Supplementary Information – Computational Methods

Data preprocessing

In this section we describe the preprocessing steps taken to establish the data matrix of hepatocyte single cell gene expression data (Table S1). The main steps were the removal of non-parenchymal cells and the subtraction of background signal for each gene. Such background is predominantly caused by barcode switching and sequencing errors[1]. The resulting data matrix consisted of $G = 27297$ genes and $N = 1415$ cells. Each cell in the matrix held the UMI count of gene g in cell c, denoted UMI_g^c.

As a first preprocessing step, we removed non-parenchymal cells. To this end we considered cell-type specific sets of genes – the Kupffer cell genes: Clec4f, Csf1r, C1qc, C1qa and C1qb, the endothelial cell genes: Kdr, Egfl7, Igfbp7 and Aqp1, and the hepatocyte genes: Apoa1, Apob, Pck1, G6pc and Ttr. Cells for which the aggregated transcript counts of either the Kupffer set or the endothelial set exceeded the aggregated counts of the hepatocyte set were removed from further analysis. We next calculated the background expression level for each gene, based on wells in which RNA extraction or amplification failed. These were defined as wells for which the aggregated sum of all ERCC spike-in molecules was greater than 0.04 of the aggregated sum of the non-ERCC molecules. These wells invariably included the four empty wells in each plate. The background expression of each gene was then set to the mean number of molecules across all background cells, and was subtracted from the expression of that gene in the remaining cells. Negative counts were set to zero.

Following subtraction of background levels, we filtered-out cells with total number of molecules smaller than 1000 UMI or larger than 30,000 UMI. Finally, we discarded cells in which the expression level of Albumin was lower than 1% of the total cellular UMI. This threshold was chosen based on previous bulk estimates that estimated the average Albumin gene expression to be ~10% of the total number of hepatocyte cellular mRNA[2], and our findings using smFISH that Albumin levels rarely decreased below 10% of this average. Our preprocessing steps yielded 1415 high-confidence hepatocytes.

Algorithm for spatial reconstruction of liver zonation profiles

In this section we describe our algorithm for reconstructing zonation profiles by combining smFISH measurements of landmark genes and single cell RNAseq measurements. Inference of spatial coordinates of cells from single cell RNAseq and traditional binary in-situ hybridization have been recently described[3,4]. Our method differs from these studies in two main aspects: 1) Here we used single molecule FISH rather than traditional FISH, yielding precise continuous cellular gene expression levels of individual cells at defined spatial coordinates, rather than binary expression. 2) While our inference provides the estimated lobule position of each cell, our main goal is reconstruction of zonation profiles. We thus utilize the complete posterior probability vectors of cells to belong to any coordinate, thus maximizing the information used to reconstruct these profiles.

1. Lobule geometry

For simplicity we considered one-dimensional lobule geometry, where lobules consist of hexagonal shaped columns with infinite height and radial symmetry. We further assumed that the relevant coordinate is the distance of each cell from the closest central vein, discretized into 9 lobule layers (layer z =1 begins at the central vein whereas layer z =9 ends at the portal node). More complicated topologies that include additional dimensions, such as distance of each cell to the closest portal node / portal tract or vertical distance along the lobule column can be considered in future work. Our topology defines a prior probability of sampling a cell at lobule layer z, $P_{prior}(z)$ (Extended Data Fig. 3c,d).

2. Probabilistic reconstruction of zonation profiles

Our goal was to infer the gene zonation matrix, which held the average expression level (in fraction of total cellular mRNA) attributed to every gene g at lobule layer z, i.e. $E_{g,z}$. Since the RNA yield in scRNAseq experiments is variable among cells, we normalized the background subtracted expression matrix UMI_g^c by dividing the number of UMI of each gene in each cell by the total number of UMI for that cell to obtain the data matrix:

$[1] D_{g,c} = UMI_g^c / \sum_{g=1}^{G} UMI_g^c$

with g=1..G genes and c=1.. N cells (G=27297 and N=1415). This normalization facilitated pooling multiple cells to estimate the average expression in each lobule layer.

To compute the gene zonation matrix, we multiplied the data matrix $D_{g,c}$ (the expression of each gene in each cell) by a weighted probability matrix, $W_{c,z}$ (the weighted probability for each cell to be at each zone),

$[2] E_{g,z} = D_{g,c} \cdot W_{c,z}$,

where we used bootstrapping to obtain standard errors for the mean zonation profiles.

The key step in our algorithm was to estimate weighted probability matrix, $W_{c,z}$. To this end we estimated the posterior probability matrix:

$$[3]P_{c,z} = P^{c}_{posterior}(z) = \frac{P^{c}_{sampling}\left(\overrightarrow{U^{c}}\middle|z\right) \cdot P_{prior}(z)}{\sum_{z=1}^{9} P^{c}_{sampling}\left(\overrightarrow{U^{c}}\middle|z\right) \cdot P_{prior}(z)}$$

The matrix $P_{c,z}$ describes the probability of each cell to belong to each lobule layer z, given the vector of expression of the 6 landmark genes $\overrightarrow{U^{c}} = \{U^{c}_{1}, \dots, U^{c}_{6}\}$. It consists of N rows representing cells and Z columns representing lobule layers (N=1415 and Z=9, Table S2).

For each cell, the posterior probability is the product of the sampling distribution, namely the probability to have the given landmark gene expression vector at each layer, $P^{c}_{sampling}\left(\overrightarrow{U^{c}}\middle|z\right)$, and the prior probability of the cell to belong to that layer, $P_{prior}(z)$, see Extended Data Fig. 3. The sampling distribution was obtained by measuring the distributions of cellular expression of each of the landmark genes in each layer using smFISH, $P^{c}_{sampling}\left(\overrightarrow{U^{c}}\middle|z\right) = \prod_{g=1}^{6} P\left(U^{c}_{g}\middle|z\right)$, assuming that the expression of the landmark genes is independent (for further details, see section 3).

The sampling distribution encapsulates both the gene-specific uncertainty introduced by the spatial variability in gene expression, as well as the cell-specific sampling uncertainty introduced by the sparse sampling of the scRNAseq method. Lastly, we normalized the posterior matrix by the column sums to obtain the weight matrix:

$$[4]W_{c,z} = \frac{P_{c,z}}{\sum_{c=1}^{N} P_{c,z}}$$

This normalization ensured that the number of cells in each lobule layer did not affect the average layer expression. We next describe our method for obtaining the sampling distribution $P_{sampling}\left(\vec{U}|z\right)$ that was used to obtain the weight matrix $W_{c,z}$.

3. Computing the sampling distribution based on smFISH measurements

Since RNAseq yields a sparse sampling of the cellular mRNA we chose genes with high levels of expression for our landmark gene panel, to ensure their representation in each of the sequenced cells. As a result, individual mRNA molecules could rarely be discerned in the smFISH images. Thus, we quantified expression as average cellular fluorescence intensity, and converted it to estimates of cellular mRNA counts, as explained below.

For each of our landmark genes, segmented cells from the smFISH images were pooled and divided into 9 equidistant layers according to their normalized distance from the central vein. At every layer z we computed the histogram of expression levels, yielding the sampling distribution in units of fluorescence intensity concentrations $P_{sampling}\left(I_g|z\right)$, where I_g is the cellular fluorescence intensity of gene g. The normalized distance was defined as the ratio between the distance of each cell from the central vein and the distance between the portal node and central vein in each quantified lobule. This normalization corrected for lobules that were not sectioned perpendicularly to the lobule vertical axis.

We sought to convert the distributions of cellular expression levels from units of fluorescence intensity to absolute mRNA counts per cell. To this end, we first used our RNAseq data to compute the average fraction of total cellular mRNA attributed to each of our landmark genes $\langle F_g \rangle$. We multiplied this fraction by an estimate of the total number of mRNA molecules in a tetraploid hepatocyte (the most abundant hepatocyte ploidy class in the mouse ages studied[5]), T, to obtain the average number of mRNA molecules of gene g in a tetraploid hepatocyte, $\langle M_g \rangle = \langle F_g \rangle \cdot T$. The cellular fluorescence intensity of each cell was divided by the average fluorescence intensity over all cells and multiplied by $\langle M_g \rangle$:

$$[5] M_g^c = \frac{I_g^c}{\langle I_g \rangle} \cdot \langle M_g \rangle$$

Using equation [5] and the normalized distance of each cell we obtained the sampling distribution in units of absolute mRNA molecules $P_{sampling}(M_g|z)$. We fit these sampling distributions with gamma functions (Table S7).

To estimate T, the average total number of mRNA molecules in a tetraploid hepatocyte, we used previous smFISH-based absolute measurements of the steady state cellular mRNA content for the genes Ass1, G6pc and Pck1 in liver of fasted mice[5]. We divided the average mRNA content of these genes by their corresponding fraction of the total transcriptome, as obtained from bulk RNAseq measurements of liver tissue[2]. This analysis yielded an estimate of 787,000 mRNA molecules per typical tetraploid hepatocyte.

In the RNAseq procedure, we do not detect all mRNA molecules but only a subsample of them. Subsampling the cellular mRNA broadens the distributions of expression levels. A key feature of our algorithm is that cells with lower levels of sampling have broader sampling distributions due to sampling noise, and therefore contribute less to the reconstructed zonation profiles (since their corresponding values in $W_{c,z}$ used in equation [4] will be smaller). To incorporate this feature we sought to estimate a cell-specific sampling distribution, $P\left(U_g^c \middle| z\right)$ defined as the probability of observing U_g molecules of gene g in cell c in each zone z, given that a sparse sampling of a fraction β_c of the cellular molecules has been applied. We first estimated β_c for each cell, as the ratio between the total UMI molecules for that cell and the average number of mRNA molecules per hepatocyte T. The range of sampling levels was $\beta_c = 1.49\% \pm 0.95\%$.

For computational efficiency we discretized the sampling levels into 8 bins representing cells with similar sampling levels and computed a median sampling for that bin, defined as $\beta^d, d = 1 \dots 8$. For each sampling bin we built the sampling distributions as follows – for every landmark gene g in every lobule layer we drew 50,000 values, M_g from the relevant gamma distribution of cellular mRNA expression in that layer $P_{sampling}\left(M_g \middle| z\right)$ (Table S7) and performed a Poisson sampling of this value with parameter $\lambda = \beta^d \cdot \nu \cdot M_g$ to obtain the sampled value m_g . $\nu = 10$ was a factor that corrected for the fact that the smFISH measurements do not capture the entire hepatocyte volume, but rather 0.1 ± 0.04 (median $\pm$ median absolute deviation) of the hepatocyte volumes, and therefore represents by itself a sampling which broadens the true distribution of cellular mRNA[5]. We multiplied the obtained sampled values m_g by $\frac{\beta_c}{\beta^d} \cdot \frac{1}{\nu}$, to ensure that they matched the cellular UMI modeled. This generated the sampling distribution $P\left(U_g^c \middle| z\right)$, used in equation [3] to generate for each cell c the posterior probability $P_{c,z}$ of originating from any of the $z = 1 \dots 9$ layers.

Data visualization

For the tSNE data visualization (Fig. 3a,b, Extended Data Fig. 2) we used an adjusted version of RaceID software[6], on all of the single liver cells acquired, including the non-parenchymal cells filtered out (see Data Preprocessing section). We excluded hepatocytes with less than 1% Albumin out of the total transcript counts. We also excluded cells with less than 600 or more than 50,000 UMI per cell. For each cell the UMI counts of every gene were divided by the summed UMI count of all genes, and multiplied by the median across all cells. Following addition of a pseudocount of 0.1 to the expression data, genes that contained less than a single transcript or more than 1000 transcripts in at least 1100 cells were removed. Additionally, genes that had more than 1000 transcripts in any cell after normalization, were removed as well, resulting in 1724 cells and 753 genes. Dimensionality reduction and visualization were performed with t-distributed stochastic neighbor embedding (t-SNE[7]). Selected gene sets were colored according to the aggregated log-expression of the normalized data.

1. Jaitin, D. A. *et al.* Massively parallel single cell RNA-Seq for marker-free decomposition of tissues into cell types. *Science* **343,** 776–779 (2014).
2. Atger, F. *et al.* Circadian and feeding rhythms differentially affect rhythmic mRNA transcription and translation in mouse liver. *Proc. Natl. Acad. Sci. U. S. A.* **112,** E6579-6588 (2015).
3. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33,** 495–502 (2015).
4. Achim, K. *et al.* High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* **33,** 503–509 (2015).

5. Bahar Halpern, K. *et al.* Bursty Gene Expression in the Intact Mammalian Liver. *Mol. Cell* **58,** 147–156 (2015).

6. Grün, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525,** 251–255 (2015).

7. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn.* **9,** 2579–2605 (2008).