nature

https://doi.org/10.1038/s41586-019-1773-3

Supplementary information

Gene expression cartography

In the format provided by the authors and unedited

Mor Nitzan, Nikos Karaiskos, Nir Friedman & Nikolaus Rajewsky

Supplementary Note

Gene expression cartography

Mor Nitzan[1,2,3#], Nikos Karaiskos[4#], Nir Friedman[3,5*] and Nikolaus Rajewsky[4*]

Supplementary Methods 2
1. The spatial organization principle and its global character 2
2. Integrating the continuity assumption into reference-guided reconstructions 2
3. Choosing the target space shape 3
4. Computing graph-based distances in expression and physical space 3
5. Single cell embedding using optimal transport 3
6. Justification of probabilistic mapping 5
7. Evaluation of spatial reconstruction. 6
8. Generative model for spatial gene expression. 6
9. Genes contained within each nutrient class in the intestinal epithelium 7
10. *De novo* reconstruction is possible up to inherent symmetries of the target space 7
11. Gene ontology analysis for genes extracted as highly zonated in the intestine and liver 7
12. Sample sizes for main text figures 8

Supplementary Discussion 9

Supplementary Tables 11

References 15

[1]John A. Paulson School of Engineering and Applied Sciences, Harvard University, 29 Oxford St, Cambridge, Massachusetts 02138, USA. [2]Broad Institute of MIT and Harvard, 415 Main St, Cambridge, Massachusetts 02142, USA. [3]School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem 9190401, Israel. [4]Systems Biology of Gene Regulatory Elements, Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Hannoversche Str. 28, Berlin 10115, Germany. [5]Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem 9190401, Israel.
[#]These authors contributed equally
[*]Correspondence: nir.friedman@mail.huji.ac.il, rajewsky@mdc-berlin.de

Supplementary Methods

1. The spatial organization principle and its global character

In this manuscript, we explored the assumption that gene expression between nearby cells is generally more similar than gene expression between cells which are separated by larger distances. Biologically, this phenotype can result from multiple mechanisms – gradients of morphogens and nutrients, trajectory of cell maturation, and communication between neighboring cells. While all of these can induce either smooth gradients or sharp boundaries (or combinations thereof) in gene expression patterns, as long as there are spatial shifts between sharp boundaries exhibited by different genes, our hypothesis would hold since *closeness* is a combined property of all genes in the transcriptome. In other words, this is an assumption about overall gene expression across space. Individual genes may very well have sharp expression territories from one cell to a neighboring cell. Our assumption just states that overall, expression of individual genes should only rarely look like salt and pepper patterns but should be organized, for most genes, in (gene specific) spatial territories.

This assumption can be readily tested. Indeed, we showed that at different stages of the developmental process of organisms, or in different tissues in matured organisms, cells that are physically close are also close in expression space, and vice versa (Figs. 2b,f, and 3b). This occurs despite existing sharp boundaries in expression patterns for different genes, since closeness, which is properly defined below, depends on the combined effect of genes composing the full transcriptome.

2. Integrating the continuity assumption into reference-guided reconstructions

Our framework enables the incorporation of both structural and reference guided information. When reconstructing only by a reference atlas (which corresponds to alpha=1 in the manuscript), however, the structural correspondence assumption is already integrated to a certain extent. Although counter-intuitive, this stems from the fact that novoSpaRc exploits a new framework to reconstruct spatial information based on marker genes. There are two major advances implemented into novoSpaRc. First, existing methods (Seurat[1] and DistMap[2]) require binarization of the reference atlas: a gene is considered to be either ON or OFF in a given location of the tissue. In contrast, novoSpaRc works with continuous values and therefore exploits subtle gradients in gene expression that might be present in the data. Second, Seurat and DistMap map individually one cell at a time. By using the framework of optimal transport, novoSpaRc finds the optimal reconstruction by mapping all cells simultaneously, choosing the reconstruction that best satisfies the constraints and is consistent with the marginal distributions (see "Single cell embedding using optimal transport" supplementary section, and "Mathematical formulation of novoSpaRc" Methods section). Taking into account the use of gradients, this process favors reconstructions that best respect the continuity assumption in gene expression. This rather intricate issue will be further discussed in a future manuscript.

3. Choosing the target space shape

A geometrical representation of the tissue will be in general unknown. In these cases, we can flexibly represent the target space as a regular lattice covering the shape of the original tissue (or, in fact, any distribution of finite support with predefined desirable properties). novoSpaRc supports target spaces of any shape and density, but we found that the reconstruction is greatly benefited when an appropriate target shape is selected. Therefore, any prior information or a good educational guess will result in better reconstruction of the investigated tissue.

If the (effective) dimension of the tissue to be reconstructed is unknown, it can be possibly approximated by computing the intrinsic dimensionality of the manifold spanned by single cells lying in the high dimensional expression space (by using for example a maximum likelihood-based approach[3]).

4. Computing graph-based distances in expression and physical space

As the expression profiles are represented in high-dimensional space, metric distances are prone to multiple limitations. Instead, we use steps motivated by non-linear dimensionality reduction methods (e.g., Isomap[4]). However, at this stage we do not require finding low-dimensional coordinates, but rather constructing a robust distance matrix. For symmetry we apply the same procedure to both cells and locations independently (Extended Data Fig. 1a, first column). We start by computing pairwise distances between entities. We chose as a distance metric the Euclidean distance for the physical space (locations) and the correlation-based distance for the expression space (cells), but other measures can be used. These however do not capture the true geometry of nonlinear low- dimensional manifolds. Thus, we use these pairwise distances to construct a k-nearest neighbors graph (Extended Data Fig. 1a, second column). From these graphs, we compute the shortest path lengths for each pair of cells, resulting in graph-based distance matrices for cells and for locations. (Extended Data Fig. 1a, third column).

5. Single cell embedding using optimal transport

As was discussed in the main text, the optimal probabilistic coupling $T^* \in R_+^{N\times M}$ between N single cell expression profiles and M cellular locations can be framed as the solution to the following optimization problem:

$$T^* = \mathrm{argmin}_{T\in C_{p,q}}\ (1-\alpha)D_1(T) + \alpha D_2(T) - \epsilon H(T)$$

where

$$D_1(T) = \sum_{i,j,k,l} L\left(D_{i,k}^{\mathrm{exp}}, D_{j,l}^{\mathrm{phys}}\right) T_{i,j} T_{k,l}\ ,$$

$$D_2(T) = \sum_{i,j} D_{i,j}^{\mathrm{exp,phys}}\ T_{i,j}\ ,$$

ϵ is a non-negative regularization constant, and $\alpha \in [0,1]$ is a constant interpolating between the first two objectives, and can be set to $\alpha = 0$ when no reference atlas is available. The set of coupling between the distribution over expression profiles, $\boldsymbol{p} \in \{p \in R_+^N; \sum_i p_i = 1\}$, and the distribution over locations, $\boldsymbol{q} \in \{q \in R_+^M; \sum_i q_i = 1\}$, is $C_{p,q} \equiv \{T \in R_+^{NxM}; \sum_j T_{i,j} = p_i \ \forall i \in \{1, \dots, N\}, \sum_i T_{i,j} = q_i \ \forall j \in \{1, \dots, M\}\}$.

To retrieve the coupling T^*, we extend upon the results for entropically regularized optimal transport[5] and Gromov-Wasserstein distance-based mapping between metric-measure spaces[6], and use projected gradient descent, where the projection is based on the Kullback-Leibler (KL) metric. Each iteration of the projected exponentiated gradient method consists of two steps; in the first step the current estimate of T is updated by exponentiated gradient descent step, similarly to[6], to yield $\tilde{T}$:

$$\tilde{T} \leftarrow T \odot e^{-\tau \nabla \{\alpha D_1(T) + (1-\alpha) D_2(T) - \epsilon H(T)\}}$$

where $\odot$ is an element-wise multiplication, $e^{(x)}$ is element-wise operation, and $\tau > 0$ is a small step size. In the second step, $\tilde{T}$ is projected back into the set $C_{p,q}$ according to the KL metric:

$$T \leftarrow \mathrm{Proj}_{C_{p,q}}^{KL}\left(\tilde{T}\right),$$

where the KL projection is

$$Proj_{C_{p,q}}^{KL}(K) \equiv \mathrm{argmin}_{T \in C_{p,q}} KL(T||K) \equiv \mathrm{argmin}_{T \in C_{p,q}} \sum_{i,j} T_{i,j} \log \frac{T_{i,j}}{K_{i,j}} \ .$$

It was shown in[7] that the KL projection can be rewritten as an instance of entropically-regularized optimal transport:

$$\mathrm{Proj}_{C_{p,q}}^{KL}(K) \equiv \mathrm{argmin}_{T \in C_{p,q}} < -\epsilon \log K, T > -\epsilon H(T) \ .$$

The gradient of the objective function can be written as

$$\nabla \{\alpha D_1(T) + (1-\alpha) D_2(T) - \epsilon H(T)\} = \alpha L\left(D^{\exp,}, D^{\mathrm{phys}}\right) \otimes T + (1-\alpha) D^{\exp,\mathrm{phys}} + \epsilon \log T \ ,$$

where $\log(x)$ is an element-wise operation, and the tensor product is defined as

$$L \otimes T \equiv \left(\sum_{k,l} L_{i,j,k,l} T_{k,l}\right)_{i,j} \ .$$

Altogether, we have:

$$\mathrm{Proj}_{C_{p,q}}^{KL}\left(T \odot e^{-\tau \nabla \{\alpha D_1(T) + (1-\alpha) D_2(T) - \epsilon H(T)\}}\right) = \mathrm{argmin}_{T \in C_{p,q}}$$
$$< -\epsilon \log T + \epsilon\tau \left\{\alpha L\left(D^{\exp,}, D^{\mathrm{phys}}\right) \otimes T + (1-\alpha) D^{\exp,\mathrm{phys}} + \epsilon \log T\right\}, T > -\epsilon H(T) \ .$$

Therefore, if we set $\tau = 1/\epsilon$, each iteration of the algorithm can be simplified to a Sinkhorn projection,

$$T \leftarrow \mathrm{argmin}_{T \in C_{p,q}} < \alpha L\left(D^{\exp,}, D^{\mathrm{phys}}\right) \otimes T + (1-\alpha) D^{\exp,\mathrm{phys}}, T > -\epsilon H(T) \ .$$

Each of these iteration steps can be computed using Sinkhorn's fixed point algorithm[5]. Specifically,

$$T \leftarrow \mathrm{diag}(a) K \mathrm{diag}(b) \ ,$$

where the Gibbs kernel associated with $\{\alpha L\left(D^{\exp,}, D^{\mathrm{phys}}\right) \otimes T + (1-\alpha) D^{\exp,\mathrm{phys}}\}$

is $K \equiv \mathrm{e}^{-\frac{1}{\epsilon}\{\alpha L\left(D^{\exp,}, D^{\mathrm{phys}}\right) \otimes T + (1-\alpha) D^{\exp,\mathrm{phys}}\}} \in R_+^{NxM}$.

Finally, $a \in R_+^N$ and $b \in R_+^M$ can be computed using Sinkhorn's fixed point iteration[8] involving element-wise division:

$$a \leftarrow \frac{p}{Kb} \ , \ b \leftarrow \frac{q}{K^T a} \ .$$

We provide a Python package for the implementation of novoSpaRc at https://github.com/rajewsky-lab/novosparc. Parts of the code are based on modifications of the Python Optimal Transport Package (https://pot.readthedocs.io/).

6. Justification of probabilistic mapping

We posed the spatial mapping problem as finding a probabilistic embedding between the cells and locations. That is, each single cell is to be assigned a distribution over cellular locations. A probabilistic mapping is preferable for several reasons.

Single cell data does not yield an exact 1-to-1 matching problem. (i) When a tissue is dissociated into single cells, we would generally not be able to retrieve information for the full batch of single cells, but only for a certain fraction of them, due to experimental constraints. (ii) There would generally not be information about the number of original single cells in the tissue and their exact location, meaning we would need to resort to assignment of single cells over a grid. (iii) Even in cases where there are known, reproducible cellular locations, and there is the possibility to dissociate many nearly-identical tissues to increase the single cell coverage, we would still expect to have cellular locations that correspond to multiple single cells, and cellular locations that do not correspond to any of the single cells in the dataset.

Probabilistic mapping would yield smoother expression patterns and would be more robust to the noisy, partially sampled single cell data. Given imperfect data, as is the case for experimental setups, we may be uncertain about the exact location of a dissociated single cell and would rather place it in a certain neighborhood of the tissue (or, probabilistically spread it over several locations in that area). This is motivated both by noise and dropouts in the original data, and the fact that if we are mapping single cells to a grid, their true original location may be in between several nodes (cellular locations) on the grid, in which case their true mapping should be distributed over the grid nodes surrounding the original location, weighted by their corresponding distance from that location.

Probabilistic matching is more efficient computationally. Intuitively, we replace a discrete optimization problem over a large combinatorial space with a continuous optimization of a smooth function, which allows us to employ more efficient optimization methods. Details can be found in the Supplementary section 'Single cell embedding using optimal transport'.

Finally, we are interested in the reconstructed *expression patterns over stereotypical tissues*, and not necessarily in assigning single cells their exact original location.

7. Evaluation of spatial reconstruction.

We evaluate the quality of reconstruction by novoSpaRc by three different measures: (a) *Correlation of expression patterns*. The reconstructed spatial gene expression of all genes (vISH) can be compared to the original expression patterns by computing the Pearson correlation between them, averaged over all genes, such as in Fig. 3c. (b) *Alignment of single cell assignment*. For the tissues with 1d symmetry we also compute the fraction of cells correctly assigned to their original spatial zone. To do this, we compare for each cell its original spatial zone to its reconstructed zone according to novoSpaRc. More specifically, the zone that the cell is assigned to with highest probability. This notion can be extended to the fraction of cells assigned to a spatial zone that is found at most at a certain distance from their original zone. We show this evaluation for increasing distances for the reconstruction of the intestinal epithelium and the liver (Extended Data Figs. 2a,b). (c) *Probability heatmap*. In Fig. 2c,g we quantify the assignment of single cells to their corresponding 1d spatial zones by a probabilistic version of a confusion matrix (the probability heatmap). For each original zone (on the x-axis), we average over the reconstructed spatial probability distribution of single cells originating from that zone and display that on the y-axis.

8. Generative model for spatial gene expression.

To systematically evaluate novoSpaRc's performance, we generated synthetic spatial expression data using a simple generative model that is based on independent Gaussian spatial expression patterns for each gene, for either a 1d (line), 2d (squre) or 3d (cube) shaped synthetic tissue.

For 1d tissues, the expression E of each gene g over the spatial zones is proportional to a gaussian distribution, $E(x|\mu_g, \sigma_g) \propto e^{-\frac{(x-\mu_g)^2}{2\sigma g^2}}$, where μ_g is the mean of the gaussian, sampled uniformly across the 1d grid, and σ_g is the standard deviation. For 2d and 3d tissues, the expression is proportional to a multivariate normal distribution, $E(\boldsymbol{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \propto e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_g)^T {\boldsymbol{\Sigma}_g}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_g)}$, where $\boldsymbol{\mu}_g$ is the mean vector (sampled uniformly across the 2d or 3d grid), and $\boldsymbol{\Sigma}_g$ is the covariance matrix.

After generating the synthetic expression matrix, we add gaussian noise to the expression values with 0 mean and $\sigma_{noise}\sigma_{expression}$ standard deviation, where $\sigma_{expression}$ is the standard deviation of the entire expression matrix, and σ_{noise} is a parameter that sets the signal to noise ratio.

The expression of 'spatially informative' genes is set according to the model above, while the expression of 'spatially non-informative' genes is randomly permuted across the synthetic tissue.

The default parameters for the simulations and novoSpaRc reconstructions are: 1000 single cells (or closest approximation for the 2d grid), 100 grid locations (or closest approximation for the 3d grid), 100 genes, $\sigma = 10$, $\sum_g \sigma I$ (where I is the identity matrix), $\alpha = 0.5$, number of marker genes = 5, and $\sigma_{expression} = 0.1$.

9. Genes contained within each nutrient class in the intestinal epithelium

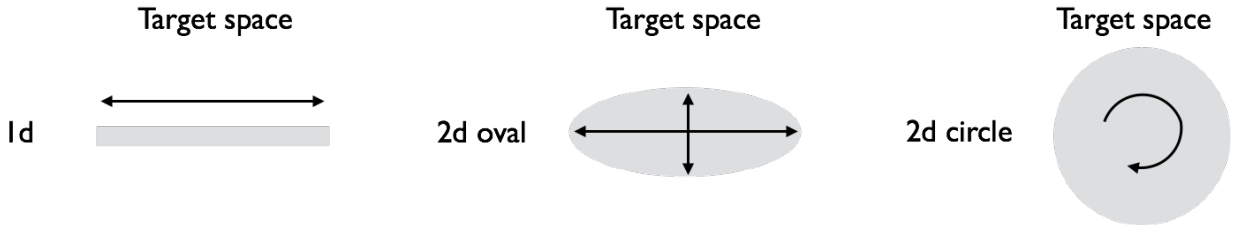
Table corresponding to Fig. 2d.

Nutrient class	**Associated genes**
Apolipoproteins Cholesterol	Apobec1, Apob, Apoa4, Apoa1, Npc1l1
Carbohydrates	Slc5a1, Slc2a5, Slc2a2
Peptides	Slc15a1
Amino acids	Slc7a9, Slc7a8, Slc7a7

10. *De novo* reconstruction is possible up to inherent symmetries of the target space

When reconstructing a tissue *de novo*, meaning without any reference atlas, the reconstruction can be established up to the symmetry axes of the target space. This is not specific to novoSpaRc, but an inherent property of the problem. For example, when reconstructing a 1-dimensional tissue, then without additional prior knowledge there is no way of distinguishing between a 'right-to-left' and 'left-to-right' reconstruction (see illustration in the Figure below, left). This can be remedied by including prior information that would break the symmetry, such as marker gene expression information (as long as they do not exhibit the same underlying symmetries). Such prior information would correctly "anchor" the reconstruction and might also appear more intuitive to the user.

Below are a few examples for such symmetries (specified using a double-arrowed line on each target space) that fundamentally cannot be resolved *de novo*:



11. Gene ontology analysis for genes extracted as highly zonated in the intestine and liver

Based on novoSpaRc's *de novo* reconstruction of the mammalian intestine [9] and liver [10] single cell datasets, we extracted a list of genes whose expression is localized to different layers of the two tissues. We used gene ontology (GO) enrichment analysis [11] to show that these groups of genes are enriched for distinct biological processes, many of which are consistent with the respective expression localization.

We filtered for genes whose maximum value in the sDGE is among the top 20% of all genes. We then chose the genes whose maximum spatial expression is localized at one of the two ends of the 1d tissue and whose associated Kendall Tau p-value < 0.05, to be set as the groups of genes highly expressed in either the 'top' or the 'bottom' of the tissue. We chose the genes whose maximum spatial expression is

localized at either one of the middle layers of the tissue (one layer apart from the two border layers of the tissue at either side), to be set as the 'intermediate' gene group. We used the Gorilla package [11] to run GO enrichment analysis, by performing pairwise comparisons between two unranked lists of genes at a time: a target set and a background set (composed of the complement of the target set).

In the intestine, the reconstructed crypt was enriched for transcription, translation, RNA splicing, and cell cycle related genes. The middle part of the reconstructed intestinal crypt-to-villus axis was enriched for amino acid and carbohydrates transport and processing, ion transport and intestinal absorption related genes. The reconstructed tip of the villus (V6) was enriched for lipoprotein metabolic processes, catabolic processes, extracellular matrix organization, cholesterol transport and processing, and stress related processes. In the liver, the reconstructed pericentral layer was enriched for xenobiotic metabolism, fatty acid metabolism and catabolic processes, while the reconstructed periportal layer was enriched for carboxylic processes, oxidation-reduction processes and ATP biosynthesis. The full lists of enriched genes and their associated p-values and FDR q-values, for both the intestine and the liver, can be found as Supplementary Files.

12. Sample sizes for main text figures

Respective sample sizes for data along x-axis:

Figure 2b: [154730, 227485, 167832, 137948, 149135, 84683, 33840].
Figure 2f: [34256, 43408, 52544, 30083, 28042, 14723, 7416, 1678, 76].
Figure 3b: [83541, 723081, 1001329, 1184469, 741496, 528972, 240138, 107831, 8423].
Extended Data Figure 9b: [49873, 92934, 80571, 69714, 57566, 46117].
Extended Data Figure 9f: [49184, 90843, 79501, 68159, 56817, 45475].

Supplementary Discussion

novoSpaRc's advantages, limitations, and direct comparison to existing reconstruction methods.

novoSpaRc offers several features which cannot be exploited as a whole by existing methods: (a) it enables incorporation and interpolation of both structural information (such as the structural correspondence assumption) and a reference atlas, (b) it naturally provides probabilistic embedding of single cells onto their original spatial context, which provides a more robust reconstruction, (c) it allows to incorporate prior structural information regarding the structure of the tissue from which the cells were dissociated, (d) it does not require any tailored pre-processing steps and can utilize continuous expression data directly, (e) and finally, it is flexible in terms of its structural assumption (which can be potentially adjusted in future work) and allows to incorporate marginal information (effectively incorporating prior knowledge about different aspects such as varying local density of cells across the tissue and varying quality of sequenced single cells).

We directly compare novoSpaRc to two available spatial reconstruction methods that fully rely on a reference atlas: Seurat[1] and DistMap[2]. A comparison of the intrinsic characteristics of the three approaches, as well as their corresponding reconstruction results for the BDTNP data[12], scRNA-seq data of the *Drosophila*[2] and zebrafish embryos[1] and the cerebellum[13] are shown in Extended Data Fig. 10. This comparative analysis is performed for varying numbers of marker genes and shows how, for the same number of marker genes, novoSpaRc generally outperforms other available methods. Both DistMap and Seurat require a large number of marker genes to reconstruct the BDTNP dataset, whereas the Pearson correlations for novoSpaRc saturate at perfect reconstruction with only 2 marker genes. novoSpaRc outperforms Seurat and DistMap in the case of the *Drosophila* embryo and performs comparably to them for the zebrafish embryo, while it should be stressed that DistMap and Seurat were developed and tailored for these two datasets, respectively. Finally, novoSpaRc substantially outperforms DistMap and Seurat for the reconstruction of the brain cerebellum, where both DistMap and Seurat struggle to form meaningful reconstructions. It should be noted that DistMap requires a threshold to produce the expression patterns, which is in principle unknown. We selected the threshold which maximizes the Pearson correlations, thus giving DistMap an unfair advantage in these comparisons.

It is important, however, to keep in mind novoSpaRc's limitations. novoSpaRc works by embedding the single cells into a predefined shape, and so does not allow to learn a latent representation of the data that was not used as input. In addition, as mentioned in the main text, *do novo* reconstruction can be achieved up to global transformations relative to symmetries of the shape of the target space. This is not a limitation specific to novoSpaRc but inherent to the problem of *de novo* reconstruction without additional prior information, such as marker gene data (see "*De novo* reconstruction is possible up to inherent symmetries of the target space" supplementary section). Finally, novoSpaRc employs an assumption about spatial gene expression (here we use the structural correspondence assumption) to reconstruct cellular locations. In general, we found the structural correspondence assumption to hold to a certain extent in all tissues and organisms we looked into so far, including highly heterogeneous and challenging

tissues like the brain. We believe this hints that spatial gene expression is much more structured and informative than currently believed, and that external signaling gradients and cell-to-cell communication provide stronger signals for spatial patterning than expected. In cases where this is a weak assumption, challenged for example by complex tissues with multiple cell types or multiple domains, novoSpaRc may struggle. However, it is important to stress that novoSpaRc's flexibility allows it to employ alternative principles or assumptions that would fit different biological scenarios or incorporate diverse experimental prior information.

Supplementary Tables

Intestine: predicted by NovoSpaRc to be zonated towards the crypt

	Hist1h2ap	2810417H13Rik	Top2a	Kcne3	Cenpa	Ccdc34	Hmgb2	Impdh2	Ptma	Cdca3
Reconstructed as zonated towards the crypt [9]	X	X	X	X	X	X	X	X	X	X
Reported to be expressed in the crypt				[14]						
Reported to be overexpressed in the crypt vs the villus (in human)			[15, 16]							
Reported to be functionally associated with crypt					[b]	[c]			[d]	[e]
Additional support	[a]						[a]			

[a] Was found to be expressed similarly to Top2a in single cells [17].
[b] Associated with cell division
[c] Reported to regulate cell proliferation, apoptosis and migration in bladder [18].
[d] Inferred to be involved in regenerative process, proliferation, or stem cell identity [19].
[e] Gene ontology process: cell cycle and cell division [20].

Intestine: predicted by NovoSpaRc to be zonated towards the tip of the villus

	Clca4a	Tubb2a	Pmp22	Apol9b	Tm4sf4	Enpp3	Apol9a	Isg15	Abhd2	Apoa4
Reconstructed as zonated towards V6 [9]	X	X	X	X	X	X	X	X	X	X
Protein abundance was associated with V6 [9]	X			X		X	X	X	X	X
Reported to be overexpressed in the villus vs the crypt (in human)										[15, 16]

Supplementary Table 1 | Literature-based support for highly zonated genes in the intestinal epithelium revealed by novoSpaRc. All 20 genes recovered by novoSpaRc to rank highest among

zonated genes (10 top zonated genes towards the crypt, and 10 top zonated genes towards V6), were either independently reconstructed (based on a reference atlas) to be zonated, and/or have direct experimental support for their zonation profiles, and/or were shown to be functionally related to processes associated with their respective zonation profiles. Selection of top zonated genes is described in Methods.

Liver: predicted by NovoSpaRc to be pericentral (zonated towards layer 1)

	Oat	Cyp2a5	Glul	Lhpp	Fitm1	Cyp2c37	Rdh1	Cyp2e1	Cyp2c29	Lect2
Reconstructed as zonated towards CV [10]	X	X	X	X	X	X	X	X	X	X
Reported as zonated towards CV	[21-24]		[25, 26] [24]					[27, 28] [24, 26]		
Differentially methylated towards CV [24]	X		X	X				X		
Higher expression in Axin2$^+$ pericentral hepatocytes [29]		X	X	X		X				X
Additional support							[a]			

[a] A gene found to increase in liver of mice exposed to chronic hypoxia [30].

Liver: predicted by NovoSpaRc to be periportal (zonated towards layer 9)

	Serpina12	Sds	Tstd1	Ly6e	Mfsd2a	Pigr	Prdx4	Gm5506	Sdc1	Itih3
Reconstructed as zonated towards PV [10]	X	X	X	X	X	X	X	X	X	X
Reported as zonated towards PV		[24]								
Differentially methylated [24]	X	X		X		X			X	X
Lower expression in Axin2$^+$ pericentral hepatocytes [29]	X				X					
Additional support		[c]			[a]		[b]			[c]

[a] A gene found to increase in liver of mice exposed to chronic hypoxia [30].
[b] secretory antioxidase that protects against oxidative damage, whose overexpression reduced local and systemic oxidative stress generated by BDL [31].
[c] Reported as differentially expressed genes between PV and CV zone that were associated with differentially methylated regions featuring hypomethylation coinciding with a transcriptional upregulation in the respective zone [24].

Supplementary Table 2 | Literature-based support for highly zonated genes in the liver lobule revealed by novoSpaRc. All 20 genes recovered by novoSpaRc to rank highest among zonated genes (10 top zonated genes towards the CV, and 10 top zonated genes towards PV), were either independently reconstructed (based on a reference atlas) to be zonated, and/or have direct experimental support for their

zonation profiles, and/or were shown to be functionally related to processes associated with their respective zonation profiles. Selection of top zonated genes is described in Methods.

References

1. Satija, R., et al., *Spatial reconstruction of single-cell gene expression data.* Nat Biotechnol, 2015. **33**(5): p. 495-502.
2. Karaiskos, N., et al., *The Drosophila embryo at single-cell transcriptome resolution.* Science, 2017. **358**(6360): p. 194-199.
3. Levina, E. and P.J. Bickel. *Maximum likelihood estimation of intrinsic dimension*. in *Advances in neural information processing systems*. 2005.
4. Tenenbaum, J.B., V. de Silva, and J.C. Langford, *A global geometric framework for nonlinear dimensionality reduction.* Science, 2000. **290**(5500): p. 2319-23.
5. Cuturi, M. *Sinkhorn distances: Lightspeed computation of optimal transport*. in *Advances in neural information processing systems*. 2013.
6. Peyré, G., M. Cuturi, and J. Solomon. *Gromov-Wasserstein averaging of kernel and distance matrices*. in *International Conference on Machine Learning*. 2016.
7. Benamou, J.-D., et al., *Iterative Bregman projections for regularized transportation problems.* SIAM Journal on Scientific Computing, 2015. **37**(2): p. A1111-A1138.
8. Sinkhorn, R., *Diagonal equivalence to matrices with prescribed row and column sums.* The American Mathematical Monthly, 1967. **74**(4): p. 402-405.
9. Moor, A.E., et al., *Spatial Reconstruction of Single Enterocytes Uncovers Broad Zonation along the Intestinal Villus Axis.* Cell, 2018. **175**(4): p. 1156-1167 e15.
10. Halpern, K.B., et al., *Single-cell spatial reconstruction reveals global division of labour in the mammalian liver.* Nature, 2017. **542**(7641): p. 352-356.
11. Eden, E., et al., *GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists.* BMC Bioinformatics, 2009. **10**: p. 48.
12. *Berkeley drosophila transcription network project*. Available from: http://bdtnp.lbl.gov/Fly-Net/bioimaging.jsp.
13. Rodriques, S.G., et al., *Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution.* Science, 2019. **363**(6434): p. 1463-1467.
14. Preston, P., et al., *Disruption of the K+ channel beta-subunit KCNE3 reveals an important role in intestinal and tracheal Cl- transport.* J Biol Chem, 2010. **285**(10): p. 7165-75.
15. Gassler, N., et al., *Molecular characterisation of non-absorptive and absorptive enterocytes in human small intestine.* Gut, 2006. **55**(8): p. 1084-9.
16. Olsen, L., et al., *CVD: the intestinal crypt/villus in situ hybridization database.* Bioinformatics, 2004. **20**(8): p. 1327-8.
17. Grootjans, J., et al., *Epithelial endoplasmic reticulum stress orchestrates a protective IgA response.* Science, 2019. **363**(6430): p. 993-998.
18. Gong, Y., et al., *CCDC34 is up-regulated in bladder cancer and regulates bladder cancer cell proliferation, apoptosis and migration.* Oncotarget, 2015. **6**(28): p. 25856-67.
19. Tetteh, P.W., et al., *Replacement of lost Lgr5-positive stem cells through plasticity of their enterocyte-lineage daughters.* Cell stem cell, 2016. **18**(2): p. 203-213.
20. Eppig, J.T., et al., *Mouse Genome Informatics (MGI): Resources for Mining Mouse Genetic, Genomic, and Biological Data in Support of Primary and Translational Research.* Methods Mol Biol, 2017. **1488**: p. 47-73.
21. Stanulovic, V.S., et al., *Hepatic HNF4alpha deficiency induces periportal expression of glutamine synthetase and other pericentral enzymes.* Hepatology, 2007. **45**(2): p. 433-44.

22. Kuo, F.C., et al., *Colocalization in pericentral hepatocytes in adult mice and similarity in developmental expression pattern of ornithine aminotransferase and glutamine synthetase mRNA.* Proc Natl Acad Sci U S A, 1991. **88**(21): p. 9468-72.
23. Bennett, A.L., et al., *Acquisition of antigens characteristic of adult pericentral hepatocytes by differentiating fetal hepatoblasts in vitro.* J Cell Biol, 1987. **105**(3): p. 1073-85.
24. Brosch, M., et al., *Epigenomic map of human liver reveals principles of zonated morphogenic and metabolic control.* Nat Commun, 2018. **9**(1): p. 4150.
25. Preziosi, M., et al., *Endothelial Wnts regulate beta-catenin signaling in murine liver zonation and regeneration: A sequel to the Wnt-Wnt situation.* Hepatol Commun, 2018. **2**(7): p. 845-860.
26. Braeuning, A., et al., *Differential gene expression in periportal and perivenous mouse hepatocytes.* FEBS J, 2006. **273**(22): p. 5051-61.
27. Hailfinger, S., et al., *Zonal gene expression in murine liver: lessons from tumors.* Hepatology, 2006. **43**(3): p. 407-14.
28. Gebhardt, R., *Metabolic zonation of the liver: regulation and implications for liver function.* Pharmacol Ther, 1992. **53**(3): p. 275-354.
29. Wang, B., et al., *Self-renewing diploid Axin2(+) cells fuel homeostatic renewal of the liver.* Nature, 2015. **524**(7564): p. 180-5.
30. Baze, M.M., K. Schlauch, and J.P. Hayes, *Gene expression of the liver in response to chronic hypoxia.* Physiol Genomics, 2010. **41**(3): p. 275-88.
31. Zhang, J., et al., *Protective Effects of Peroxiredoxin 4 (PRDX4) on Cholestatic Liver Injury.* Int J Mol Sci, 2018. **19**(9).