









# Single-nucleus RNA-seq2 reveals functional crosstalk between liver zonation and ploidy

M. L. Richter <sup>1,9</sup>, I. K. Deligiannis<sup>1,9</sup>, K. Yin<sup>1,2</sup>, A. Danese <sup>3</sup>, E. Lleshi<sup>2</sup>, P. Coupland <sup>2</sup>, C. A. Vallejos <sup>4</sup>, K. P. Matchett <sup>5</sup>, N. C. Henderson <sup>4,5</sup>, M. Colome-Tatche <sup>3,6,7</sup>✉ & C. P. Martinez-Jimenez <sup>1,8</sup>✉

Single-cell RNA-seq reveals the role of pathogenic cell populations in development and progression of chronic diseases. In order to expand our knowledge on cellular heterogeneity, we have developed a single-nucleus RNA-seq2 method tailored for the comprehensive analysis of the nuclear transcriptome from frozen tissues, allowing the dissection of all cell types present in the liver, regardless of cell size or cellular fragility. We use this approach to characterize the transcriptional profile of individual hepatocytes with different levels of ploidy, and have discovered that ploidy states are associated with different metabolic potential, and gene expression in tetraploid mononucleated hepatocytes is conditioned by their position within the hepatic lobule. Our work reveals a remarkable crosstalk between gene dosage and spatial distribution of hepatocytes.

<sup>1</sup>Helmholtz Pioneer Campus (HPC), Helmholtz Zentrum München, Neuherberg, Germany. <sup>2</sup>University of Cambridge, Cancer Research UK Cambridge Institute, Robinson Way, Cambridge, United Kingdom. <sup>3</sup>Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany. <sup>4</sup>MRC Human Genetics Unit, Institute of Genetics & Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh, United Kingdom. <sup>5</sup>Centre for Inflammation Research, The Queen's Medical Research Institute, University of Edinburgh, Little France Crescent, Edinburgh, United Kingdom. <sup>6</sup>TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany. <sup>7</sup>Biomedical Center (BMC), Physiological Chemistry, Faculty of Medicine, LMU Munich, Munich, Germany. <sup>8</sup>TUM School of Medicine, Technical University of Munich, Munich, Germany. <sup>9</sup>These authors contributed equally: M. L. Richter, I. K. Deligiannis. ✉email: [maria.colome@helmholtz-muenchen.de](mailto:maria.colome@helmholtz-muenchen.de); [celia.martinez@helmholtz-muenchen.de](mailto:celia.martinez@helmholtz-muenchen.de)

The liver performs a wide variety of physiological functions, including the metabolism of xenobiotics<sup>1–5</sup> and the regulation of energy homeostasis<sup>6,7</sup>, among others. These functions are mostly performed by parenchymal hepatocytes which constitute about 70% of the liver mass. Liver non-parenchymal cells (NPCs), namely, liver sinusoidal endothelial cells, biliary cells, Kupffer cells, hepatic stellate cells, and immune cell populations constitute the remaining 30% of cell types. All these cells are organized in repetitive structures known as liver lobules<sup>8,9</sup>. The analysis of individual cells by single-cell genomics is changing our understanding of liver homeostasis and pathogenic conditions by taking into account their spatial distribution<sup>10–16</sup>. However, single-cell isolations from the liver require harsh enzymatic or mechanical dissociation protocols that perturb mRNA levels<sup>17</sup>. In particular, the two-step collagenase perfusion generally used to isolate hepatocytes from human livers leads to the downregulation of liver-specific transcription factors such as *Hnf4a*, *Cebpa*, *Hnf1a*, and *Foxa3*<sup>18</sup>, as well as their downstream target genes such as *Cyp2c9*, *Cyp2e1*, *Cyp2B6*, *Cyp2D6*, *Cyp3A5*, and *Cyp3A4*<sup>17,19</sup>.

Single-nucleus RNA-seq (snRNA-seq) has emerged as a complementary approach to study complex tissues at single-cell level<sup>20,21</sup>, including brain<sup>21–27</sup>, lung<sup>28</sup>, kidney<sup>29–32</sup>, liver<sup>33</sup>, and heart<sup>34,35</sup> in mouse and human frozen samples<sup>36,37</sup>. However, there are no snRNA-seq methods tailored for frozen liver tissues. The nuclear transcriptome of individual cells has shown a high correlation to cytoplasmic RNA<sup>23,38</sup>, indicating that snRNA-seq is a powerful tool to study tissues from which intact and fresh cells are difficult to obtain. Here, we have developed a robust single-nucleus RNA-seq2 (snRNA-seq2) approach that relies on efficient lysis of the nuclear membrane. Our approach permits the unbiased characterization of all major cell types present in the liver from frozen archived samples with high resolution.

With snRNA-seq2, we explore at single-cell resolution, a defining feature of hepatocytes ploidy<sup>39–44</sup>. At birth, all hepatocytes are diploid, with a single nucleus containing two copies of each chromosome. During development, polyploidization gradually increases, leading to hepatocytes with several levels of ploidy. Hepatocyte ploidy depends on the DNA content of each nucleus (e.g., diploid, tetraploid, etc.) and the number of nuclei per hepatocyte (e.g., mono- or bi-nucleated)<sup>41</sup>. Here we present a thorough analysis of diploid (2n) and tetraploid (4n) nuclei from the mouse liver and demonstrate that ploidy is an additional source of hepatocyte heterogeneity, linking gene dosage and liver zonation.

## Results

**snRNA-seq2 allows deep and robust characterization of single nuclei isolated from frozen livers.** In order to explore archived samples associated with health and disease conditions, we have developed a robust methodology that combines transcriptomics and efficient low-volume reactions in single nuclei isolated from frozen livers.

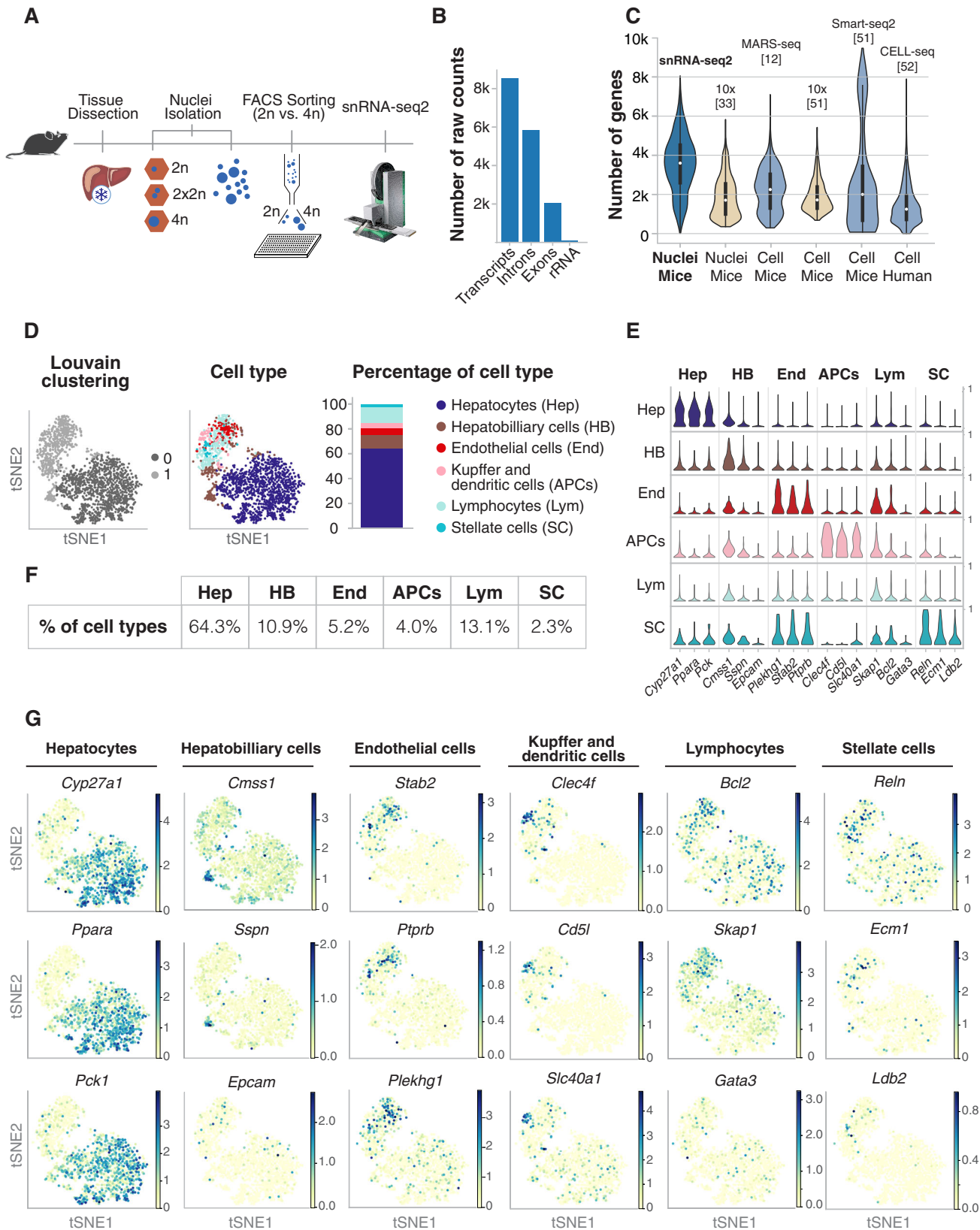
Purified diploid (2n) and tetraploid (4n) nuclei were FACS sorted according to their genome content, using a gating strategy as previously described<sup>39</sup> and followed by a modified version of Smart-seq chemistry using liquid handling robots for volume miniaturization (Fig. 1A, see “Methods”). This approach allows the detection of over 550 000 reads per nucleus, leading to an average detection of more than 11 000 transcripts (~4000 genes) per nucleus (Supp. Fig. 1). As expected for the nuclear transcriptome, the percentage of reads mapping to intronic regions was more than 68%, and the ribosomal reads were 1.2% (Fig. 1B). To show the robustness and reproducibility of this method, four different mice of three months old and C57BL/6J

background were used (Animal ID). Different nuclei isolations were performed from the same liver (Plate ID). Two technical replicates were used to assess technical variability due to sequencing and were removed for downstream analysis (Technical replicate) (Supp. Fig. 1A–D, Supp. Dataset 1 and “Methods”). Additionally, ERCC RNA spike-in mix was used to address technical noise and account for plate effects (including library preparation and sequencing)<sup>45,46</sup>. ERCC normalization was used to make counts comparable across cells and minimize results purely driven by ploidy<sup>47–50</sup> (“Methods”; Supp. Fig. 1A). SnRNA-seq2 detected approximately 4000 genes per nucleus, which is a high number of genes detected per nucleus compared with other single-cell approaches in which intact cells are isolated from livers<sup>12,51,52</sup> (Fig. 1C and Supp. Fig. 1E). Furthermore, our approach showed a *Pearson* correlation of 0.62 between gene expression in the nuclei using our snRNA-seq2 and scRNA-seq from intact cells isolated from fresh livers (Supp. Fig. 1F). This correlation shows that single-nucleus RNA-seq2 is a robust and reliable approach to study transcriptional profiles of individual cells from archived frozen tissues<sup>20,23,25,37,53–55</sup>.

The main improvement of our approach relies on the addition of a supplementary lysis buffer compatible with the generation of full-length cDNA and library preparation without the need for additional clean-up steps. This second lysis buffer (LB2) is compatible with a wide range of commercially available platforms and chemistries, including C1 Fluidigm (Fluidigm), well-plate approaches combined with liquid handling robots (Mosquito HV, TTP Labtech), as well as both SMARTer and NEBNext commercially available chemistries (Supp. Fig. 2 and Supp. Dataset 2). For instance, we detected 3.6-fold more genes per nucleus with our additional LB2 compared with classic approaches using SMARTer chemistry (Supp. Fig. 2C, D) and snRNA-seq2 performed better than other droplet-based approaches (Supp. Fig. 2E, F).

T-distributed stochastic neighbor embedding (t-SNE) was used to visualize nuclei FACS-sorted in an unbiased manner (Fig. 1D). Low-resolution *Louvain* clustering showed that the nuclei were clustering into two groups (0, 1), corresponding to hepatocytes and non-hepatocyte cells (Fig. 1D, left). After higher resolution clustering and identification of the top differential expressed genes between clusters, we identified hepatocytes (64.3%), hepatobiliary cells (10.9%), Kupffer and dendritic cells (APCs, 4%), endothelial cells (5.2%), stellate cells (2.3%), and lymphocytes (13.1%) as main cell clusters (Fig. 1D, right and Fig. 1E; Supp. Fig. 1F, G, Supp. Dataset 1 and Supp. Dataset 2)<sup>11,12,52,56</sup>. The expression distribution of key markers characteristic of those cell types showed that all liver cell populations can be identified with this methodology (Fig. 1F). Additionally, visualization of key representative markers such as *Cyp27a1*, *Ppara*, and *Pck1* for hepatocytes; *Sspn*, *Cmss1*, and *Epcam* for hepatobiliary cells; *Plekhhg1*, *Stab2*, and *Ptprb* for endothelial cells; *Clec4f*, *Cd5l*, and *Slc40a1* for APCs; *Bcl2*, *Skap1*, and *Gata3* for lymphocytes; and *Reln*, *Ecm1*, and *Ldb2* for stellate cell showed that transcriptional heterogeneity can be captured from frozen liver tissues for each cell type (Fig. 1G). Higher resolution clustering on non-hepatocytes identified additional three cell types including *Epcam*-positive epithelial cells (0.73%), and different sub-groups of lymphocytes, including T—(4.00%), Plasma B—(4.61%), and B-cells (4.49%) (Fig. 1D, right; Supp. Fig. 3, Supp. Dataset 3 and Supp. Dataset 4).

Further in-depth characterization of liver cell populations revealed additional features related to their genome content and respective ploidy levels. Heat map of the top five differentially expressed genes showed that hepatocytes and hepatobiliary cells were mainly composed of 2n and 4n nuclei while other cell types were primarily associated to 2n nuclei (Fig. 2A and Supp. Fig. 4A).



Consistent with the polyploid nature of hepatocytes<sup>39,57</sup>, we found as expected, that the hepatocyte cluster was enriched with a mixture of 2n and 4n nuclei (Fig. 2A). Nonetheless, few 4n nuclei were found in the non-hepatocyte clusters. Thus, we investigated whether cell cycle state could explain the presence of 4n nuclei in non-hepatocyte clusters (Fig. 2B and Supp. Fig. 4B–D)<sup>45,58,59</sup>.

Cyclone was used to identify nuclei associated with G1, G2/M, and S phase of cell cycle<sup>59</sup>. As expected in the liver tissue, the majority of nuclei were in G1 phase (Fig. 2B Supp. Fig. 4B–D). This resulted in 94.5% of the hepatocyte cluster assigned to G1 phase, while 85.5% of nuclei in the non-hepatocyte cluster were in G1 phase. The number of 4n nuclei in the non-hepatocyte cluster

**Fig. 1 snRNA-seq2 on frozen tissue identifies and characterizes all major cell types present in the liver.** **A** Schematic representation of the snRNA-seq2 pipeline. Nuclei are isolated from flash-frozen tissue, sorted by genome content in 384 well-plates, and subjected to the generation of full-length cDNA and sequencing. **B** Quality control of sequenced data shows the number of raw counts per genome feature in the nuclear transcriptome. **C** Violin plots comparing the high number of genes detected using snRNA-seq2 to publicly available single-cell RNA-seq datasets. Light brown: droplet-based methods ( $n = 10,395$  nuclei [33];  $n = 1,026$  cells [51]), blue: plate-based approaches ( $n = 2,496$  nuclei snRNA-seq2;  $n = 1,736$  cells [12];  $n = 981$  cells [51];  $n = 12,622$  cells [52]). In the violin plot, boxplots indicate median number of genes per method (white dot). The lower and upper ends of the boxes correspond to first and third quartiles, respectively, while the whiskers extend to minimum (first quartile minus  $1.5 \times$  the inter-quartile range (IQR)), and maximum (third quartile plus  $1.5 \times$  IQR). Data points beyond the whiskers are only represented through their density distribution in the surrounding violin plots. For clearer visualization, the violins are cut at 0. **D** t-distributed stochastic neighborhood embedding (tSNE) of the nuclei by low-resolution *Louvain* clusters (left), with cell-type annotation for all major cell types (middle), and bar plot showing the percentage of all cell types (right) identified by snRNAseq2. **E** Stacked violin plot depicting selected illustrative marker genes per cell type (y-axis: scaled, log-transformed, normalized counts). **F** Percentage of cells types identified by snRNA-seq2. **G** tSNE colored by marker gene expression revealed cellular heterogeneity in all major cell population of the mouse liver.

was below 7%, which could be explained either by cell cycle stage (11 nuclei were in division) or technical bias during FACS sorting (e.g., nuclei tend to clump together). Therefore, FACS sorting of nuclei according to their DNA content is a robust and accurate strategy to analyze hepatocytes with different levels of ploidy<sup>39</sup>.

These results demonstrate that our method is highly sensitive with regards to the identification of the main cell types and simultaneous examination of the cellular heterogeneity present in intact frozen liver tissues.

**snRNA-seq2 uncovers transcription factors and rate-limiting enzymes involved in chronic liver disease.** In order to investigate whether the nuclear transcriptome can be used to address functional responses in the context of health and disease, we studied the expression levels of key liver-specific transcription factors (e.g., *Hnf4a*, *Ppara*, *Mlxpl*/ChREBP, and *Cebpa*), nuclear receptors (e.g., *Rxa*, *Nr1l2*/PXR, *Nr1i3*/CAR, and *Nr1h4*/FXR) and coactivators (e.g., *Ppargc1a*/PGC1A, *Crebbp*/CBP/p300, and *Nco1*/SRC-1)<sup>1–3,6,60</sup> (Fig. 2C and Supp. Fig. 5A). Our approach allowed deep characterization of the expression levels of upstream metabolic regulators in hepatocytes bearing two (2n) and four (4n) complete genomes. We inspected the gene expression levels of key transcription factors and downstream effectors between 2n and 4n nuclei isolated from hepatocytes (Fig. 2D and Supp. Fig. 5)). For instance, peroxisome proliferator-activated receptor  $\alpha$  (*Ppara*) is the most abundant isoform expressed in hepatocytes with key roles in lipid metabolism and displays protective roles against non-alcoholic fatty liver disease (NAFLD)<sup>61–63</sup>. *Ppara* expression was upregulated in 4n nuclei in comparison to 2n nuclei (Fig. 2D; Supp. Fig. 5, Supp. Dataset 4 and Supp. Dataset 5). Similarly, Carbohydrate responsive element binding protein (ChREBP), encoded by the *Mlxpl* gene, is a carbohydrate-signaling transcription factor highly expressed in liver and adipose tissues and regulates the synthesis de novo of fatty acids<sup>64–66</sup>. *Mlxpl* is upregulated in 4n hepatocytes (Fig. 2D, Supp. Fig. 5, Supp. Dataset 4 and Supp. Dataset 5). The precise interplay between upstream transcription factors and rate-limiting enzymes during hepatic lipid and glucose metabolism is likely to determine the final outcome in NAFLD-induced chronic lipid overload. Therefore, investigating the functional differences between 2n and 4n hepatocytes is crucial to understand the development and progression of complex liver disease.

Recently, changes in the expression patterns of transcriptional regulators and downstream target genes have been associated with the development and progression of chronic liver diseases such as NAFLD<sup>67,68</sup>, non-alcoholic steatohepatitis (NASH)<sup>69</sup>, fibrosis<sup>15</sup>, and human hepatocellular carcinoma (HCC)<sup>52</sup>. From those studies, well-known markers associated with chronic liver disease, were selected to illustrate that our methodology is highly sensitive to detect those genes from the nuclear transcriptome (Fig. 2E, F and Supp. Fig. 5). In particular, dysregulation of *Akr1d1* has been

associated with human NAFLD<sup>68</sup>, and upregulation of *Yap1* (paralogue of *Wwtr1*) in human and murine NASH liver<sup>70</sup>. Furthermore, *Tead1* has been proposed as a marker of NASH in murine mouse models<sup>71</sup>, whereas *Vegfa*, *Lifr*, and *Nrp1* have been associated with the intrahepatic ligand-receptor signaling network involved in the pathogenesis of NASH<sup>69</sup> (Fig. 2E, F and Supp. Fig. 5B). Similarly, downregulation of *Bic1* and *Fstl1* has been associated with advance fibrosis<sup>72</sup>. More recently, Aizarani et al.<sup>52</sup> have shown that the perturbation of gene signatures in individual cells is associated with HCC, for instance, changes in *Serpin1c* and *Igfbp1* (Fig. 2E, F; Supp. Fig. 5B).

In summary, we are able to detect marker genes associated with chronic liver disease identified from bulk and single-cell transcriptomic analysis. This methodology has the potential to investigate previously archived frozen murine and human samples, and interrogate how changes in gene expression correlate with the development and progression of liver diseases, taking into account cellular crosstalk and signaling pathways.

**snRNA-seq2 reveals differential transcriptional variability between 2n and 4n nuclei.** Although recent single-cell transcriptomic studies have addressed cellular heterogeneity in the liver with respect to hepatocyte zonation<sup>11,12,14,15,52</sup>, and the variability among non-parenchymal cells during chronic liver disease<sup>69,73–76</sup>, little is known about the cellular heterogeneity in hepatocytes with different levels of ploidy<sup>57,77–79</sup>.

Polyploidy occurs in the liver during normal development<sup>80</sup>, and is also associated with pathological conditions such as cancer and chronic liver disease<sup>81,82</sup>. Moreover, ploidy increases with age<sup>83–85</sup>, as well as in chronic liver conditions<sup>86</sup>. More specifically, the presence of tetraploid mononucleated hepatocytes has been associated with poor prognosis in human hepatocellular carcinoma (HCC)<sup>87,88</sup>. Accordingly, we further characterized the transcriptional profile of individual 2n and 4n nuclei from the hepatocyte cluster and investigated whether their transcriptomic profile differs among those two populations.

Across all cell types, the median number of genes detected in 4n nuclei was 1.36-fold higher than in 2n nuclei (Supp. Fig. 6A and “Methods”). In the hepatocyte cluster, a 1.25-fold increase was detected in 4n compared to 2n nuclei (Fig. 3A), most likely due to the increase in cellular volume in 4n which has been associated with an increase in global transcription in larger cells<sup>89</sup>. We saw that 2n and 4n nuclei from the hepatocyte cluster do not separate in tSNE embedding based on their global gene expression profile (Fig. 3B and Supp. Fig. 6B); thus, they would be indistinguishable without a prior knowledge of their ploidy status. Herein, 2n and 4n nuclei from the hepatocyte cluster will be named 2n and 4n hepatocytes respectively.

Differential expression analysis showed that 2n and 4n hepatocytes are globally very similar. We detected 248 genes upregulated and 64 genes downregulated in 4n (Fig. 3C, Supp.

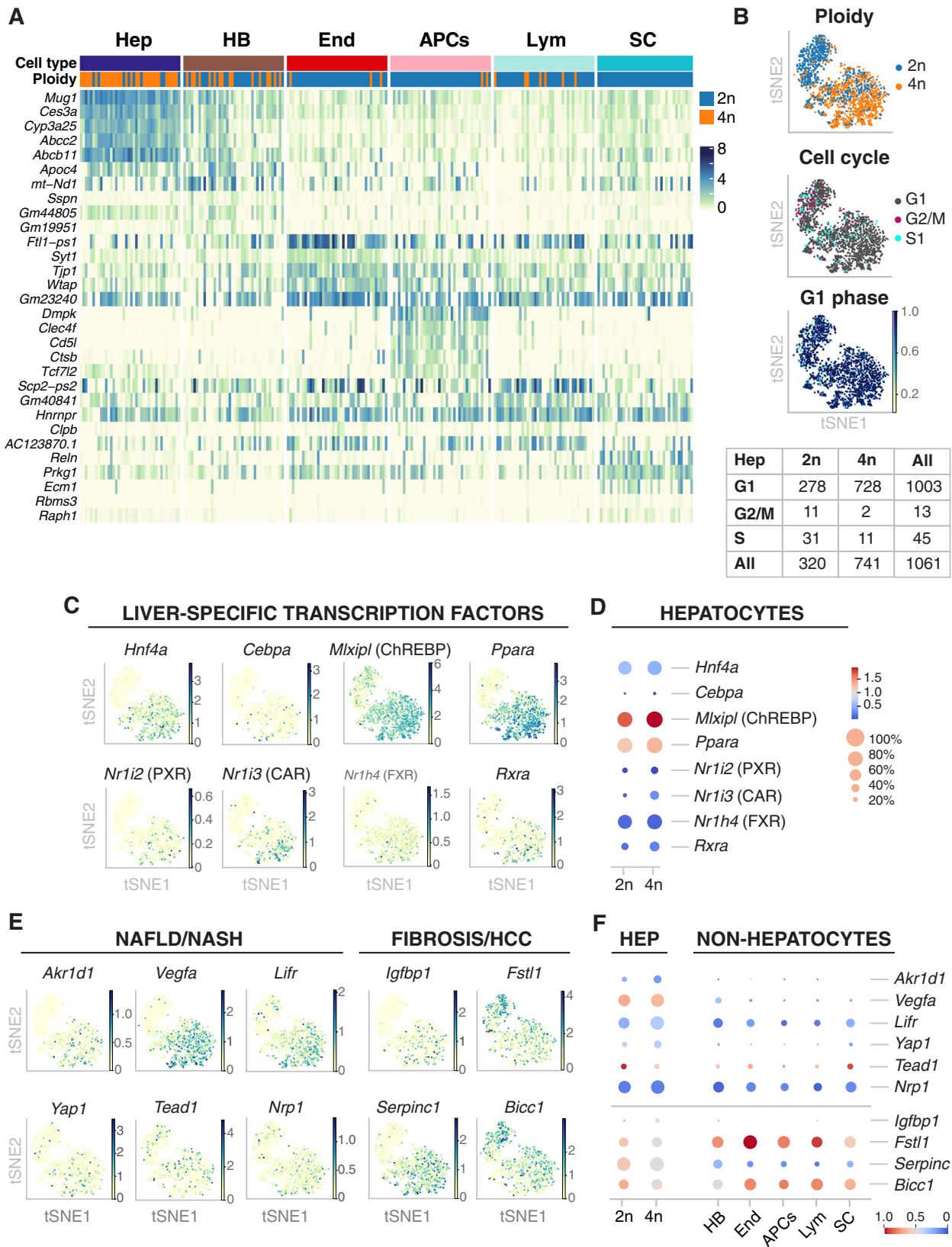
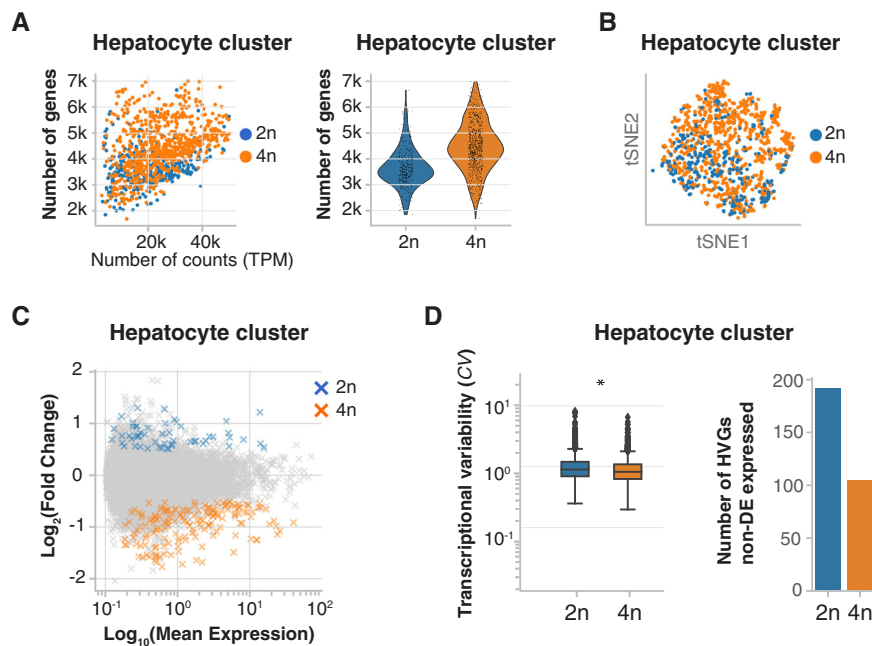


Fig. 6C, D and Supp. Dataset 5). Only two of the 248 genes upregulated in 4n had zero counts in the 2n population (“Methods” and Supp. Fig. 6D, Supp. Dataset 5).

Subsequently, the transcriptional variability was estimated as the *coefficient of variation* using log-transformed data<sup>90–92</sup> (Supp. Information). We confirmed that 4n hepatocytes showed a

significantly lower *coefficient of variation* compared to 2n hepatocytes (2n are 1.09-times more variable than 4n, Mann–Whitney U-test,  $p$ -value  $2.097e-14$ ) (Fig. 3D and Supp. Dataset 6). Accordingly, the number of highly variable genes (HVGs) was higher in the 2n hepatocytes compared to the 4n hepatocytes (Fig. 3D and Supp. Dataset 6).

**Fig. 2 snRNA-seq2 allows deep profiling of single nuclei including key liver-specific transcription factors and downstream target genes involved in healthy homeostasis and chronic liver disease.** **A** Heatmap showing the gene expression of the top five differentially expressed genes in forty randomly selected nuclei per cell type (colored by log-transformed, normalized counts). Ploidy analysis shows that 4n nuclei are enriched in the hepatocyte cluster. **B** Cell cycle analysis using *Cyclone* shows that the majority of nuclei are in G1 phase. tSNEs colored by ploidy (top), assigned cell cycle phase (middle), and G1 score (bottom). Table showing the number of nuclei that are in each assigned phase for diploid and tetraploid nuclei from the hepatocyte cluster. **C** tSNE colored by the expression of key liver-specific transcription factors involved in liver homeostasis and hepatic function. **D** Dot plot shows that the expression levels (color scale) and the percentage (dot size) of cells expressing key transcription factors can be dissected between 2n and 4n nuclei. **E** tSNEs colored by the expression of disease-related marker genes, separated into two main categories of liver disease: non-alcoholic fatty liver disease/non-alcoholic steatohepatitis (NAFLD/NASH) and fibrosis/hepatocellular carcinoma (HCC), showing cellular heterogeneity in different cell populations. **F** Dot plot showing the expression of the disease-related marker genes across cell types in young wild-type mice livers.



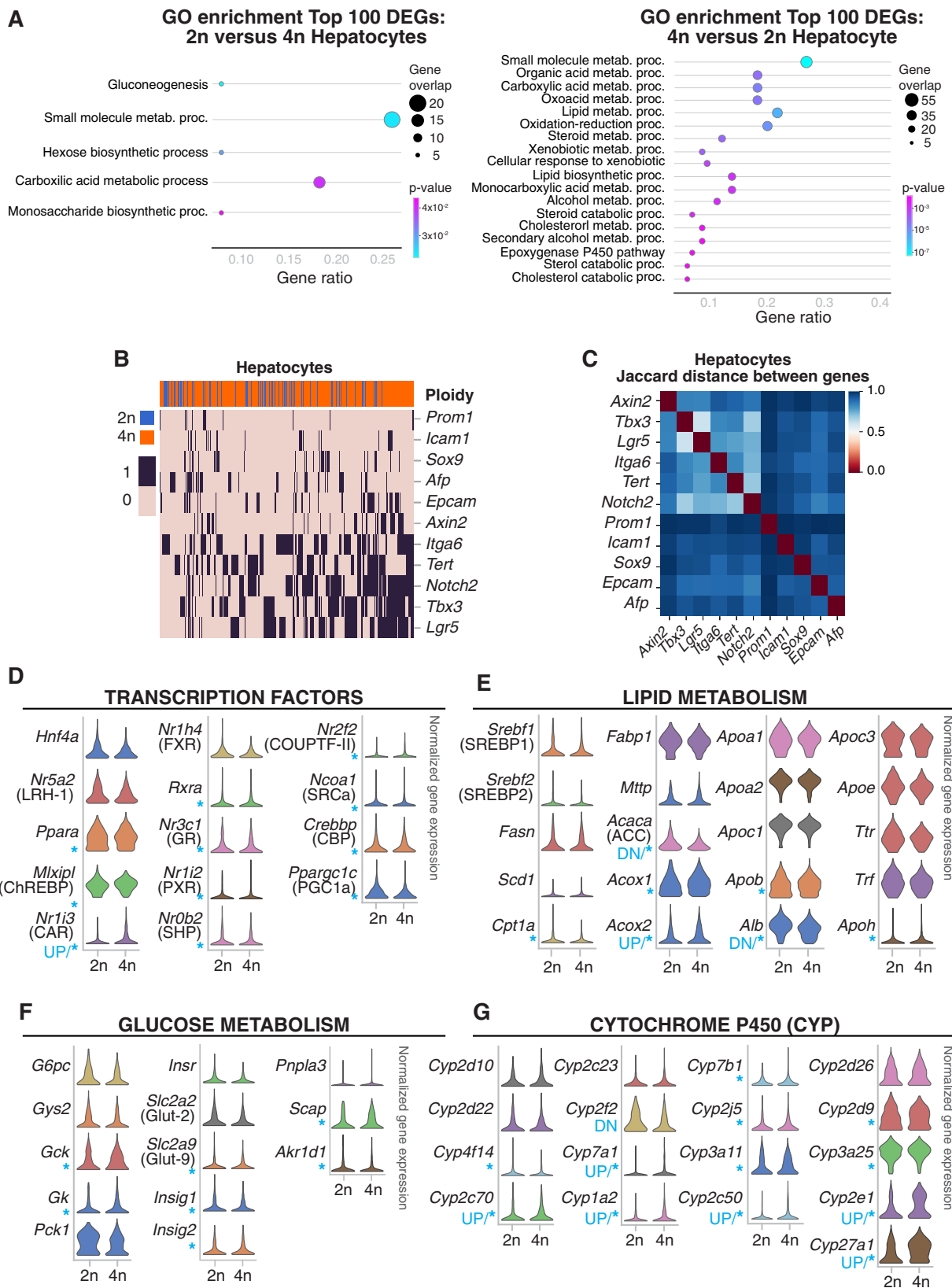
**Fig. 3 snRNA-seq2 reveals differences in gene expression between diploid and tetraploid hepatocytes.** **A** Scatterplot showing the higher number of genes per gene-length normalized counts (TPM) detected in 4n hepatocytes compared with 2n hepatocytes (left). The violin plot showing that the number of genes in 4n nuclei is 1.25-times higher than in 2n hepatocytes (right). **B** tSNE embedding of the hepatocyte cluster showing that 2n and 4n nuclei are clustered together. **C** MA plot showing average logarithmically transformed mean expression (x-axis) versus the  $\text{log}_2$  fold change (y-axis) for pairwise comparison between 2n and 4n hepatocytes; 312 differentially expressed genes (DEGs) are depicted with crosses: 64 genes are upregulated in 2n nuclei (blue) and 248 genes are upregulated in 4n (orange). **D** Box plot of the coefficient of variation showing the transcriptional variability between 2n ( $n = 320$ ) and 4n ( $n = 741$ ) (left; two-sided Mann–Whitney U-test, \*:  $p$ -value =  $2.09 \times 10^{-14}$ ). Box plots show the median coefficient of variation per ploidy status. The lower and upper ends of the boxes correspond to first and third quartiles, respectively, while the whiskers extend to minimum (first quartile minus  $1.5 \times$  the inter-quartile range (IQR)), and maximum (third quartile plus  $1.5 \times$  IQR). Data points beyond the whiskers are shown as individual points. Bar plot showing the higher number of highly variable genes (HVGs) that are not differentially expressed (non-DE) in 2n vs 4n (right).

In summary, 2n and 4n hepatocytes showed a similar transcriptional profile in young mice; however, 312 DEGs were detected between these two cellular states. These results revealed the concealed cellular heterogeneity present in hepatocytes with different levels of ploidy.

**Tetraploid nuclei show extensive co-expression of liver stem cell markers.** Polyploid hepatocytes have been associated with terminal differentiation and senescence<sup>41</sup>. However, a growing body of evidence indicates that polyploid hepatocytes retain their proliferative potential<sup>79,93–96</sup>. To further analyze the functional characteristics of 2n and 4n hepatocytes, Gene Ontology (GO) analysis was first performed on significant DEGs (Supp. Fig. 7A and Supp. Dataset 5). DEG in 2n hepatocytes were enriched in one category (i.e., small molecule metabolic process), while DEG in 4n hepatocytes were enriched in twenty different processes (Supp. Fig. 7A and Supp. Dataset 5). Secondly, we extended our GO analysis to the top hundred most differentially expressed

genes (Fig. 4A and Supp. Dataset 5). Comparisons between 2n and 4n hepatocytes showed that 4n hepatocytes were further enriched in pathways involved in lipid, cholesterol, and xenobiotic metabolism (Fig. 4A and Supp. Dataset 5).

In order to further characterize the basal differences between 2n and 4n hepatocytes, we studied the regenerative and proliferative properties of these two cellular states. It has been suggested that diploid hepatocytes, located in proximity to the central vein of the liver lobule, act as stem cells during homeostasis and in response to injury<sup>97</sup>. However, it has recently been shown using a multicolor reporter allele system that polyploid hepatocytes proliferate in chronically injured livers<sup>79,93,98</sup>. Independent studies by Su *et al.* using AXIN2 lineage tracing have shown that hepatocytes upregulate *Axin2* and *Lgr5* after injury throughout the liver lobe<sup>79,93,98</sup>. To investigate if there are specific subpopulations of cells with stem cell properties enriched in 2n or 4n hepatocytes, several stem/progenitor cell-like marker genes were selected, including *Icam1*,



*Afp*, *Sox9*, *Epcam*, *Axin2*, *Tbx3*, *Itga6*, *Tert*, *Lgr5*, and *Notch2*<sup>78</sup> (Fig. 4B and Supp. Fig. 7B). Five of these markers, *Notch2*, *Tbx3*, *Itga6*, *Lgr5*, and *Tert* were expressed in more than 40% of the nuclei analyzed (Supp. Fig. 7C). While we observed that both 2n and 4n hepatocytes expressed several stem/progenitor markers, we did not find an enrichment of those genes in diploid

hepatocytes (Fig. 4B and Supp. Fig. 7C). In order to investigate whether these markers were co-expressed in the same nuclei, the Jaccard index was used to measure the probability of those genes being co-expressed in the same nucleus<sup>99,100</sup>. In particular, the Jaccard distance measures dissimilarity between genes, and lower distance indicates a higher probability of co-expression (Fig. 4C).

**Fig. 4 Functional characterization of 2n and 4n hepatocytes at the single-cell level.** **A** Gene ontology (GO) analysis of the top 100 genes upregulated in 2n hepatocytes in comparison to 4n hepatocytes (left), and vice versa (right). **B** Heatmap showing co-expression of stem cell markers in hepatocytes (binary expression: 1-detected, 0-non-detected). Ordering of the cells and genes is based on hierarchical clustering between nuclei (columns), and stem cell markers (rows), revealing the stem cell markers are co-expressed in 2n and 4n hepatocytes. **C** Heatmap showing pairwise Jaccard distances between stem cell markers, reveals one main module of co-expression in hepatocytes. **D** Violin plots showing changes in gene expression level and distribution in key liver-specific transcription factors; **E** regulators of the hepatic lipid metabolism; **F** regulators of hepatic glucose metabolism and; **G** representative genes from the cytochrome P450 family involved in the xenobiotic metabolism of 2n and 4n hepatocytes. (y-axis: scaled, log-transformed, normalized counts; \*: significant changes in gene expression distribution, UP: significant upregulation in 4n, DN: significant downregulation in 4n; Bonferroni-adjusted  $p$ -value < 0.05; specific  $p$ -values can be found in Supplementary Dataset 4 and 7).

With the Jaccard distance one module was identified for *Axin2*, *Tbx3*, *Lgr5*, *Itga6*, *Tert*, and *Notch2* (Fig. 4C and “Methods”). This module showed higher pair-wise similarity independently of gene expression levels. Additional genes showed higher distance and lower probability of co-expression: *Afp*, *Epcam*, *Sox9*, and *Icam1* (Fig. 4C and “Methods”). Furthermore, the percentage of nuclei expressing those gene markers was similar in both 2n and 4n hepatocytes (Supp. Fig. 4C) indicating that 2n and 4n hepatocytes have similar stem cell marker gene expression profiles. More than two stem/progenitor markers were co-expressed in both 2n and 4n hepatocytes (Fig. 4B, and Supp. Fig. 4D; “Methods”), strongly supporting the notion that polyploid hepatocytes have regenerative potential. Immunofluorescence (IF)/RNA-FISH co-detection analysis was used to simultaneously detect beta-catenin and *Lgr5* mRNA signals in the liver tissue (Supp. Fig. 7E). A higher percentage of *Lgr5* positive hepatocytes surrounding the central vein (CV) was detected compared to the periportal hepatocytes. These findings are in agreement with previous observations by Sun *et al.*<sup>93</sup>. Quantification of *Lgr5* mRNA copies per nucleus showed no significant differences in expression levels between 2n and 4n hepatocytes (Supp. Fig. 7E, right), thus validating our results generated from our snRNA-seq2 analyses.

In summary, no subpopulation of diploid hepatocytes enriched in stem/progenitor markers was detected in young healthy livers. Furthermore, polyploid hepatocytes co-express genes associated with proliferative and regenerative functions, and therefore have the potential to contribute to organ regeneration and liver homeostasis.

**Changes in expression pattern distribution in 4n nuclei indicate a high capacity for adaptation and regeneration.** A focused analysis on energy homeostasis also revealed changes in the distribution pattern of key metabolic genes involved in lipid and glucose metabolism (Fig. 4D–G, Supp. Fig. 8 and Supp. Dataset 7). Some liver-specific transcription factors showed no change in mean expression or distribution pattern between 2n and 4n hepatocytes, for instance, *Hnf4a*, *Nr5a2* (LRH-1), *Nr1h4* (FXR), and *Rxa* (Fig. 4D). However, *Mlxipl* (ChREBP), *Nr3c1* (GR), *Nr1i2* (PXR), *Nr0b2* (SHP), and *Nr2f2* (COUPTF-II), and the coactivators *Ncoa1* (SRCa), *Crebbp* (CBP), and *Ppargc1c* (PGC1a) showed a different distribution pattern between 2n and 4n but no significant changes in mean expression. In some cases, such as *Nr1i3* (CAR), changes in distribution were associated with upregulation of its expression levels in 4n hepatocytes (Fig. 4D, Supp. Dataset 5 and Supp. Dataset 7). We further studied critical regulators of lipid metabolism and found that *Cpt1*, *Acox1*, *Apob*, and *ApoB* showed changes in distribution pattern but no significant changes in mean expression. Albeit, in *Acaca* (ACC) and *Alb*, we found that their distribution pattern changed and its expression was downregulated in 4n hepatocytes, while *Acox2* was upregulated in 4n hepatocytes.

For genes involved in glucose metabolism, statistically significant changes in the distribution pattern of *Gck*, *Gk*, *Slc2a9* (Glut9), *Insig1*, *Insig2*, *Scap*, and *Akr1d1* were found (Fig. 4F,

Supp. Fig. 8 and Dataset 7). We observed considerable changes in the metabolism of xenobiotics and cytochrome P450 superfamily (Fig. 4G, Supp. Fig. 8 and Supp. Dataset 5). For instance, *Cyp4f14*, *Cyp2j5*, *Cyp2d9*, and *Cyp3a25* showed changes in gene expression distribution, while *Cyp2c70*, *Cyp7a1*, *Cyp1a2*, *Cyp2c50*, *Cyp2e1*, and *Cyp27a1* were additionally upregulated in 4n hepatocytes (Fig. 4G).

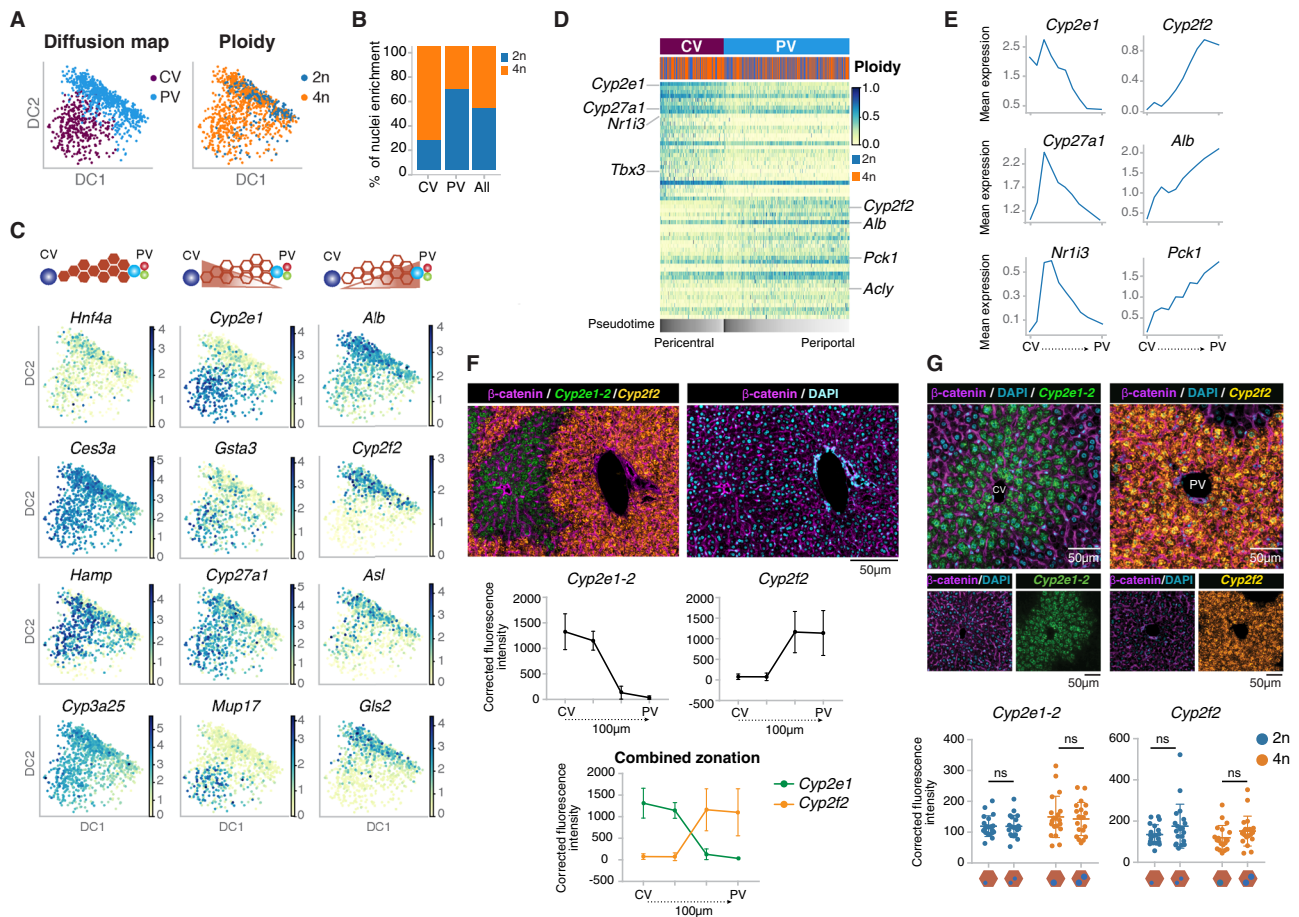
We further extended our analysis to hepatocytes with high levels of ploidy (>4n), and octaploid (8n), and decahexaploid (16n) nuclei were FACS sorted followed by snRNA-seq2 (Supp. Fig. 9 and Supp. Information). As expected, young mice had a low percentage of 8n (3%) and 16n (0.008%) hepatocytes in comparison to 2n (56%) and 4n (36%) hepatocytes (Supp. Fig. 9A, B). Similarly, we observed genes significantly upregulated or downregulated with high levels of ploidy, as well as statistically significant changes in gene expression distribution (Supp. Fig. 9C, D) again indicating high levels of cellular heterogeneity in the hepatocyte compartment.

In conclusion, rate-limiting enzymes involved in energy homeostasis and metabolism of drugs showed a statistically significant change in expression distribution between 2n and 4n hepatocytes. Changes in gene expression distribution have been associated with genetic plasticity and higher adaptation<sup>101–104</sup>, which could be important when the liver is faced with a chronic or overwhelming insult.

**Hepatic metabolic zonation determines gene expression levels independently of the ploidy status.** Zonation of hepatocytes and endothelial cells along the portal-central axis of the liver lobule has previously been investigated<sup>11,12</sup>. Recently, 2n and 4n hepatocytes have been localized to specific metabolic zones (periportal or pericentral) or randomly interspersed throughout the hepatic lobe<sup>97,105,106</sup>. We used diffusion pseudotime (dpt), visualized in diffusion maps, to infer the pseudospacial ordering of nuclei according to defined markers associated with liver zonation and their metabolic specialization<sup>12,107,108</sup>. *Louvain* clustering on these zonation markers established three clusters (Fig. Supp. 10A and Supp. Information). After visual investigation of pericentral and periportal marker genes on the diffusion map, two clusters (0 and 2) were assigned to a periportal cluster, and cluster 1 to a pericentral cluster (Fig. 5A, Supp. Fig. 10A, and Supp. Information). Then, the percentage of 2n and 4n hepatocytes in the pericentral (CV) and periportal (PV) clusters were calculated respectively (Fig. 5B and Supp. Information)<sup>12</sup>. We observed a 1.3-fold relative enrichment of 4n hepatocytes in the pericentral cluster, indicating that they reside close to the central vein. This is in agreement with recent studies in which polyploid hepatocytes have been shown by histological analysis to be enriched in the pericentral zone, and diploid hepatocytes in the periportal zone using cell-lineage tracing<sup>79,109</sup>.

We further studied the expression levels of classic markers of liver zonation such as *Cyp2e1*, *Gsta3*, *Cyp27a1*, and *Mup17* for the central vein, and *Alb*, *Cyp2f2*, *Asl*, and *Gls2* for the portal vein (Fig. 5C; Supp. Fig. 10). Non-zonated markers were also selected,





**Fig. 5 Pseudospacial ordering of hepatocytes along the liver lobe shows functional crosstalk between ploidy and zonation.** **A** Diffusion map based on zonation markers, colored by annotated zones (left) and ploidy (right). **B** Bar plot showing a higher percentage of nuclei in the pericentral cluster in comparison to the periportal cluster, colored by ploidy. **C** Diffusion maps colored by the expression of zonation marker genes, for representative non-zonated genes (*Hnf4a*, *Ces3a*, *Hamp*, and *Cyp3a25*), pericentral genes (*Cyp2e1*, *Gsta3*, *Cyp27a1*, and *Mup17*), and periportal genes (*Alb*, *Cyp2f2*, *Asl*, and *Gls2*). **D** PAGA path heatmap showing the top 30 differentially expressed genes per zone between the periportal (PV) and pericentral (CV) clusters of nuclei, where 4n hepatocytes are enriched in the CV zone (colored by log-transformed, normalized counts). **E** Line plots depicting the mean expression of representative zonation markers along the diffusion pseudospace vector ordered from CV to PV within a liver lobule. **F** Representative image of IF/RNA-FISH co-detection analysis, showing expression levels of *Cyp2e1-2* (green) and *Cyp2f2* (yellow) mRNA in a liver lobule ( $n = 3$ ). The fluorescence intensity of each gene was calculated and normalized against background intensity. The corrected fluorescence intensity is plotted against a linear distance (100  $\mu\text{m}$ ) from CV to PV. **G** Corrected fluorescence intensity of *Cyp2e1-2* and *Cyp2f2* signals were categorized according to 2n or 4n mononucleated or binucleated hepatocytes, determined by  $\beta$ -catenin (magenta) and 4',6-diamidin-2-phenylindol (DAPI; cyan) staining. No significant changes were detected between ploidy levels or the number of nuclei from an unpaired  $t$ -test analysis (ns not significant). **F, G** Data points are expressed as mean  $\pm$  s.d.

for instance, *Hnf4a* whose mRNA expression is non-zonated as opposed to its protein levels<sup>10</sup>. Additional non-zonated genes including *Ces3a*, *Hamp*, and *Cyp3a25* were also studied (Fig. 5C, Supp. Fig. 10). In order to verify the spatial distribution of 4n hepatocytes, we generated a partition-based graph abstraction (PAGA) to visualize coordinated changes in expression across the pseudospace in the hepatic lobule<sup>110</sup> (Supp. Information). The thirty most differentially expressed genes in each of the two zonation clusters were selected along the diffusion pseudotime, again showing an enrichment of 4n hepatocytes in the pericentral cluster and the expected changes of pericentral and periportal markers, respectively (Fig. 5D, E, Supp. Fig. 11 and Supp. Dataset 8). Out of 224 zonation markers that were upregulated in the pericentral cluster, 55 were upregulated only in 4n hepatocytes. Meanwhile, out of 68 zonation markers that were upregulated in the periportal cluster 12 were upregulated in 2n hepatocytes. We binned the vector of diffusion pseudotime into 10 bins and visualized the mean expression of specific zonation markers along these bins (Fig. 5E, Supp. Fig. 11). This approach is

suitable to visualize simultaneously the decrease and increase in mean expression of periportal and pericentral genes along these bins. These results showed that liver zonation can be investigated by diffusion pseudotime using our snRNA-seq2 approach. In particular, hepatic metabolic zonation was found in both 2n and 4n hepatocytes indicating that spatial zonation determines gene expression levels independently of the ploidy status.

In order to validate the spatial relationship shown by changes in the nuclear transcriptome, a thorough analysis of cell size and ploidy was done based on nuclear content using ImageStream Mk II (Supp. Fig. 9E–H). Additional immunofluorescence staining for Lyve1, a marker of endothelial cells indicated that their nuclei are smaller in size (3–6  $\mu\text{m}$  diameter) and represent the smaller peak within the 2n population (Supp. Fig. 9E–H). After analysis of the nuclei size ( $\mu\text{m}^2$ ) and diameter ( $\mu\text{m}$ ) of 2n, 4n, and 8n hepatocytes, DNA-FISH with a custom-designed probe against *Hnf4a* gene in chromosome 2 was used to confidently identify nuclei with 2, 4, and 8 copies of *Hnf4a* gene locus (Supp. Fig. 12A–E). Additional immunostainings with DAPI and beta-

catenin, to annotate nuclei and the cytoplasmic membrane respectively, allowed tissue mapping and quantification of diploid (2n), tetraploid mononucleated (4n) and binucleated (2×2n), and octaploid binucleated (2×4n) cells (Fig. 12F, G). This approach allowed us to confidently identify diploid and tetraploid cells (mono- and binucleated), verify the predictions of the diffusion maps, and validate our findings using snRNA-seq2 in which 4n nuclei are enriched in the pericentral vein area (Fig. 5B).

Importantly, IF/RNA-FISH co-detection was used to quantify the number of *Cyp2e1* and *Cyp2f2* mRNA molecules in four zones within the liver lobule (from central to portal vein) (Fig. 5F, G). We observed that the mRNA level of *Cyp2e1* decreases towards the periportal area, whereas *Cyp2f2* mRNA levels increase (Fig. 5F). These results in tissue sections validated our findings using snRNA-seq2 and computational approaches (Fig. 5E). The combination of RNAscope staining for *Cyp2e1* and *Cyp2f2* mRNAs with DAPI and beta-catenin allowed accurate assessment of mRNA levels in polyploid hepatocytes. We found no significant differences in the mRNA levels between 2n nuclei in diploid and tetraploid binucleated cells using RNAscope (Fig. 5G), and those populations cannot be distinguished by snRNA-seq2 (Fig. 3 and Supp. Fig. 6B).

In summary, our results indicate a tight coordination between gene expression in polyploid hepatocytes and their spatial location within the liver lobule. Therefore, hepatocytes adjust their gene expression according to their spatial distribution, demonstrating a functional crosstalk between liver zonation and ploidy. We then further demonstrated that zonation is the major determinant of gene expression levels in polyploid hepatocytes by using a mouse model of centrilobular injury-induced liver fibrosis (CCl<sub>4</sub>-induced liver fibrosis)<sup>14,111</sup>. We found that the relative percentages of 2n and 4n hepatocytes changed following chronic liver injury, with an increased number of 4n hepatocytes in the pericentral area following pseudotime ordering, an observation that was supported by FACS analysis (Supp. Fig. 13A–F). Furthermore, the diffusion maps, PAGA representation, and histological analysis clearly showed an alteration in the zonation pattern in the CCl<sub>4</sub> treated mice, with profound changes observed in the pericentral hepatocyte compartment (Supp. Fig. 13A, B, Supp. Fig. 14A, B, and Supp. Dataset 9). Despite the fact that zonation was markedly affected by CCl<sub>4</sub> treatment, 2n and 4n hepatocytes coordinate their gene expression levels according to their spatial distribution in the liver lobule. More precisely, 4n hepatocytes are enriched in the peri-central vein area, suggesting that polyploid hepatocytes undergo genetic plasticity and higher adaptation to deal with iterative liver injury<sup>39,43,112</sup>.

Taken together, our data demonstrate that hepatocyte ploidy and zonation are two coordinated but independent processes in which polyploid hepatocytes adjust their gene expression level according to their position within the liver lobule.

## Discussion

Single-cell genomics allows the unbiased exploration of cell states and cell types at single-cell resolution, leading to a revolutionary change in our understanding of liver biology and disease pathogenesis<sup>16</sup>. However, single-cell RNA-seq (scRNA-seq) in the liver is associated with some caveats. First, dissociation protocols including enzymatic or mechanical dissociation lead to changes in the cellular transcriptome<sup>17–19</sup>. Second, dissociation may lead to the underrepresentation of certain cell types due to cell fragility or large cell-size, such as hepatocytes<sup>113</sup>. Third, it relies on the isolation of intact cells from fresh tissues, which hinders its clinical application on human samples. In this context, single-nucleus RNA-seq has emerged as a complementary approach that relies on the unbiased assessment of nuclei from all cells present in a

tissue<sup>20,21,37</sup>. The analysis of the nuclear transcriptome has proven to be very powerful in the study of cell-type diversity in the mouse and human tissues, including brain<sup>21,23–27,114</sup>, spinal cord<sup>53</sup>; breast cancer<sup>54</sup>, kidney<sup>29–32</sup>, lung<sup>28</sup>, heart<sup>34,35</sup>, and a variety of human tumor samples<sup>37</sup>.

We have developed a single-nucleus RNA-seq2 method that improves the lysis of the nuclear membrane and significantly increases the number of transcripts detected per nuclei, comparing favorably with transcript data derived from whole-cell scRNAseq (Fig. 1 and Supp. Fig. 1). Our method is highly sensitive, reproducible and it has been specifically designed for frozen, archived liver samples (Fig. 1). The unbiased sorting of single-nuclei allows the study of all the main cell types within the liver, and this approach can be used to interrogate cellular crosstalk and transcription factor networks (Figs. 1 and 2). Indeed, snRNA-seq2 has the potential for comprehensive analysis of fresh and archived flash-frozen liver samples from both mice and humans.

Notably, scRNA-seq has revealed the high degree of functional specialization of multiple cell types within the liver lobule, for instance for hepatocytes, endothelial cells<sup>9,11,12,52</sup>, and macrophages<sup>14,15,73</sup>. Additionally, it has been shown how changes in mRNA correlate closely with changes in protein levels in the liver<sup>10</sup>. However, little is known about the functional role of polyploid hepatocytes at the single-cell level. Recently, Katsuda *et al.* have shown in rats how genes associated with hepatic zones have differential expression patterns in polyploid hepatocytes<sup>78</sup>. Here, we present a genome-wide analysis of the nuclear transcriptome of individual 2n and 4n nuclei in mice (Figs. 3–5). Our study is focused on the analysis of ploidy based on the DNA content of each nucleus (2n and 4n), and we cannot resolve whether 2n nuclei belong to either diploid cells or binucleated tetraploid cells (2nx2)<sup>41,82</sup>. Moreover, it has been shown that hepatocyte volume does not depend on the number of nuclei but rather on their ploidy status, whereas mono-nucleated hepatocytes (2n) and bi-nucleated hepatocytes (2nx2) have the same cellular volume<sup>8,115</sup>. In young wild-type mice, an incomplete cytokinesis leads to tetraploid cells (2nx2) that are generally associated with the fidelity of chromosome transmission<sup>41,82</sup>. For these reasons, we have focused our analysis on the 2n and 4n nuclei, for which we do not expect chromosomal abnormalities<sup>116,117</sup>.

Additionally, cell size is the major determinant in biochemical reaction rates<sup>118,119</sup> and cell size increases with ploidy. It has been reported that transcript abundance correlates with cellular volume at the single-cell level due to an increase in global transcription in larger cells<sup>89</sup>. During partial hepatectomy, hepatocytes increase cell size before proliferation<sup>120</sup>. In particular, in the liver-specific *Cdk1* knockout mice, larger hepatocytes displayed alterations in lipid and mitochondrial metabolism<sup>121</sup>, suggesting that polyploid hepatocytes could have a more prominent role in energy homeostasis compared with diploid hepatocytes.

Our transcriptomic analyses revealed that 2n and 4n hepatocytes are globally very similar, but 312 genes are found to be differentially expressed between these two groups (Fig. 3). This suggests that if polyploid hepatocytes increase during pathological processes, they could lead to gene expression imbalances of functionally relevant genes (Figs. 3 and 4). Additionally, the transcriptional variability in young mice is also lower in 4n hepatocytes, which has previously been shown using single-molecule fluorescence in situ hybridization (smFISH) for the beta-actin (*Actb*) gene<sup>122</sup>. Whether transcriptional variability changes during ageing<sup>47</sup> or in chronic liver disease still remains to be investigated. Interestingly, the presence of mononucleated tetraploid hepatocytes has been associated with human hepatocellular carcinoma (HCC)<sup>40,86,87,123</sup>. Likewise, the number of

polyploid hepatocytes also increases in models of NAFLD, including ob/ob mice and wild-type mice fed with methionine-choline-deficient diet (MCD) or high-fat diet (HFD)<sup>86,124</sup>. Therefore, alterations in the number of 2n and 4n hepatocytes are observed during NAFLD in both animal models and patients. The transcriptomic analysis of 2n and 4n hepatocytes in archived, frozen samples could open an entirely new strategy to understand disease pathogenesis and search for new therapies to treat liver disease.

The liver is characterized by its regenerative potential. However, the cellular origin that triggers liver homeostasis and repair is still under debate. Recently, complementary studies using different cell lineage-tracing models have shown that all hepatocytes have comparable self-renewal potential<sup>79,93,98</sup>. Our transcriptomic analysis also supports the notion that the vast majority of hepatocytes, regardless of their ploidy level and location within the hepatic lobule, co-expresses stem/progenitor gene markers and have the potential to contribute to liver homeostasis after hepatic injury (Fig. 4B, C).

Furthermore, the broad spatial heterogeneity in the liver shows that key liver functions are zoned<sup>9</sup>, and recently it has been shown that the zonation is perturbed during liver fibrosis<sup>14,15</sup>. In this context, the spatial distribution of polyploid hepatocytes and their functional consequences are still debated<sup>8,41,79,87,109</sup>. We used diffusion pseudotime to infer the pseudospacial ordering of 2n and 4n hepatocytes according to defined markers of liver zonation, and we found that 4n nuclei are distributed throughout the hepatic lobe with an enrichment in the pericentral zone (Fig. 5). These results support previous studies in which polyploid hepatocytes were quantified by lineage-tracing<sup>79</sup> or smFISH for the glutamine synthetase (*Glu1*) gene<sup>109</sup>. Additionally, we have shown that 2n and 4n hepatocytes adjust their gene expression levels in a coordinated manner depending on their spatial location within the liver lobule. In summary, we have found that 4n hepatocytes are enriched 1.3-fold in the pericentral zone, and that hepatocyte ploidy and liver zonation are tightly regulated. Altogether, we have discovered that the division of labor in hepatocytes is linked to both their spatial distribution and ploidy levels, and this crosstalk could be affected during ageing and chronic liver disease.

In summary, our methodology has the potential to investigate previously archived frozen murine and human samples, and interrogate how changes in gene expression correlate with the development and progression of liver disease, taking into account cellular crosstalk and signaling pathways. Our analysis shows that hepatocytes in young livers, without injury or selective pressure, express stem cell markers across the portal-central axis independently of their ploidy status. These findings support recent reports showing that polyploid hepatocytes are proliferative and contribute to liver regeneration. Moreover, hepatocytes with different levels of ploidy adjust their gene dosage according to their position in the liver, showing a functional crosstalk between liver zonation and ploidy. We anticipate that changes in the number of 2n and 4n hepatocytes in the overall liver cell composition will lead either too protective or deleterious outcomes during ageing and chronic liver diseases.

## Methods

**Ethics statement.** This investigation was approved by the Animal Welfare and Ethics Review Board and followed the Cambridge Institute guidelines for the use of animals in experimental studies under Home Office licences PPL 70/7535 until February 2018 and PPL P9855D13B from March 2018. All animal experimentation was carried out in accordance with the Animals (Scientific Procedures) Act 1986 (United Kingdom) and conformed to the Animal Research: Reporting of In Vivo Experiments (ARRIVE) guidelines developed by the National Center for the Replacement, Refinement, and Reduction of Animals in research (NC3Rs).

**Mice and liver tissue collection.** All wild-type C57BL/6 mice were purchased from Charles River UK Ltd (Margate, United Kingdom) and were maintained under specific pathogen-free conditions at the University of Cambridge, CRUK—Cambridge Institute under the auspices of a UK Home Office license. These animal facilities are approved by and registered with the UK Home Office. All mice were maintained in a specific pathogen-free environment in Individually Ventilated Caging units. Mice were kept in a positive pressure system, maintaining a temperature between 19 and 23°C, 55% humidity ( $\pm 10\%$ ), 20 total air changes per hour, and under a 12 h light/dark cycle. Mice had free access to a standard laboratory diet (PicoLab Mouse Diet 20, 5R58) and water. All male C57BL/6 animals were sacrificed by an approved scientist in accordance with Schedule 1 of the Animals (Scientific Procedures) Act 1986. Young 3-month-old male mice were sacrificed and individual pieces of the harvested liver were immediately flash-frozen. A small piece of liver tissue was collected in 4% paraformaldehyde. All young animals for this study were macroscopically inspected for any signs of pathologies or abnormalities.

**Liver fibrosis model.** Carbon tetrachloride (CCl<sub>4</sub>)-induced liver fibrosis was induced as described previously<sup>111</sup>. Mice were injected i.p. with 1 mL/g body weight CCl<sub>4</sub> in a 1:3 ratio with olive oil (0.25  $\mu$ L/g CCl<sub>4</sub>) twice weekly for 6 weeks; livers were harvested 48 h after the final injection<sup>14</sup>.

**Single nuclei isolation/homogenization.** For the single nuclei isolation, the protocol described by Krishnaswami et al.<sup>21</sup> was followed with several modifications. In summary, fresh frozen liver tissues (~3 mm<sup>3</sup>) were homogenized using a 2 mL Dounce homogenizer (Lab Logistics, #9651632) in 1 mL of ice-cold homogenization buffer, HB (250 mM sucrose, 25 mM KCl, 5 mM MgCl<sub>2</sub>, 10 mM Tris buffer, 1  $\mu$ M DTT), 1 x protease inhibitor tablet (Sigma Aldrich Chemie), 0.4 U/ $\mu$ L RNaseIn (Thermo Fisher Scientific), 0.2 U/ $\mu$ L Superasin (Thermo Fisher Scientific), 0.1% Triton X-100 (v/v) and 10  $\mu$ g/mL Hoechst 33342 (Thermo Fisher Scientific) in RNase-free water.

Before the dounce homogenization, the frozen tissue was placed in a pre-chilled petri dish on ice containing 1 mL of HB and finely cut with a pre-chilled scalpel blade until all the pieces can be transferred to the homogenizer using a wide orifice 1 mL tip. Throughout the procedure, all pipetting steps of transferring the samples were made using wide orifice tips (Rainin, #17014297) to minimize shear force.

Always on ice, we performed five very slow strokes with the loose inner tolerance pestle and ten more strokes using the tight inner tolerance pestle. Collecting 10  $\mu$ L of the solution, the nuclei suspension was inspected under a microscope on a Neubauer chamber with 10  $\mu$ L of Trypan blue (0.4%). If aggregates were predominant, up to five more strokes were performed using the tight pestle. The suspension was passed through a 50  $\mu$ m sterilized filter (CellTrics, Symtex, #04-004-2327) into a 5 mL Eppendorf pre-chilled tube washing the Dounce homogenizer with additional 500  $\mu$ L of cold homogenization buffer. Subsequently, the suspension was centrifuged at 1000  $\times$  g for 8 min in an Eppendorf centrifuge at 4°C (5430R). The pellet was then resuspended in 250  $\mu$ L of pre-chilled homogenization buffer. To ensure high quality of single-nuclei suspensions, we performed a density gradient centrifugation clean-up for 20 min, using Iodixanol gradient (Optiprep, D1556, Sigma Aldrich Chemie). The final pellet was resuspended gently in 200  $\mu$ L of nuclei storage buffer (NSB) (166.5 mM sucrose, 5 mM MgCl<sub>2</sub>, 10 mM Tris buffer pH 8.0) containing additional RNase inhibitors 0.2 U/ $\mu$ L Superasin (Thermo Fisher Scientific, #AM2696) and 0.4 U/ $\mu$ L Recombinant RNase Inhibitor (Takara Clontech #2313A). A final visual inspection and counting were performed under a microscope and the single nuclei suspension was filtered through a 35  $\mu$ m cell strainer cap into a pre-chilled FACS tube prior FACS sorting.

**Flow cytometry.** Hoechst dye (Life technologies, #H3570) was used to stain all nuclei during nuclei isolation, allowing to distinguish between diploid and tetraploid nuclei using a FACS sorter (BD FACSAria Fusion 1 and/or BD FACSAria 3) with a 100  $\mu$ m nozzle. Before sorting, 384-well thin-walled PCR plates (BioRad, #HSP3901) were freshly prepared with 940 nL of reaction buffer (following the manufacturer's instructions, only 1  $\mu$ L of 10X reaction buffer is diluted in 2,75  $\mu$ L of water). Herein, this reaction buffer will be termed Lysis Buffer 1—LB1 (Takara kit SMART-Seq v4 Ultra Low input RNA) using a liquid miniaturization robot (Mosquito HV, STP Labtech) and kept on ice. Reaction buffer was prepared following the manufacturer's instruction adding 1  $\mu$ L of RNase Inhibitor in 19  $\mu$ L of 10X Lysis Buffer. The FACS droplet delay and cut-off point were optimized prior to every sorting, the plate holder was cooled at 4°C and all settings and calibrations were done by the FACS operator while samples were processed to avoid additional delays that could lead to RNA degradation of the samples.

In brief, the gating strategy used FSC-A/SSC-A to select intact nuclei stained by Hoechst, in which events above 10<sup>4</sup> are considered Hoechst positive. Subsequently, selection of singlets was performed by FSC-A/FSC-H followed by exclusion of doublets by FSC-A/FSC-W. Identification of different levels of ploidy was performed by FSC-A/Hoechst and FSC-A/Modal as previously described<sup>39</sup> (Supp. Fig. 13).

The sorting accuracy in the 384 well-plates was assessed using a colorimetric method with tetramethyl benzidine substrate (TMB, BioLegend, #421501) and

50 µg/mL of Horseradish Peroxidase (HRP, Life Technologies, #31490)<sup>125</sup>. Only when the plate alignment testing results to single events to more than 95% of the wells, we proceed with the fresh isolated nuclei. The nuclei suspended in NSB was diluted at  $\sim 10 \times 10^4$  nuclei per mL and kept on average 200–1000 events per second as flow rate. In the sorting layout, we sorted diploid (2n) cells in half of the 384 well plate and tetraploid (4n) in the remaining half for each individual. After sorting, every plate was sealed (MicroAmp Thermo Seal lid, #AB0558), shortly vortexed (10 s), centrifuged (prechilled at 4 °C, 2000 x g for 1 min), frozen on dry ice, and stored at –80 °C, until cDNA synthesis.

**snRNAseq-2.** For the generation of double-stranded full-length cDNA, we used the Smart-Seq2 chemistry from a commercial kit (SMART-Seq v4 Ultra Low input RNA, Takara) and we optimized the lysis of the nuclear membrane by adding a supplementary lysis buffer (Lysis Buffer 2, LB2). Firstly, using a low volume liquid handling robot, we miniaturized the volumes reducing the amount four-times and keeping the same ratios for all the reagents, significantly lowering the experimental cost. Liquid handling robots allowed to increase pipetting accuracy, maintained a sterile and temperature-controlled environment, and reduced the user variability and potential cross-contamination. Our additional Lysis Buffer 2 (LB2) consist in a mixture of 0.4 % NP-40 (v/v) final concentration (Life Tech, #85124) and 0.1% Triton-X100 (v/v) final concentration (Fisher, #10671652). Using the Mosquito HV (Labtech STP), we added 2190nL of our additional snRNAseq-2 lysis mix in each well together with the first strand synthesis primers and the spike-in controls (ERCC). Ratios to the final volume (3.125 µL) of the total Lysis Buffer (1 and 2) are shown in parentheses (NP40 2% (2.5/12.5), Triton-X100 1% (1.25/12.5), ERCC spike--in at 1/300.000 dilution (1/12.5) and 3' SMART-seq CDS Primer II A (2/12.5) and additional water (2/12.5). Every plate was thawed directly on a –20 °C chilled holder while LB2 was added by the Mosquito HV robot. Then, the plate was sealed, vortexed vigorously (20 s in a Mixmate (Eppendorf) 2000 rpm), centrifuged (30 s in a pre-chilled Eppendorf 5430R at 4 °C, 2000 x g), and placed in a Thermal cycler (BioRad C1000) for 6 min at 72 °C. We suggest to use the same lot (e.g., #00769049) of ERCCs (Life Technologies, #4456740) per project. ERCC spike-ins were diluted 1 in 10, in water with 0.4 U/µL Recombinant RNase Inhibitor (Takara Clontech #2313A), aliquoted, and stored at –80 °C. Fresh dilution of 1 in 300 000 and 1 in 100 000 were prepared immediately before the first strand synthesis.

Next, reverse transcription and Pre-PCR amplification steps were followed as described by the manufacturer. We maintained our four-times reduced volumes for all steps. As a final modification on the Takara kit protocol, we optimized the PCR cycles program for the cDNA amplification. We used 21 cycles and a PCR programs consisting of: 1 min at 95 °C, [20 s at 95 °C, 4 min at 58 °C, 6 min at 68 °C] × 5, [20 s at 95 °C, 30 s at 64 °C, 6 min at 68 °C] × 9, [30 s at 95 °C, 30 s at 64 °C, 7 min at 68 °C] × 7, 10 min at 72 °C.

Internal ERCC spike-ins were used as positive controls and the cDNA yield was assessed in an Agilent Bioanalyzer with a High Sensitivity DNA kit. In our protocol, we did not perform a bead clean-up before final library preparation. The miniaturization details and Mosquito programs used can be found in the supplementary information. RNA-seq library preparation and sequencing are described in the supplementary information.

**RNA-seq library preparation and sequencing.** Sequencing libraries were prepared using the standard Illumina Nextera XT. DNA Sample Preparation kit (Illumina, #FC-131-1096) and the combination of 384 Combinatorial Dual Indexes (Illumina- Set A to D, #FC-131-2001 to FC-131-2004). Using the Mosquito liquid handling robot, the Nextera XT chemistry was miniaturized<sup>126,127</sup> (Supp. Dataset 10–14 and Supp. Information). All the final 384-pooled libraries were sequenced using Illumina HiSeq4000 NGS sequencer in a paired-end—150 bases length.

**Fluorescence in situ hybridization.** Detection of mouse *Hnf4a* was performed on FFPE sections. Briefly, sections were cut at 3 µm thick, baked for 1 h at 60 °C before deparaffinization in xylene and rehydration through graded ethanol. Pretreatments were carried out using Kretech Tissue Digestion Kit II reagents (Cat # KBI-60004, Leica Biosystems) according to manufacturer's instructions: using 0.2 M HCl for 30 min at room temperature, followed by 3 min wash in milliQ water, 30 min incubation in Pretreatment solution B at 80 °C, 3 min wash in 2X SSC, 30 min in pepsin solution at room temperature, 1 min in milliQ water and 5 min wash in 2X SSC. Slides were dehydrated through graded ethanol. Custom design HNF4 probe (Cat # HNF4A-20-OR, Empire Genomics) was applied to each slide and coverslips were sealed with Fixogum rubber cement (Cat # ICNA11FIXO0125, VWR). Slides were denatured for 5 min at 80 °C before hybridization overnight in a humid chamber at 37 °C. Coverslips were carefully removed and slides were washed in Buffer II (Kretech Tissue Digestion Kit II, KBI6004) at room temperature for 2 min, followed by Buffer I at 72 °C for 2 min and then Buffer II again at room temperature for 2 min. The slides were dehydrated through graded ethanol prior to incubation with Sudan Black B (Cat # 199664, Sigma Aldrich, 0.5% in 70% ethanol, stirred for 2 h in the dark, filtered with 0.22 µm filter) for 10 min. The slides were washed for 5 min each in three changes of PBS prior to mounting with Prolong Gold Antifade Mountant with DAPI (ThermoFisher Scientific, Cat No. P36931,

1:1000). Additional immunofluorescence and image analysis are described in the supplementary information.

**Multiplexed RNAscope and immunofluorescence co-detection.** Simultaneous detection of mRNA for mouse *Cyp2e1-2* and *Cyp2f2*, or *Lgr5* and mouse beta-catenin protein was performed on FFPE sections using Advanced Cell Diagnostics (ACD) RNAscope® 2.5 LS Multiplex Reagent Kit (Cat No. 322800), RNAscope® 2.5 LS Probe Mm-Cyp2f2 (Cat No. 451858), RNAscope® 2.5 LS Probe Mm-Cyp2e1-C2 (Cat No. 402788 -C2), or RNAscope® 2.5 LS Probe Mm-Lgr5 (Cat No. 312178) in combination with ACD Co-Detection Antibody Diluent (Cat No. 323160) (ACD, Hayward, CA, USA), and a mouse monoclonal beta-catenin antibody from BD Biosciences (Cat No. 610154). Briefly, sections were cut at 3 µm thickness, baked for 1 h at 60 °C before loading onto a Bond RX instrument (Leica Biosystems). Slides were deparaffinized and rehydrated on-board prior to heat pre-treatment with Epitope Retrieval Solution 2 (Cat No. AR9640, Leica Biosystems) at 95 °C for 30 min. The antibody was diluted to 1:500 using the ACD Co-Detection Antibody diluent and incubated for 15 min at ambient temperature, before post fixation with 10% neutral buffered formalin (NBF) for 30 min and protein digestion using ACD Enzyme from the Multiplex Reagent kit at 40 °C for 30 min. Probe hybridization and signal amplification were performed according to the manufacturer's instructions. TSA plus-Cy5 (Akoya Biosciences Cat No. NEL745001KT) detection at 1:500 dilution of beta-catenin protein, TSA plus-Cy3 (Akoya Biosciences Cat No. NEL744001KT) detection at 1:500 dilution of *Cyp2e1-2* or *Lgr5*, and TSA plus—Fluorescein detection at 1:500 dilution of *Cyp2f2* (Akoya Biosciences Cat No. NEL741001KT) were performed on the Bond Rx according to the ACD protocol. Slides were then removed from the Bond Rx and mounted using Prolong Diamond (ThermoFisher Cat No P36965). The slides were imaged on the AxioScan (Zeiss) to create whole slide images. Images were captured at 40× magnification, with a resolution of 0.25 microns per pixel. Image analyses were performed using the HALO (Indicalabs) software.

**Read alignment and pre-processing.** Raw sequencing reads were mapped against a customized genome containing both mm10 (GRCm38, assembly version 93), and the ERCC92 sequences<sup>128</sup>. Mapping was done using STAR-2.7.1a with the following parameters --outFilterMultimapNmax 1 --outSAMtype BAM SortedByCoordinate. Potential PCR-duplicates were identified and removed using MarkDuplicates from picard tools version 2.20.2 with the option REMOVE\_DUPLICATES=true. For every single nucleus, reads mapping to individual transcripts were counted and summed per gene using htseq-count version 0.11.3 with the following parameters: -m intersection-nonempty -f bam -r pos -s no --nonunique all -t transcript -i gene\_id --additional-attr=gene\_name.

The resulting raw count matrix was loaded into python and stored as an AnnData object, anndata version 0.7.1. The downstream analysis described below was adapted from Luecken et al.<sup>5,5</sup>. Unless described otherwise, functions implemented in scanpy (version 1.4.5.2.dev6+gfa408dc7) were used in downstream analysis<sup>110</sup>. Additionally, to the count matrix based on full transcripts, count matrices for exonic and intronic reads, respectively were built using featureCounts<sup>129</sup>.

**Quality control and filtering.** Starting with a matrix of 2496 single nuclei and 54 329 genes, nuclei were kept if the percentage of mapped reads corresponding to ERCC spike-in transcripts was higher than 5% but lower than 90%, and if nuclei had >1000 and <7000 genes detected. Then, genes sequenced in fewer than 25 cells (about 1% of the population) and with read count below 250 reads were filtered out. Finally, only the nuclei having less than 7000 genes detected, and a library size of 10 000 to 300 000 reads were kept. This processing yielded a matrix of 2016 single nuclei times 19 340 genes.

**ERCC size factor calculation and normalization.** Two different ERCC dilutions were used (1:100 000 and 1:300 000 respectively). When sequencing to saturation, the proportional amounts of sequenced endogenous transcripts and synthetic spike-ins depend on the input number of endogenous transcripts. The first step in this normalization approach is to calculate ERCC size factors as the sum of ERCC reads per nucleus divided by the mean ERCC reads across all nuclei within one dilution. Subsequently, to normalize the expression matrix, the read counts were first divided by the corresponding gene length in kilobases. Then, the cell coverage was corrected by the following approach: endogenous counts were summed per nucleus, and the sum was divided by 10 000 times the ERCC size factor, yielding the normalization factor per nucleus. Finally, the endogenous reads per nucleus were divided by this factor. Thereby, reads stemming from nuclei with few endogenous reads and many ERCC reads are divided by a smaller factor than reads stemming from nuclei with many endogenous reads and proportionally few ERCC reads.

$$x'_{ij} = \frac{\frac{x_{ij}}{L_j}}{\frac{\sum_i x_{ij}}{10000 \cdot sf_i}} \quad (1)$$

where  $x'_{ij}$  is the normalized count of gene  $j$  in cell  $i$ ,  $L_j$  is the length of gene  $j$  and  $sf_i$  is the ERCC size factor of cell  $i$ .

After normalization, cells with more than 50 000 gene length-normalized total counts were removed. Technical replicates SNI-234(R2) and SNI-235(R2) that contain the same cells as SNI-160 and SNI-116, respectively, were not further considered (Supp. Dataset 1). This resulted in a final matrix of 1649 nuclei times 19 258 genes. The normalized counts were log-transformed by applying a natural logarithm to one plus the values in the count matrix. Batch effects were corrected for using *combat* with plates as a covariate<sup>130</sup>. Description of clustering and visualization are further described in the supplementary information.

**Differential expression analysis.** In order to find differential expressed genes between groups of interest, Welch's t-tests were done as implemented in the scanpy function `rank_genes_groups` with the parameters `method="t-test"` and `n_genes=19258`. First, this was done between cell types, considering 2n and 4n hepatocytes together as one group, and then, only hepatocytes were taken in order to compare 2n to 4n. Genes that had a log<sub>2</sub> fold change of greater than 0.5 and a Bonferroni-adjusted *p*-value below 0.05 were considered significantly upregulated; genes with a log<sub>2</sub> fold change smaller than -0.5 and a Bonferroni-adjusted *p*-value below 0.05 were considered significantly downregulated.

To further investigate changes in the distribution of a given gene between 2n and 4n hepatocytes, a Kolmogorov–Smirnov test was performed. Genes were called as significantly changing their gene expression distribution between 2n and 4n if they had a *p*-value below 0.05 and a test statistic of >0.15.

For visualization purposes, 40 nuclei were randomly selected per cell type (for cell types with <40 nuclei, all nuclei were taken) and their respective top five differential expressed genes (based on their score calculated by `rank_genes_groups`) were taken to create a heatmap (Fig. 2A) using the *ComplexHeatmap* package in R. For visualization purposes, genes with a mean expression between 0.1 and 100 were selected in the MA plot (Fig. 3C).

**Co-expression analysis of liver stem cell markers.** For this analysis, only hepatocytes were used. A heatmap of known hepatic stem cell markers showed expression of these markers not only in diploid but also in tetraploid nuclei. To investigate whether these markers are expressed in the same nuclei, the matrix was binarized after log-transformation and stored as an additional layer in the *AnnData* object. This binary matrix was a subset to only contain the stem cell markers. Nuclei not expressing any of the stem cell markers were removed, resulting in a matrix of 364 nuclei times 11 stem cell markers. Based on this, pairwise Jaccard distance between the stem cell markers was calculated.

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \quad (2)$$

where *X* is the binary expression vector of gene *j*<sub>1</sub> across the nuclei, and *Y* is the binary expression vector of gene *j*<sub>2</sub> across the nuclei.

Furthermore, pairwise Jaccard distances between nuclei and genes, respectively, were used to calculate linkage for hierarchical clustering. By this approach, the gene's expression level per nucleus is neglected and only its presence or absence is counted. Additional downstream analyses are described in the supplementary information.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

All raw sequencing data is deposited and publicly available in ArrayExpress under accession numbers "E-MTAB-9333"<sup>131</sup> and "E-MTAB-10223"<sup>132</sup>. Additional publicly available data used for method comparison was obtained from either GEO (Accession numbers "GSE84498": GSE84498\_umatib.txt.gz<sup>12</sup>, "GSE124395": GSE124395\_Normalhumanlivercellatlasdata.txt.gz<sup>52</sup>, "GSE148339": GSE148339\_liver.integrated.final.RData.gz<sup>33</sup> or from <https://doi.org/10.6084/m9.figshare.5829687.v7> and <https://doi.org/10.6084/m9.figshare.5968960.v2><sup>51</sup>). All other relevant data supporting the key findings of this study are available within the article and its Supplementary Information files or from the corresponding author upon reasonable request. Source data are provided with this paper. A reporting summary for this Article is available as a Supplementary Information file. Source data are provided with this paper.

## Code availability

The scripts (jupyter notebook and R) containing code to reproduce the analysis are publicly available at [http://github.com/celiantez/snRNA-seq2\\_young](http://github.com/celiantez/snRNA-seq2_young)<sup>133</sup>. Packages used for the analysis include *ComplexHeatmap* (v2.6.2)<sup>131,134</sup>, *matplotlib* (v3.2.1)<sup>132,135</sup>, *pandas* (v1.0.3)<sup>133,136</sup>, *numpy* (v1.18.1)<sup>14</sup>, *seaborn* (v0.10.0)<sup>137</sup>, and *scipy* (v1.4.1)<sup>138</sup>. The version of the code at the time of publication is stored in Zenodo: <https://doi.org/10.5281/zenodo.4694749><sup>133</sup>.

Received: 3 August 2020; Accepted: 24 June 2021;

Published online: 12 July 2021

## References

- Martinez-Jimenez, C. P., Castell, J. V., Gomez-Lechon, M. J. & Jover, R. Transcriptional activation of CYP2C9, CYP1A1, and CYP1A2 by hepatocyte nuclear factor 4alpha requires coactivators peroxisomal proliferator activated receptor-gamma coactivator 1alpha and steroid receptor coactivator 1. *Mol. Pharm.* **70**, 1681–1692 (2006).
- Martinez-Jimenez, C. P., Gomez-Lechon, M. J., Castell, J. V. & Jover, R. Underexpressed coactivators PGC1alpha and SRC1 impair hepatocyte nuclear factor 4 alpha function and promote dedifferentiation in human hepatoma cells. *J. Biol. Chem.* **281**, 29840–29849 (2006).
- Martinez-Jimenez, C. P., Jover, R., Donato, M. T., Castell, J. V. & Gomez-Lechon, M. J. Transcriptional regulation and expression of CYP3A4 in hepatocytes. *Curr. Drug Metab.* **8**, 185–194 (2007).
- Castell, J. V., Jover, R., Martinez-Jimenez, C. P. & Gomez-Lechon, M. J. Hepatocyte cell lines: their use, scope and limitations in drug metabolism studies. *Expert Opin. Drug Metab. Toxicol.* **2**, 183–212 (2006).
- Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
- Martinez-Jimenez, C. P., Kyrmizi, I., Cardot, P., Gonzalez, F. J. & Talianidis, I. Hepatocyte nuclear factor 4alpha coordinates a transcription factor network regulating hepatic fatty acid metabolism. *Mol. Cell Biol.* **30**, 565–577 (2010).
- Schmidt, D. et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**, 1036–1040 (2010).
- Morales-Navarrete H., et al. A versatile pipeline for the multi-scale digital reconstruction and quantitative analysis of 3D tissue architecture. *Elife* **4**, e11214 (2015).
- Ben-Moshe, S. & Itzkovitz, S. Spatial heterogeneity in the mammalian liver. *Nat. Rev. Gastroenterol. Hepatol.* **16**, 395–410 (2019).
- Ben-Moshe, S. et al. Spatial sorting enables comprehensive characterization of liver zonation. *Nat. Metab.* **1**, 899–911 (2019).
- Halpern, K. B. et al. Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells. *Nat. Biotechnol.* **36**, 962–970 (2018).
- Halpern, K. B. et al. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature* **542**, 352–356 (2017).
- Moor, A. E. & Itzkovitz, S. Spatial transcriptomics: paving the way for tissue-level systems biology. *Curr. Opin. Biotechnol.* **46**, 126–133 (2017).
- Dobie, R. et al. Single-cell transcriptomics uncovers zonation of function in the mesenchyme during liver fibrosis. *Cell Rep.* **29**, 1832–1847 (2019). e1838.
- Ramachandran, P. et al. Resolving the fibrotic niche of human liver cirrhosis at single-cell level. *Nature* **575**, 512–518 (2019).
- Ramachandran P., Matchett K. P., Dobie R., Wi I lson-Kanamori J. R., Henderson N. C. Single-cell technologies in hepatology: new insights into liver biology and disease pathogenesis. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 457–472 (2020).
- van den Brink, S. C. et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 (2017).
- Gomez-Lechon, M. J., Donato, M. T., Castell, J. V. & Jover, R. Human hepatocytes in primary culture: the choice to investigate drug metabolism in man. *Curr. Drug Metab.* **5**, 443–462 (2004).
- Gomez-Lechon, M. J., Donato, M. T., Castell, J. V. & Jover, R. Human hepatocytes as a tool for studying toxicity and drug metabolism. *Curr. Drug Metab.* **4**, 292–312 (2003).
- Grindberg, R. V. et al. RNA-sequencing from single nuclei. *Proc. Natl Acad. Sci. USA* **110**, 19802–19807 (2013).
- Krishnaswami, S. R. et al. Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. *Nat. Protoc.* **11**, 499–524 (2016).
- Kalish, B. T. et al. Single-nucleus RNA sequencing of mouse auditory cortex reveals critical period triggers and brakes. *Proc. Natl Acad. Sci. USA* **117**, 11744–11752 (2020).
- Lake, B. B. et al. A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. *Sci. Rep.* **7**, 6031 (2017).
- Lake, B. B. et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586–1590 (2016).
- Habib, N. et al. Div-seq: single-nucleus RNA-seq reveals dynamics of rare adult newborn neurons. *Science* **353**, 925–928 (2016).
- Habib, N. et al. DroNc-seq: deciphering cell types in human archived brain tissues by massively-parallel single nucleus RNA-seq. *bioRxiv* **9**, 1–12 (2017).
- Hu, P. et al. Dissecting cell-type composition and activity-dependent transcriptional state in mammalian brains by massively parallel single-nucleus RNA-seq. *Mol. Cell* **68**, 1006–1015 (2017). e1007.
- Koenitzer, J. R., Wu, H., Atkinson, J. J., Brody, S. L. & Humphreys B. D. Single-nucleus RNA-sequencing profiling of mouse lung, reduced dissociation bias and improved rare cell-type detection compared with single-cell RNA sequencing. *Am. J. Respir. Cell. Mol. Biol.* **63**, 739–747 (2020).
- Wu, H., Kiritani, Y., Donnelly, E. L. & Humphreys, B. D. Advantages of single-nucleus over single-cell RNA sequencing of adult kidney: rare cell types and novel cell states revealed in fibrosis. *J. Am. Soc. Nephrol.* **30**, 23–32 (2019).

30. Wilson, P. C. et al. The single-cell transcriptomic landscape of early human diabetic nephropathy. *Proc. Natl Acad. Sci. USA* **116**, 19619–19625 (2019).
31. Denisenko, E. et al. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol.* **21**, 130 (2020).
32. Lake, B. B. et al. A single-nucleus RNA-sequencing pipeline to decipher the molecular anatomy and pathophysiology of human kidneys. *Nat. Commun.* **10**, 2832 (2019).
33. Nault, R., Fader, K. A., Bhattacharya, S. & Zacharewski, T. R. Single-nuclei RNA sequencing assessment of the hepatic effects of 2,3,7,8-Tetrachlorodibenzo-p-dioxin. *Cell Mol. Gastroenterol. Hepatol.* **11**, 147–159 (2021).
34. Wolfien M., et al. Single-nucleus sequencing of an entire mammalian heart: cell type composition and velocity. *Cells* **9**, 318 (2020).
35. Selewa, A. et al. Systematic comparison of high-throughput single-cell and single-nucleus transcriptomes during cardiomyocyte differentiation. *Sci. Rep.* **10**, 1535 (2020).
36. Ding J., et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38**, 737–746 (2020).
37. Slyper, M. et al. A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nat. Med.* **26**, 792–802 (2020).
38. Abdelmoez, M. N. et al. SINC-seq: correlation of transient gene expressions between nucleus and cytoplasm reflects single-cell physiology. *Genome Biol.* **19**, 66 (2018).
39. Duncan, A. W. et al. The ploidy conveyor of mature hepatocytes as a source of genetic variation. *Nature* **467**, 707–710 (2010).
40. Duncan, A. W. et al. Frequent aneuploidy among normal human hepatocytes. *Gastroenterology* **142**, 25–28 (2012).
41. Donne R., Saroul-Ainama M., Cordier P., Celton-Morizur S., Desdouets C. Polyploidy in liver development, homeostasis and disease. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 391–405 (2020).
42. Epstein, C. J. Cell size, nuclear content, and the development of polyploidy in the Mammalian liver. *Proc. Natl Acad. Sci. USA* **57**, 327–334 (1967).
43. Kreutz, C. et al. Hepatocyte ploidy is a diversity factor for liver homeostasis. *Front. Physiol.* **8**, 862 (2017).
44. Lu, P. et al. Microarray analysis of gene expression of mouse hepatocytes of different ploidy. *Mamm. Genome* **18**, 617–626 (2007).
45. Buettner, F. et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155–160 (2015).
46. Lun, A. T. L. & Marioni, J. C. Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data. *Biostatistics* **18**, 451–464 (2017).
47. Martinez-Jimenez, C. P. et al. Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science* **355**, 1433–1436 (2017).
48. Lun, A. T. L., Calero-Nieto, F. J., Haim-Vilmovsky, L., Gottgens, B. & Marioni, J. C. Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. *Genome Res.* **27**, 1795–1806 (2017).
49. Eling, N., Morgan, M. D. & Marioni, J. C. Challenges in measuring and understanding biological noise. *Nat. Rev. Genet.* **20**, 536–548 (2019).
50. Svensson, V. et al. Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14**, 381–387 (2017).
51. Tabula Muris, C. et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
52. Aizarani, N. et al. A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature* **572**, 199–204 (2019).
53. Sathyamurthy, A. et al. Massively parallel single nucleus transcriptional profiling defines spinal cord neurons and their activity during behavior. *Cell Rep.* **22**, 2216–2225 (2018).
54. Gao, R. et al. Nanogrid single-nucleus RNA sequencing reveals phenotypic diversity in breast cancer. *Nat. Commun.* **8**, 228 (2017).
55. Mereu E., et al. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat. Biotechnol.* **38**, 747–755 (2020).
56. Segal, J. M. et al. Single cell analysis of human foetal liver captures the transcriptional profile of hepatobiliary hybrid progenitors. *Nat. Commun.* **10**, 3350 (2019).
57. Celton-Morizur, S. & Desdouets, C. Polyploidization of liver cells. *Adv. Exp. Med. Biol.* **676**, 123–135 (2010).
58. McDavid, A., Finak, G. & Gottardo, R. The contribution of cell cycle to heterogeneity in single-cell RNA-seq data. *Nat. Biotechnol.* **34**, 591–593 (2016).
59. Scialdone, A. et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85**, 54–61 (2015).
60. Martinez-Jimenez, C. P., Gomez-Lechon, M. J., Castell, J. V. & Jover, R. Transcriptional regulation of the human hepatic CYP3A4: identification of a new distal enhancer region responsive to CCAAT/enhancer-binding protein beta isoforms (liver activating protein and liver inhibitory protein). *Mol. Pharm.* **67**, 2088–2101 (2005).
61. Montagner, A. et al. Liver PPARalpha is crucial for whole-body fatty acid homeostasis and is protective against NAFLD. *Gut* **65**, 1202–1214 (2016).
62. Lee, S. S. et al. Targeted disruption of the alpha isoform of the peroxisome proliferator-activated receptor gene in mice results in abolishment of the pleiotropic effects of peroxisome proliferators. *Mol. Cell Biol.* **15**, 3012–3022 (1995).
63. Pawlak, M., Lefebvre, P. & Staels, B. Molecular mechanism of PPARalpha action and its impact on lipid metabolism, inflammation and fibrosis in non-alcoholic fatty liver disease. *J. Hepatol.* **62**, 720–733 (2015).
64. Yamashita, H. et al. A glucose-responsive transcription factor that regulates carbohydrate metabolism in the liver. *Proc. Natl Acad. Sci. USA* **98**, 9116–9121 (2001).
65. Benhamed, F. et al. The lipogenic transcription factor ChREBP dissociates hepatic steatosis from insulin resistance in mice and humans. *J. Clin. Invest* **122**, 2176–2194 (2012).
66. Samuel, V. T. & Shulman, G. I. Nonalcoholic fatty liver disease as a nexus of metabolic and hepatic diseases. *Cell Metab.* **27**, 22–41 (2018).
67. Casagrande, V. et al. Hepatocyte specific TIMP3 expression prevents diet dependent fatty liver disease and hepatocellular carcinoma. *Sci. Rep.* **7**, 6747 (2017).
68. Nikolaou, N. et al. AKR1D1 is a novel regulator of metabolic phenotype in human hepatocytes and is dysregulated in non-alcoholic fatty liver disease. *Metabolism* **99**, 67–80 (2019).
69. Xiong, X. et al. Landscape of intercellular crosstalk in healthy and NASH liver revealed by single-cell secretome gene analysis. *Mol. Cell* **75**, 644–660 (2019). e645.
70. Wang, X. et al. Hepatocyte TAZ/WWTR1 promotes inflammation and fibrosis in nonalcoholic steatohepatitis. *Cell Metab.* **24**, 848–862 (2016).
71. van Koppen, A. et al. Uncovering a predictive molecular signature for the onset of NASH-related fibrosis in a translational NASH mouse model. *Cell Mol. Gastroenterol. Hepatol.* **5**, 83–98 (2018). e10.
72. Ramnath D., et al. Hepatic expression profiling identifies steatosis-independent and steatosis-driven advanced fibrosis genes. *JCI Insight* **3**, e120274 (2018).
73. MacParland, S. A. et al. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat. Commun.* **9**, 4383 (2018).
74. Krenkel O., Hundertmark J., Ritz T. P., Weiskirchen R., Tacke F. Single cell RNA sequencing identifies subsets of hepatic stellate cells and myofibroblasts in liver fibrosis. *Cells* **8**, 503 (2019).
75. Pepe-Mooney, B. J. et al. Single-cell analysis of the liver epithelium reveals dynamic heterogeneity and an essential role for YAP in homeostasis and regeneration. *Cell Stem Cell* **25**, 23–38 (2019). e28.
76. Su, X. et al. Single-cell RNA-Seq analysis reveals dynamic trajectories during mouse liver development. *BMC Genomics* **18**, 946 (2017).
77. Fox, D. T. & Duronio, R. J. Endoreplication and polyploidy: insights into development and disease. *Development* **140**, 3–12 (2013).
78. Katsuda, T. et al. Transcriptomic dissection of hepatocyte heterogeneity: linking ploidy, zonation, and stem/progenitor cell characteristics. *Cell Mol. Gastroenterol. Hepatol.* **9**, 161–183 (2020).
79. Matsumoto, T., Wakefield, L., Tarlow, B. D. & Grompe, M. In vivo lineage tracing of polyploid hepatocytes reveals extensive proliferation during liver regeneration. *Cell Stem Cell* **26**, 34–47 (2020). e33.
80. Davoli, T. & de Lange, T. The causes and consequences of polyploidy in normal development and cancer. *Annu. Rev. Cell Dev. Biol.* **27**, 585–610 (2011).
81. Wang, M. J., Chen, F., Lau, J. T. Y. & Hu, Y. P. Hepatocyte polyploidization and its association with pathophysiological processes. *Cell Death Dis.* **8**, e2805 (2017).
82. Gjelsvik, K. J., Besen-McNally, R. & Losick, V. P. Solving the polyploid mystery in health and disease. *Trends Genet.* **35**, 6–14 (2019).
83. Turner, R. et al. Human hepatic stem cell and maturational liver lineage biology. *Hepatology* **53**, 1035–1045 (2011).
84. Schwartz-Arad, D., Zajicek, G. & Bartfeld, E. The streaming liver IV: DNA content of the hepatocyte increases with its age. *Liver* **9**, 93–99 (1989).
85. Kudryavtsev, B. N., Kudryavtseva, M. V., Sakuta, G. A. & Stein, G. I. Human hepatocyte polyploidization kinetics in the course of life cycle. *Virchows Arch. B Cell Pathol. Incl. Mol. Pathol.* **64**, 387–393 (1993).
86. Genric, G. et al. Oxidative stress promotes pathologic polyploidization in nonalcoholic fatty liver disease. *J. Clin. Invest.* **125**, 981–992 (2015).
87. Bou-Nader M., et al. Polyploidy spectrum: a new marker in HCC classification. *Gut* **69**, 355–364 (2019).
88. Cobbina, E. & Akhlaghi, F. Non-alcoholic fatty liver disease (NAFLD) - pathogenesis, classification, and effect on drug metabolizing enzymes and transporters. *Drug Metab. Rev.* **49**, 197–211 (2017).
89. Padovan-Merhar, O. et al. Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Mol. Cell* **58**, 339–352 (2015).
90. Grun, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11**, 637–640 (2014).
91. Goolam, M. et al. Heterogeneity in Oct4 and Sox2 targets biases. *Cell Fate 4-Cell Mouse Embryos. Cell* **165**, 61–74 (2016).

92. Canchola, J. A., Tang, S., Hemyari, P., Paxinos, E. & E., M. Correct use of percent coefficient of variation (%CV) formula for log-transformed data. *MOJ Proteom. Bioinform.* **6**, 316–317 (2017).
93. Sun, T. et al. AXIN2(+) pericentral hepatocytes have limited contributions to liver homeostasis and regeneration. *Cell Stem Cell* **26**, 97–107 (2020). e106.
94. Wilkinson, P. D. et al. Polyploid hepatocytes facilitate adaptation and regeneration to chronic liver injury. *Am. J. Pathol.* **189**, 1241–1255 (2019).
95. Zhang, S. et al. Knockdown of anillin actin binding protein blocks cytokinesis in hepatocytes and reduces liver tumor development in mice without affecting regeneration. *Gastroenterology* **154**, 1421–1434 (2018).
96. Zhang, S. et al. The polyploid state plays a tumor-suppressive role in the liver. *Dev. Cell* **44**, 447–459 (2018). e445.
97. Wang, B., Zhao, L., Fish, M., Logan, C. Y. & Nusse, R. Self-renewing diploid Axin2(+) cells fuel homeostatic renewal of the liver. *Nature* **524**, 180–185 (2015).
98. Chen, F. et al. Broad distribution of hepatocyte proliferation in liver homeostasis and regeneration. *Cell Stem Cell* **26**, 27–33 (2020). e24.
99. Levandowsky, M. & Winter, D. Distance between sets. *Nature* **234**, 34–35 (1971).
100. Fuxman Bass, J. I. et al. Using networks to measure similarity between genes: association index selection. *Nat. Methods* **10**, 1169–1176 (2013).
101. Li, X., Lalic, J., Baeza-Centurion, P., Dhar, R. & Lehner, B. Changes in gene expression predictably shift and switch genetic interactions. *Nat. Commun.* **10**, 3886 (2019).
102. Shalek, A. K. et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
103. Chalancon, G. et al. Interplay between gene expression noise and regulatory network architecture. *Trends Genet.* **28**, 221–232 (2012).
104. Anatskaya, O. V. & Vinogradov, A. E. Genome multiplication as adaptation to tissue survival: evidence from gene expression in mammalian heart and liver. *Genomics* **89**, 70–80 (2007).
105. Font-Burgada, J. et al. Hybrid periportal hepatocytes regenerate the injured liver without giving rise to cancer. *Cell* **162**, 766–779 (2015).
106. Lin, S. et al. Distributed hepatocytes expressing telomerase repopulate the liver in homeostasis and injury. *Nature* **556**, 244–248 (2018).
107. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).
108. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
109. Tanami, S. et al. Dynamic zonation of liver polyploidy. *Cell Tissue Res.* **368**, 405–410 (2017).
110. Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).
111. Henderson, N. C. et al. Targeting of alpha integrin identifies a core molecular pathway that regulates fibrosis in several organs. *Nat. Med.* **19**, 1617–1624 (2013).
112. Duncan, A. W. et al. Aneuploidy as a mechanism for stress-induced liver adaptation. *J. Clin. Invest.* **122**, 3307–3315 (2012).
113. Lafzi, A., Moutinho, C., Picelli, S. & Heyn, H. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nat. Protoc.* **13**, 2742–2757 (2018).
114. Lacar, B. et al. Nuclear RNA-seq of single neurons reveals molecular signatures of activation. *Nat. Commun.* **7**, 11022 (2016).
115. Miyaoka, Y. & Miyajima, A. To divide or not to divide: revisiting liver regeneration. *Cell Div.* **8**, 8 (2013).
116. Knouse, K. A., Wu, J., Whittaker, C. A. & Amon, A. Single cell sequencing reveals low levels of aneuploidy across mammalian tissues. *Proc. Natl Acad. Sci. USA* **111**, 13409–13414 (2014).
117. Knouse, K. A., Lopez, K. E., Bachofner, M. & Amon, A. Chromosome segregation fidelity in epithelia requires tissue architecture. *Cell* **175**, 200–211 (2018). e213.
118. Schmoller, K. M. & Skotheim, J. M. The biosynthetic basis of cell size control. *Trends Cell Biol.* **25**, 793–802 (2015).
119. Vargas-Garcia, C. A., Ghusinga, K. R. & Singh, A. Cell size control and gene expression homeostasis in single-cells. *Curr. Opin. Syst. Biol.* **8**, 109–116 (2018).
120. Miyaoka, Y. et al. Hypertrophy and unconventional cell division of hepatocytes underlie liver regeneration. *Curr. Biol.* **22**, 1166–1175 (2012).
121. Miettinen, T. P. et al. Identification of transcriptional and metabolic programs related to mammalian cell size. *Curr. Biol.* **24**, 598–608 (2014).
122. Bahar Halpern, K. et al. Bursty gene expression in the intact mammalian liver. *Mol. Cell* **58**, 147–156 (2015).
123. Margall-Ducos, G., Celton-Morizur, S., Couton, D., Bregerie, O. & Desdouets, C. Liver tetraploidization is controlled by a new process of incomplete cytokinesis. *J. Cell Sci.* **120**, 3633–3639 (2007).
124. Celton-Morizur, S., Merlen, G., Couton, D., Margall-Ducos, G. & Desdouets, C. The insulin/Akt pathway controls a specific cell division program that leads to generation of binucleated tetraploid liver cells in rodents. *J. Clin. Invest.* **119**, 1880–1887 (2009).
125. Rodrigues, O. R. & Monard, S. A rapid method to verify single-cell deposition setup for cell sorters. *Cytom. Part A: J. Int. Soc. Anal. Cytol.* **89**, 594–600 (2016).
126. Gasch, A. P. et al. Single-cell RNA sequencing reveals intrinsic and extrinsic regulatory heterogeneity in yeast responding to stress. *PLoS Biol.* **15**, e2004050 (2017).
127. Mora-Castilla, S. et al. Miniaturization technologies for efficient single-cell library preparation for next-generation sequencing. *J. Lab Autom.* **21**, 557–567 (2016).
128. Lee, H., Pine, P. S., McDaniel, J., Salit, M. & Oliver, B. External RNA controls consortium beta version update. *J. Genomics* **4**, 19–22 (2016).
129. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
130. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
131. Richter M. L., et al. snRNA-seq2 investigating hepatocyte ploidy using frozen liver samples from young healthy C57Bl6/J mice and from mice treated with CCl4. *ArrayExpress* <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-9333/> (2021).
132. Richter M. L., et al. snRNA-seq from frozen liver samples from young healthy C57Bl6/J mice using 10X Genomics. *ArrayExpress* <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-10223/> (2021).
133. Richter M. L., et al. SnRNA-seq2 computational pipeline. *Zenodo* <https://doi.org/10.5281/zenodo.4694749> (2021).
134. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
135. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
136. Reback J., et al. pandas-dev/pandas: Pandas 1.0.3. *Zenodo* <https://doi.org/10.5281/zenodo.3715232> (2020).
137. Waskom M. L. Seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
138. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

## Acknowledgements

This research was supported by the Helmholtz Pioneer Campus (C.P.M.-J., M.L.R.), Cancer Research UK (C.P.M.-J.—SWAG/051, E.L.I., P.C.), the Janet Thornton Fellowship (WT098051 to C.P.M.-J.), Impuls-und Vernetzungsfonds of the Helmholtz-Gemeinschaft (VH108 NG-1219 to M.C.-T.), Incubator grant sparse2big (#ZT-1-0007 to A.D.), The European Association for the Study of the Liver (Andrew K. Burroughs Short-term Training Fellowship to K.Y.), The University of Edinburgh (Chancellor's Fellowship to C.A.V.), and Helmholtz future topic Aging and Metabolic Programming (AMPro ZT-0026 to I.K.D). N.C.H. is supported by a Wellcome Trust Senior Research Fellowship in Clinical Science (ref. 219542/Z/19/Z). We thank the CRUK-CI core facilities, including Genomics (P. Coupland), Flow Cytometry (J. Markovic-Djuric, M. Strzelecki and R. Grenfell) for technical assistance, the Biological Resources Unit (A. Mowbray, N. Jacobs, M. Clayton, and M. Mitchell) for animal husbandry, and the histopathology/ISH core facility in particular J. Miller and J. Jones for assistance with the RNAscope Integrated Co-Detection Workflow and C. Brodie for Halo Image analysis. We also thank the Core Genomics at HMGU (I. de la Rosa), and the Core Facility Cell Analysis at TranslaTUM (R. Mishra). We are grateful to Dr. M. Hartman, Dr. A. Schröder, and Ms. A. Barden (Helmholtz Pioneer Campus) for their legal, managerial, and administrative support. We thank Dr. N. Eling for initial data analysis, and Dr. R. Kamies for her comments on the manuscript. We are grateful to Dr. D. T. Odom (CRUK, DKFZ) for his support and scientific advice.

## Author contributions

I.K.D., K.Y., and C.P.M.-J., designed experiments; I.D., K.Y., and C.P.M.-J. performed experiments; M.L.R. performed computational analysis. M.L.R., I.K.D., K.Y., M.C.-T., and C.P.M.-J. interpreted the data; E.L.I., and P.C. provided technical assistance; A.D. and C.V. provided computational assistance and advise; K.P.M. and N.C.H. designed and performed the mouse model of liver fibrosis. C.P.M.-J. wrote the paper with input from all co-authors. C.P.M.-J. and K.Y. provided figure editing and design. All authors commented on and approved the paper.

## Funding

Open Access funding enabled and organized by Projekt DEAL.

## Competing interests

The authors declare no competing interests.

**Additional information**

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-24543-5>.

**Correspondence** and requests for materials should be addressed to M.C.-T. or C.P.M.-J.

**Peer review information** *Nature Communications* thanks Nicholas Webster and the other anonymous reviewer(s) for their contribution to the peer review of this work.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021