and others are colored blue. The TFs/genes in the GS (middle
panel) and not in the GS (bottom panel) are plotted separately. TFs are
labeled in all capital letters. The GS was taken from the MSigDB Hallmark
database. Protein-protein interactions were taken from the STRING database
(B) Analysis in endothelial cells.
(C) Analysis in macrophages.
METHODS

Data and code availability

The single-nucleus and bulk ATAC-seq raw data generated during this study are available at DDBJ (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.

Animal experiment

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

Library preparation and sequencing for snATAC-seq

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

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

Measuring triglycerides in liver tissue

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

Measuring total cholesterol in liver tissue

Total cholesterol in liver tissue was measured enzymatically using the Total Cholesterol and Cholesteryl Ester Colorimetric Assay Kit II (BioVision, K623-100). Tissue homogenate was obtained by dispersing 15 mg tissue in 600 ul CHCl₃:IPA:NP-40 (7:11:0.1) at 2200 rpm for 1 min, using a handheld disperser (KINEMATICA, PT 1300 D). The homogenate was then centrifuged for 5 min at 15,000 g. Thereafter, we transferred 200 ul of the liquid phase to a new tube, dried with air-flow-in at 40 C for 10 min using a centrifugal concentrator (TOMY Digital Biology, CC-105), and dissolved the dried lipids with 400 ul Cholesterol Assay Buffer. Total cholesterol was measured in a colorimetric assay according to the manual.
Data cleaning for snATAC-seq

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

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

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

For each nucleus, the proportion of fragment counts from a chromosome (e.g., chromosome 1) is approximately constant, being proportional to the chromosome length, but could deviate if a droplet is contaminated by free DNA from broken nuclei (Orchard et al., 2020).
Therefore, we represented the proportion by a matrix indexed by nuclei (rows) and chromosomes (columns). To standardize the matrix, we subtracted column-wise averages and divided by the standard deviation. By taking the sum of the absolute values of a row, we defined the deviation score of a nucleus. In order to omit contaminated droplets, we discarded nuclei with a score ≥40.

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

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

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

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

Algorithm for dimensionality reduction and nucleus clustering

The tile and peak matrices are large binary matrices indexed by genomic positions (rows) and nuclei (columns). Since the column vectors are extremely high dimensional, we need to reduce the dimensions in order to properly analyze the similarity among nuclei. For dimensionality reduction, we used latent semantic indexing (Berry and Browne, 2005) and singular value decomposition (SVD). The dimensionality reduction was performed for all samples combined. Since experimental conditions, including Tn5:DNA ratio,
could differ among the samples, we introduced a sample batch correction step. We applied the same algorithm to the tile and peak matrices.

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

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

For the \(i\)-th row of the matrix weighted by IDF, the variance becomes

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

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

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

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

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

When the denominator equals zero, \(\gamma_{ij}\) is set as one. The above correction requires no knowledge of cell types. If cell type classification of nuclei is available, the correction can be refined. For the nucleus \(j\) of cell type \(c\) belonging to sample \(s\), the correction factor becomes

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

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

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

1. Compute sample batch correction factor $\gamma_{ij}$ in (1).

2. Compute the matrix $g_i \cdot d_j \cdot f_{ij} \cdot \gamma_{ij}$ and centralize each row by subtracting the average.

3. Compute truncated SVD, $UDV^T$, for the 30 largest singular values using the R package irlba (version 2.3.3) (Baglama and Reichel, 2005).

4. Take truncated SVD for the two largest singular values, project the nuclei to the $VD$ coordinates, and detect and discard outlier nuclei. Outliers were detected based on the sum of the squared distance to the five nearest neighbors.

5. Take truncated SVD for the non-noise singular values, and project the nuclei to the $VD$ coordinates. As the 11th and later singular values bottomed out and the corresponding columns of $V$ appeared randomly distributed in our data, we took the ten largest singular values.

6. Identify clusters of nuclei by computing the shared nearest-neighbor graph with the R package Seurat (version 4.0.0) (Hao et al., 2021) and optimizing modularity with the Leiden algorithm (version 0.8.4) (Traag et al., 2019). By merging the clusters, define coarse clustering that clearly separates in the $VD$ coordinates. In our data, the coarse clusters corresponded to hepatocytes, stellate cells, endothelial cells and white blood cells.

7. Compute sample batch correction factor $\gamma_{ij}$ in (2) by incorporating the coarse clustering of nuclei.
8. Compute the matrix \( (g_i \cdot d_j \cdot f_{ij} \cdot \gamma_{ij}) \) and centralize each row by subtracting the average.

9. Compute truncated SVD, \( UDV^T \), for the 30 largest singular values.

10. Take the truncated SVD for the non-noise singular values, and project the nuclei to the \( VD \) coordinates. We took the eight largest singular values.

11. Identify clusters of nuclei by the shared nearest neighbor modularity optimization.

**Cell type assignment for snATAC-seq**

For each nuclei cluster obtained above, we assigned a cell type based on chromatin accessibility at known marker genes. For each cluster, chromatin accessibility at the TSS and gene body of the genes were visually inspected using the browser of ArchR. We used the following marker genes in the rat:

- \( Alb, Apoc3, Cyp2e1 \) and \( Cyp2f4 \) for hepatocyte
- \( Epcam \) and \( Krt19 \) for cholangiocyte
- \( Stab2 \) for endothelial cell
- \( Dcn, Des, Lrat \) and \( Acta2 \) for stellate cell
- \( Cf3r \) for granulocyte
- \( Elane, Ly6g6c, Ly6g6d, Ly6g6e \) and \( Mpo \) for neutrophil
- \( Csf1r \) for macrophage (including Kupffer cell)
- \( Lyz2 \) for inflammatory macrophage
- \( Marco \) for non-inflammatory macrophage
- \( Cd3g \) and \( Gata3 \) for T cell
- \( Gzma \) and \( Prf1 \) for cytotoxic T cell and NK cell
- \( Ebf1 \) for B cell

**Data cleaning for bulk ATAC-seq**

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

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

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

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

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

We factorize the count as

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

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

We used the DeconPeaker software (version 1.0) (Li et al., 2020) to select the peaks (rows) that discriminate cell types. DeconPeaker selects
discriminating peaks based on t-tests while minimizing the conditional number. Let $\tilde{v}_{jk}$ be the vector truncated from $v_{jk}$ accordingly. The vectors $\tilde{v}_{jk}$ across the samples from the same cell type $j$ are approximately equal, and their average becomes the reference vector $w_j$ for chromatin accessibility:

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

We represent by $W$ the matrix whose columns are $w_j$. By regarding the samples as replicates, the error variance for peak $i$ was estimated as $e_i^2 = \text{mean}_j\{\text{var}_k\{\tilde{v}_{ijk}\}\}$.

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 $a_{jk}/(\sum_j a_{jk})$, and the mixture becomes

$$x_k = \sum_j \frac{a_{jk}}{(\sum_j a_{jk})} \cdot \frac{1}{c_j} w_j.$$  

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 $y_k$ for sample $k$. The elements $y_{ik}$ are taken only for the peaks $i$ that were selected by DeconPeaker. Five samples were assayed both in bulk and in single nuclei, for which we have vectors $x_k$ and $y_k$. Although, in principle, the two vectors should be proportional, the ratio $y_{ik}/x_{ik}$ actually varies by peaks, due to measurement differences between different assay technologies. Correction of between-peak variability can improve deconvolution (Jew et al., 2020). We compute the correction factor

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

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

For deconvolution of a bulk sample $k$, we first compute $\theta_k$ that approximates

$$F^{-1}y_k \approx W\theta_k$$

under $\theta_{jk} \geq 0$. Using robust linear regression (Teschendorff et al., 2017), we solve

$$F^{-1}y_k = W\theta_k + \epsilon,$$

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

where $\sigma^2$ 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 $\hat{\theta}_k$ are set to zero. When computing $\theta_k$, the independent variable $F^{-1}y_k$ was scaled to one nucleus, and the regressors $w_j$ were scale to $c_j$ nuclei. We multiply $\hat{\theta}_k$ by $C$ and obtain $C\hat{\theta}_k$, which is the fractional number of nuclei from each cell types included per nucleus of sample $k$, which in spirit is the cell type composition. Although the sum of the elements in $C\hat{\theta}_k$ was approximately equal to one, in order to be precise, we rescale to obtain the cell type composition for sample $k$ as

$$\frac{C\hat{\theta}_k}{\sum_j c_j \hat{\theta}_{jk}}.$$

**Comparing cell type composition among dietary conditions**

For each cell type, the imputed composition in bulk samples was compared among dietary conditions. We performed multiple regression where the dependent variable was the composition and independent variables were indicators of dietary condition, with the normal diet being the reference. Using Bonferroni’s correction for multiple testing of seven cell types and three...
conditions (apart from normal diet), the significance level was set to 0.05/[$7 \times 3$].

**Quantifying gene expression in snATAC-seq**

For each nucleus, the expression level of a gene was inferred from the number of Tn5 insertions in the gene body and vicinity. The targeted biological phenomenon is the chromatin opening of the gene rather than the amount of mRNA in a nucleus or cell. Using ArchR, we computed the Gene Score Matrix, which is indexed by genes (rows) and nuclei (columns) and represents the insertion counts. As the matrix is analogous to the count matrix obtained in single-cell RNA-seq, we applied library size normalization of the R package Linnorm (version 2.16.0) (Yip et al., 2017) and imputation of the R package SAVER (version 1.1.2) (Huang et al., 2018). Before imputation, 47% of the matrix elements were zero, and after imputation, 0.2% were set to the lowest value (0.001). We excluded low expression genes, where the lowest value occurred in $\geq 1\%$ of nuclei. We took 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 TF, Tn5 insertion in the peak suggests the binding of the corresponding TF. For each nucleus, by aggregating over all peaks, we can infer the genome-wide binding of the TF. This is obtained as the chromVAR (Schep et al., 2017) deviation z-score in ArchR. The CIS-BP database of TF binding motifs
(Weirauch et al., 2014) was used. In the same way as we did for the Gene Score Matrix, we applied quantile normalization to the Motif Matrix and subtracted the first two components of SVD.

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

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

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

To elucidate a global picture of TF regulation, we inferred the regulatory impacts of all TFs on all genes and then extracted major modules of the regulatory relationship between TFs and genes. For each gene, we computed its regulatory TFs using the R package GENIE3 (version 1.14.0) (Huynh-Thu
et al., 2010), which implements a random forest algorithm and thus can model nonlinear regulation. The algorithm aimed to predict the expression level of the gene based on the binding of TFs and dietary conditions. GENIE3 quantified the impacts of regulators by the proportion of explained variance. By combining the results from all genes, we obtained a regulator-regulatee matrix, in which the rows correspond to TFs and dietary conditions, the columns correspond to genes, and an element represents the importance of a TF to a gene.

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

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

**GS activity score in single nuclei**

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

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

The principal component approach was adapted from the PLAGE program (Tomfohr et al., 2005) that was developed to compute the pathway activity score of a GS in bulk tissue specimens. Instead, we applied to nuclei. Compared with the number of specimens, the large number of nuclei enabled significance testing by permutation.
REFERENCES


https://doi.org/10.1093/jac/dku071.


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


Zhang, F., Jiang, W.W., Li, X., Qi, X.Y., Wu, Z., Chi, Y.J., Cong, X., and Liu, Y.L. (2016). Role of intrahepatic B cells in non-alcoholic fatty liver disease by secreting pro-inflammatory...
cytokines and regulating intrahepatic T cells. J Digest Dis 17, 464–474.
https://doi.org/10.1111/1751-2980.12362.


Figure 1

(A) Schematic representation of the experimental design. HFD (high-fat diet) for 4 weeks or 8 weeks, followed by washout.

(B) UMAP visualization of single-cell RNA-seq data. UMAP Dimension 1 and UMAP Dimension 2. Different cell types: Hepatocytes, Endothelial cells, Stellate cells, Inflammatory macrophages, Non-inflammatory macrophages, T cells, NK cells, and B cells.

(C) Gene expression heatmap. Different cell types and conditions.

(D) Plot showing the relationship between true proportion and inferred proportion. Different color codes for different cell types.

(E) Bar graph showing the proportion of different cell types across different conditions. HFD 4 weeks, HFD 8 weeks, Washout, Normal diet (1), Normal diet (2). Significant differences indicated by asterisks. * P-value < 0.05/7/3 vs. Normal diet.
### Figure 2

Compared to normal diet.

<table>
<thead>
<tr>
<th>Database</th>
<th>Gene set</th>
<th>Hepatocyte</th>
<th>Endothelial cell</th>
<th>Stellate cell</th>
<th>Macrophage</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEGG</td>
<td>PPAR_SIGNALING_PATHWAY</td>
<td>HFD 4w: 1, HFD 8w: 2, HFD wash out: 3</td>
<td>HFD 4w: 1, HFD 8w: 2, HFD wash out: 3</td>
<td>HFD 4w: 1, HFD 8w: 2, HFD wash out: 3</td>
<td>HFD 4w: 1, HFD 8w: 2, HFD wash out: 3</td>
</tr>
<tr>
<td>GOBP</td>
<td>STEROID_METABOLIC_PROCESS</td>
<td>HFD 4w: 3, HFD 8w: 3, HFD wash out: 3</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 3</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 3</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 3</td>
</tr>
<tr>
<td>REACTOME</td>
<td>CHOLESTEROL_BIOSYNTHESIS</td>
<td>HFD 4w: 2, HFD 8w: 3, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
</tr>
<tr>
<td>REACTOME</td>
<td>REGULATION_OF_CHOLESTEROL_BIOSYNTHESIS_BY_SREBP_SREBF</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
</tr>
<tr>
<td>GOBP</td>
<td>FATTY_ACID_METABOLIC_PROCESS</td>
<td>HFD 4w: 2, HFD 8w: 3, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
</tr>
<tr>
<td>GOBP</td>
<td>LIPID_STORAGE</td>
<td>HFD 4w: 3, HFD 8w: 3, HFD wash out: 3</td>
<td>HFD 4w: 3, HFD 8w: 3, HFD wash out: 3</td>
<td>HFD 4w: 3, HFD 8w: 3, HFD wash out: 3</td>
<td>HFD 4w: 3, HFD 8w: 3, HFD wash out: 3</td>
</tr>
<tr>
<td>KEGG</td>
<td>XENOBIOIC_METABOLISM</td>
<td>HFD 4w: 2, HFD 8w: 3, HFD wash out: 3</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
</tr>
<tr>
<td>H</td>
<td>TNFA_SIGNALING_VIA_NFKB</td>
<td>HFD 4w: 1, HFD 8w: 2, HFD wash out: 1</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
</tr>
<tr>
<td>GOBP</td>
<td>CYTOKINE_PRODUCTION</td>
<td>HFD 4w: 1, HFD 8w: 2, HFD wash out: 1</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
</tr>
<tr>
<td>GOBP</td>
<td>RESPONSE_TO_CYTOKINE</td>
<td>HFD 4w: 2, HFD 8w: 1, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
</tr>
<tr>
<td>REACTOME</td>
<td>INTERLEUKIN_1_SIGNALING</td>
<td>HFD 4w: 3, HFD 8w: 3, HFD wash out: 3</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
</tr>
<tr>
<td>REACTOME</td>
<td>INTERLEUKIN_10_SIGNALING</td>
<td>HFD 4w: 3, HFD 8w: 3, HFD wash out: 3</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
</tr>
<tr>
<td>REACTOME</td>
<td>INFLAMMASOMES</td>
<td>HFD 4w: 3, HFD 8w: 3, HFD wash out: 3</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 2, HFD 8w: 2, HFD wash out: 2</td>
</tr>
<tr>
<td>GOBP</td>
<td>ADAPTIVE_IMMUNE_RESPONSE</td>
<td>HFD 4w: 0, HFD 8w: 0, HFD wash out: 0</td>
<td>HFD 4w: 1, HFD 8w: 2, HFD wash out: 1</td>
<td>HFD 4w: 1, HFD 8w: 2, HFD wash out: 1</td>
<td>HFD 4w: 1, HFD 8w: 2, HFD wash out: 1</td>
</tr>
<tr>
<td>KEGG</td>
<td>APOPTOSIS</td>
<td>HFD 4w: 1, HFD 8w: 1, HFD wash out: 1</td>
<td>HFD 4w: 1, HFD 8w: 1, HFD wash out: 1</td>
<td>HFD 4w: 1, HFD 8w: 1, HFD wash out: 1</td>
<td>HFD 4w: 1, HFD 8w: 1, HFD wash out: 1</td>
</tr>
<tr>
<td>GOBP</td>
<td>ACTIN_FILAMENT_BASED_PROCESS</td>
<td>HFD 4w: 2, HFD 8w: 0, HFD wash out: 0</td>
<td>HFD 4w: 2, HFD 8w: 0, HFD wash out: 0</td>
<td>HFD 4w: 2, HFD 8w: 0, HFD wash out: 0</td>
<td>HFD 4w: 2, HFD 8w: 0, HFD wash out: 0</td>
</tr>
<tr>
<td>GOBP</td>
<td>CELL_MIGRATION</td>
<td>HFD 4w: 3, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 3, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 3, HFD 8w: 2, HFD wash out: 2</td>
<td>HFD 4w: 3, HFD 8w: 2, HFD wash out: 2</td>
</tr>
<tr>
<td>GOBP</td>
<td>INSULINLIKE_GROWTH_FACTOR_RECEPTOR_SIGNALING_PATHWAY</td>
<td>HFD 4w: 0, HFD 8w: 0, HFD wash out: 0</td>
<td>HFD 4w: 0, HFD 8w: 0, HFD wash out: 0</td>
<td>HFD 4w: 0, HFD 8w: 0, HFD wash out: 0</td>
<td>HFD 4w: 0, HFD 8w: 0, HFD wash out: 0</td>
</tr>
</tbody>
</table>
Figure 3. Hepatocyte

A

B

- log10(P)

Module 1

Module 2

Module 3

Module 4

GLUTAMINE FAMILY AMINO ACID BIOSYNTHETIC PROCESS

FARNESOID X RECEPTOR PATHWAY

TGF BETA SIGNALING

APOPTOSIS (HALLMARK)

INTERLEUKIN 10 SIGNALING

RESPONSE TO CYTOKINE NFKB SIGNALING VIA NFKB

XENOBIOTIC METABOLISM

REGULATION OF CHOLESTEROL BIOSYNTHESIS BY SREBP SREBF

STEROID METABOLIC PROCESS

CC-BY 4.0 International license

made available under a (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is

Figure 4. Endothelial

A

Gene
Hgf
Akr3
Dkk2
Ednr1
Lama4
Pxn
Cpne5
Rap2b
Stab2
Lima1
Ccr5
Ii15
Mir30d
Dusp21
Alas2

M1
M2
M3

B

- log10(Ω)

0.0 2.5 5.0 7.5 10.0

PI3K AKT MTOR SIGNALING
MITOTIC SPINDLE
VEGFA VEGR2 SIGNALING PATHWAY
APOPTOSIS
ONCOSTATIN M SIGNALING PATHWAY
P53 AKT MTOR SIGNALING
Figure 5. Stellate

A

Module 1

Module 2

Module 3

Module 4

Gene

Sema3d
Mmp16
Gi3
Pde1a
Ntni
Maf
Adrb2
Cd93
LOC691000
Ndufs6
Mir532
Mir106a
Ces2c
Ccr5
Dusp21
Sp4
Cxxc4
Spcs3
Arsj
LOC100125362

B

- log10(P)

SEMAPHORIN PLEXIN SIGNALING PATHWAY
Figure 6. Macrophage

A

B

Gene

TF binding

0.07

0.06

0.05

0.04

0.03

0.02

0.01

M1

M2

M3

M4

0

5

10

15

20

- log10(P)

0

5

10

15

20

0

5

10

15

20

0

5

10

15

20

- log10(P)

0

5

10

15

20

- log10(P)

0

5

10

15

20

- log10(P)

0

5

10

15

20

- log10(P)

0

5

10

15

20

- log10(P)

0

5

10

15

20

- log10(P)

0

5

10

15

20

- log10(P)

0

5

10

15

20

- log10(P)

0

5

10

15

20

- log10(P)

0
Figure 7

A
Hepatocyte
TNFA SIGNALING VIA NFκB

B
Endothelial cell
TNFA SIGNALING VIA NFκB

C
Macrophage
TNFA SIGNALING VIA NFκB