

Gene expression cartography

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

Mor Nitzan^{1,2,3,6}, Nikos Karaiskos^{4,6}, Nir Friedman^{3,5*} & Nikolaus Rajewsky^{4*}

Received: 1 February 2019

Accepted: 7 October 2019

Published online: 20 November 2019

Multiplexed RNA sequencing in individual cells is transforming basic and clinical life sciences^{1–4}. Often, however, tissues must first be dissociated, and crucial information about spatial relationships and communication between cells is thus lost. Existing approaches to reconstruct tissues assign spatial positions to each cell, independently of other cells, by using spatial patterns of expression of marker genes^{5,6}—which often do not exist. Here we reconstruct spatial positions with little or no prior knowledge, by searching for spatial arrangements of sequenced cells in which nearby cells have transcriptional profiles that are often (but not always) more similar than cells that are farther apart. We formulate this task as a generalized optimal-transport problem for probabilistic embedding and derive an efficient iterative algorithm to solve it. We reconstruct the spatial expression of genes in mammalian liver and intestinal epithelium, fly and zebrafish embryos, sections from the mammalian cerebellum and whole kidney, and use the reconstructed tissues to identify genes that are spatially informative. Thus, we identify an organization principle for the spatial expression of genes in animal tissues, which can be exploited to infer meaningful probabilities of spatial position for individual cells. Our framework ('*novoSpaRc*') can incorporate prior spatial information and is compatible with any single-cell technology. Additional principles that underlie the cartography of gene expression can be tested using our approach.

Single-cell RNA sequencing (scRNA-seq) has revolutionized our understanding of the rich heterogeneous cellular populations that make up tissues, the dynamics of developmental processes and the underlying regulatory mechanisms that control cellular function^{1–4}. However, to understand how single cells orchestrate multicellular functions, it is crucial to have access not only to the identities of single cells but also to their spatial context. This is a challenging task, as tissues must commonly be dissociated into single cells before scRNA-seq can be performed, and thus the original spatial context and relationships between cells are lost. Two seminal papers tackled this problem computationally^{5,6}—the key idea being to use a reference atlas of informative marker genes as a guide to assign spatial coordinates to sequenced cells. This concept was successfully used in various tissues^{7–11}, including the early *Drosophila* embryo¹². However, such methodologies rely heavily on the existence of an extensive reference database for spatial expression patterns, which may not always be available or straightforward to construct. Moreover, in practice the number of available reference marker genes is usually not large enough to label each spatial position with a distinct combination of reference genes, making it impossible to uniquely resolve cellular positions. More generally, marker genes, even when available, convey limited information, which could possibly be enriched by the structure of single-cell data.

To this aim, we developed a new computational framework (*novoSpaRc*), which allows for de novo spatial reconstruction of single-cell gene expression, with no inherent reliance on any prior information, and the flexibility to introduce it when it does exist (Fig. 1). Similar to solving a puzzle, we seek the optimal configuration of pieces (cells)

that recreates the original image (tissue). However, contrary to a typical puzzle, here we do not have access to the image that we aim to reconstruct. Although the number of ways to spatially arrange (or 'map') sequenced cells in tissue space is enormous, our hypothesis is that gene expression in the vast majority of these arrangements will not be as organized as in the real tissue. For example, we know that typically there exist genes that are specifically expressed in spatially contiguous territories and are thus consistent with only a small subset of all possible arrangements. We therefore set out to identify simple, testable assumptions that govern how gene expression is organized in space, and to subsequently find the arrangements of cells that best respect those assumptions.

novoSpaRc charts gene expression in tissues

Here, we specifically explore the assumption that cells that are physically close tend to share similar transcription profiles, and vice versa (Extended Data Fig. 1a, Supplementary Methods). Biologically, this phenotype can result from multiple mechanisms, such as gradients of oxygen, morphogens and nutrients, the trajectory of cell development and communication between neighbouring cells. We stress that this is an assumption about overall gene expression across the entire tissue—not about individual genes and not about all cells that are physically close (Supplementary Methods). We show that, on average, the distance between cells in expression space increases with their physical distance, for diverse tissues in mature organisms or whole embryos in early development. Thus, to predict the spatial locations of sequenced cells, we

¹John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. ²Broad Institute of MIT and Harvard, Cambridge, MA, USA. ³School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel. ⁴Systems Biology of Gene Regulatory Elements, Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany. ⁵Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel. ⁶These authors contributed equally: Mor Nitzan, Nikos Karaiskos. *e-mail: nir.friedman@mail.huji.ac.il; rajewsky@mdc-berlin.de

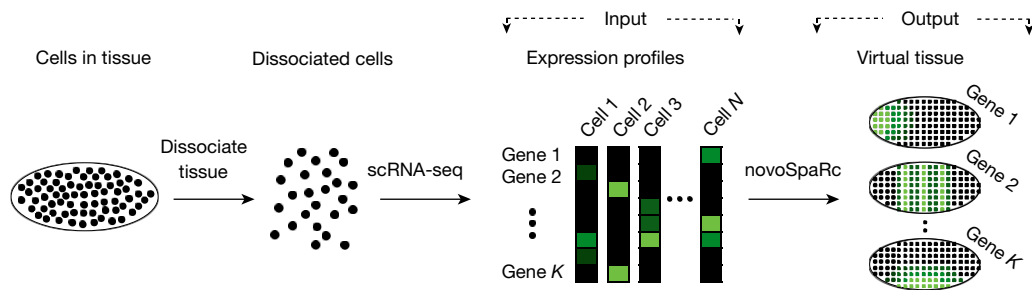


Fig. 1 | Overview of novoSpaRc. A matrix that contains single-cell transcriptome profiles, sequenced from dissociated cells, is the main input for novoSpaRc. The output is a virtual tissue of a chosen shape, which can be queried for the expression of all genes quantified in the data.

seek to find a map of sequenced cells to tissue space ('cartography') such that overall structural correspondence is preserved—meaning that, overall, cells have similar relative distances to other cells in expression and physical space. The physical space is anchored by locations that may be either known (such as the reproducible cellular locations in the early stages of development of the *Drosophila* embryo¹³) or approximated by a grid (Supplementary Methods). The distances are first computed for each pair of cells across graphs constructed over the two spaces, to account for the underlying structure of the data (Supplementary Methods). Then, novoSpaRc optimally aligns the distances of pairs of cells between the expression data and geometric features of the physical space, in a way that is consistent with spatial expression profiles of marker genes when these are available (Methods, Supplementary Methods). For reasons that are both biologically and computationally motivated, we seek a probabilistic mapping that assigns each cell a distribution over locations on the physical space (Supplementary Methods). We formulate this as a generalized optimal-transport problem^{14–16}, which has been proven to be increasingly valuable for diverse fields (including biology^{17,18}) and renders the task of reconstruction feasible for large datasets. Specifically, we formulate an interpolation between entropically regularized Gromov–Wasserstein^{19,20} and optimal-transport²¹ objectives, which serves to satisfy the assumption of structural correspondence between gene expression space and physical space, and to match prior knowledge when available (Methods). We show that this optimization problem can be efficiently solved using projected gradient descent reduced to iterations of linear optimal-transport sub-problems (Supplementary Methods). To systematically assess the performance of novoSpaRc, we used a simple generative model of spatial gene expression to show that it can robustly recover it (Supplementary Methods, Extended Data Fig. 1b–d).

novoSpaRc reconstructs tissues de novo

Focusing on real single-cell datasets, we first reconstructed tissues de novo that have inherent symmetries that render them effectively one-dimensional, such as the mammalian intestinal epithelium¹⁰ and liver lobules⁷. Schematic figures of the reconstruction process are shown in Fig. 2a, e. Cells were previously classified into seven distinct zones for the intestine, or nine layers for the liver, on the basis of robust marker gene information^{7,10}. We found that the average pairwise distances between cells in expression space increased monotonically with the pairwise distances in physical one-dimensional space (Fig. 2b, f), consistent with our structural correspondence assumption.

We used novoSpaRc to embed the expression data into one dimension. The embedded coordinates of single cells correlated well on average with their layer or zone memberships (Fig. 2c, g, Supplementary Methods). The median Pearson correlation coefficient for reconstructed expression patterns to original patterns for the top 100 variable genes was 0.99 for intestine and 0.94 for liver (Supplementary Methods), and the fraction of cells that were correctly assigned up to one layer away from their original layer was 0.98 for intestine and 0.73

for liver (Supplementary Methods, Extended Data Fig. 2a, b). novoSpaRc captured spatial expression patterns of the top zoned genes and spatial division of labour within the intestinal epithelium—as well as within the layers of the liver lobules (Methods, Fig. 2d, h, Extended Data Fig. 3a, b), in which cells in different tissue layers perform different tasks and exhibit different expression profiles. For the intestine, varying the grid resolution to include either fewer or more embedded zones did not compromise the quality of the reconstructed expression patterns (Extended Data Fig. 3c), which shows the potential for increased resolution of single-cell-based relative to atlas-based embedding.

novoSpaRc reconstructs early embryos

Next, we focused on spatially reconstructing the well-studied *Drosophila* embryo, as a more-challenging, higher-dimensional tissue. Late in stage 5 of development, the fly embryo consists of around 6,000 cells. It has been previously suggested²² that at early stages of fly development, the expression levels of gap genes can be optimally decoded into positional information. The expression levels of 84 transcription factors were quantitatively registered using fluorescence in situ hybridization (FISH) for each of the cells by the Berkeley *Drosophila* Transcription Network Project (BDTNP)¹³.

To assess the performance of novoSpaRc, we first simulated scRNA-seq data by in-silico dissociating the BDTNP dataset into single cells (Methods), and then attempted to reconstruct the original expression patterns across the tissue both de novo and by using marker genes (Fig. 3a). Similarly to the 'one-dimensional' datasets, we found a monotonically increasing relationship between the cell–cell pairwise distances in expression space and in physical space (Fig. 3b), confirming that the data adheres to our structural correspondence assumption.

The reconstructed patterns of spatial gene expression highly correlated with the original ones (Fig. 3c). We found that the novoSpaRc reconstruction that incorporated both structural and marker gene information outperformed the reconstruction based on only the latter, and that performance was saturated at two marker genes (Fig. 3c), independently of the marker genes used. As expected, the quality of the reconstruction increased with the number of genes used to provide structural information in expression space, and with the fraction of spatially informative genes (Supplementary Methods, Extended Data Fig. 4a, b). The majority of spatial patterns were recapitulated faithfully even when only a single marker gene was used (Fig. 3c, d). In addition, novoSpaRc identified the physical neighbourhoods from which cells originated when used de novo (up to inherent symmetries; see Supplementary Methods), and pinpointed their true locations ($P < 0.05$ compared to random assignment) when a handful of marker genes were used (Fig. 3e, Extended Data Fig. 5a, b).

We examined the expression patterns of four transcription factors that span the dorsal–ventral and anterior–posterior axes (Fig. 3d). The quality of the reconstruction improved when applying the structural correspondence assumption (Supplementary Methods, Extended Data Fig. 5d). The de novo reconstruction correctly identified both axes of the embryo, and

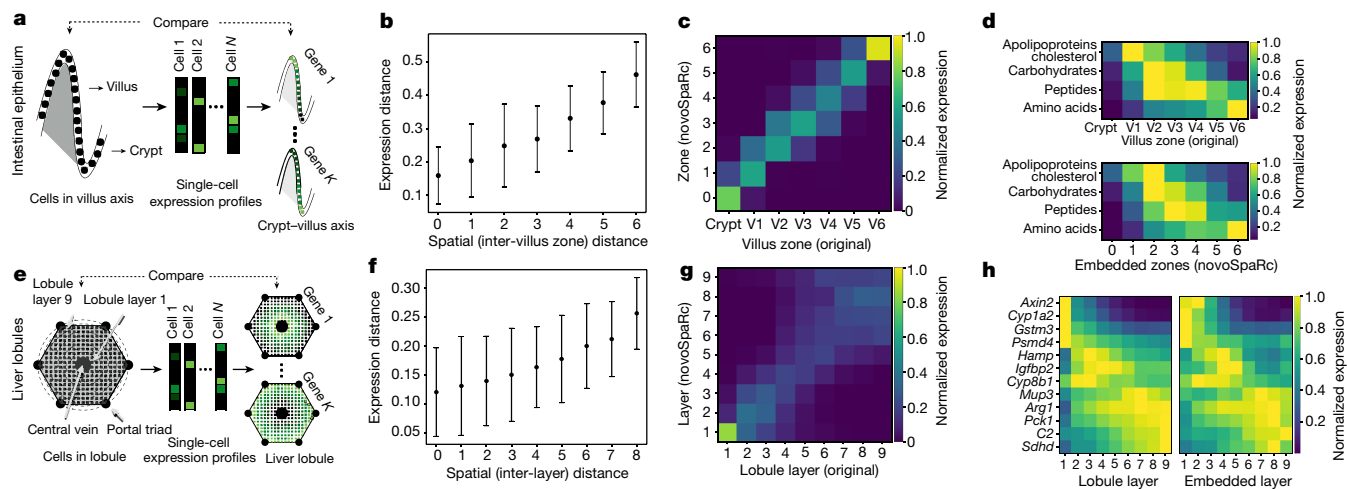


Fig. 2 | novoSpaRc successfully reconstructs complex tissues with effective one-dimensional structure de novo. **a, e.** The reconstruction scheme for the mammalian intestinal epithelium (**a**) and liver lobules (**e**). **b, f.** Demonstration of the monotonic relationship between cellular pairwise distances in expression and physical space for intestinal epithelium (**b**) and liver lobules (**f**). Distances are measured as weighted shortest paths along the graphs constructed over physical or expression spaces. Data are mean \pm s.d. **c, g.** novoSpaRc infers the original spatial context of single cells of the intestinal

epithelium (**c**) and liver lobules (**g**) with high accuracy. Heat maps show the inferred distribution over embedded layers (rows) for the cells in each of the original layers (columns). **d, h.** novoSpaRc captures the spatial division of labour of averaged expression of genes that have a role in the absorption of different classes of nutrients in the intestine (**d**) and the spatial expression patterns of a group of pericentral, periportal and non-monotonic genes in the liver lobule (**h**). The expression level of each gene in **d** and **h** is normalized to its maximum value.

the reconstructed portrait was remarkably similar to the original one (Fig. 3d). In general—because de novo reconstruction is performed without any prior information that would anchor the cells—the reconstructed configuration is similar up to global transformations (reflections, rotations and translations), relative to the respective axes of symmetry (Supplementary Methods). Consequently, the resulting patterns of gene expression might be shifted or flipped relative to the expected ones. However, there are features of a faithful reconstruction that we can test for, such that the reconstruction would be robust to small changes in the optimization parameters (Supplementary Methods, Extended Data Fig. 4i) and that

the embedding of cells onto the embryo would be relatively localized—as we would expect for a biologically meaningful embedding (Fig. 3e). This means that the distribution over locations that each cell is assigned should be localized, and indeed, the mean standard deviation of that distribution for all cells is significantly lower than that of a randomized embedding (Supplementary Methods, Extended Data Fig. 4j). Furthermore, we demonstrated that the results from novoSpaRc—as measured by correlation to observed imaging data and optimization error—were robust to optimization parameters and sources of noise, including partial sampling of cells, additive expression noise and dropouts (Extended Data Fig. 4c–h).

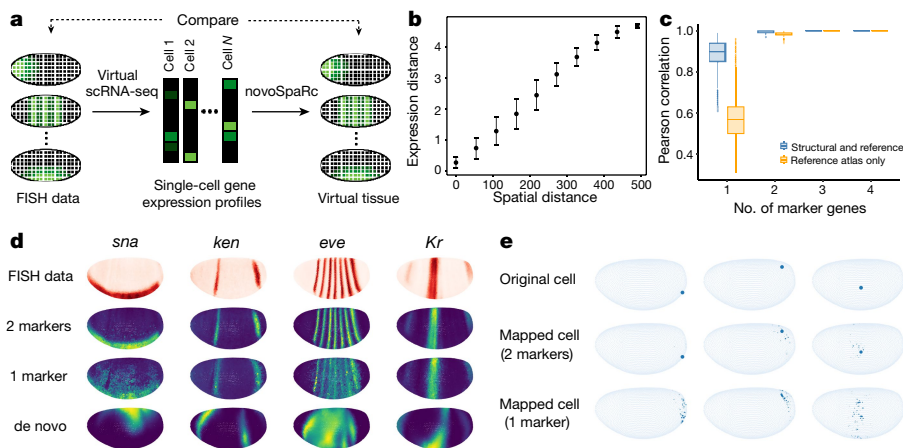


Fig. 3 | novoSpaRc accurately reconstructs the *Drosophila* embryo on the basis of the BDTNP dataset¹³. **a.** FISH data are used to create virtual scRNA-seq data, which novoSpaRc inputs to reconstruct a virtual embryo. **b.** Demonstration of the structural correspondence hypothesis. Pairwise cellular distances in expression space increase monotonically with distances in physical space. Data are mean \pm s.d. **c.** novoSpaRc spatially reconstructs the *Drosophila* embryo with only one marker gene. The quality of reconstruction (measured by Pearson correlation with FISH data) increases with the number of marker genes and saturates at perfect reconstruction at two marker genes, when using both structural information and marker gene information (blue boxes). This outperforms reconstruction that relies only on marker gene

information (yellow boxes). The results are averaged for 100 different combinations of marker genes. For the box plots, the centre line is the median, box limits are the 0.25 and 0.75 quantiles and whiskers extend to ± 2.698 s.d. **d.** Visualization of the reconstruction results for four transcription factors. The original FISH data (first row) are compared to reconstruction by novoSpaRc that exploits both structural and marker gene information (using two marker genes and one marker gene) and reconstruction without any marker gene information (de novo). **e.** The original locations of three cells are compared to their respective reconstructed locations by novoSpaRc (using two marker genes and one marker gene). The expression patterns of the marker genes used for the results in **d** and **e** are shown in Extended Data Fig. 5c.

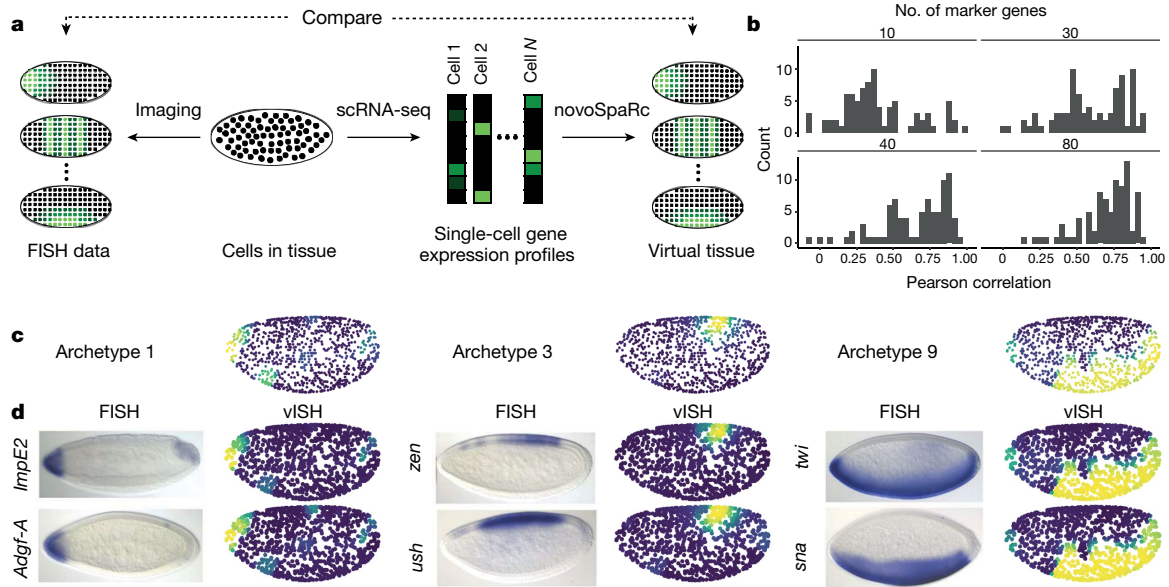


Fig. 4 | novoSpaRc identifies spatial archetypes in the *Drosophila* embryo by using scRNA-seq data. **a**, Schematic overview. The expression patterns as reconstructed by novoSpaRc are compared with the BDTNP expression values. **b**, Reconstruction of the *Drosophila* embryo using scRNA-seq data. Distributions of gene-specific Pearson correlation coefficients reflect better

reconstruction with increasing number of marker genes. **c**, Three of the spatial archetypes (1, 3 and 9) that novoSpaRc identified in the *Drosophila* embryo. **d**, Representative genes for each of the spatial archetypes depicted in **c**. FISH data (left columns) are compared against the corresponding novoSpaRc predictions ('virtual in situ hybridization' (vISH); right columns).

We next used novoSpaRc to reconstruct the stage 6 *Drosophila* embryo by using a scRNA-seq dataset¹² (Fig. 4a). In that study, 84 marker genes were required to reconstruct a virtual embryo by distributing 1,297 cells over 3,039 locations. When we used novoSpaRc with the combination of both structural information and the reference atlas, the accuracy of reconstruction increased with the number of marker genes, reaching high correlation (Pearson correlation coefficient, 0.74) with the FISH data (Fig. 4b, Extended Data Fig. 5e). The de novo, atlas-free reconstruction accurately separated the major post-gastrulation spatial domains (mesoderm, neurogenic ectoderm and dorsal ectoderm), as well as finer spatial domains (Fig. 4c, d). We clustered the reconstructed patterns of the highly variable genes and averaged to obtain a representative pattern for each cluster, which we term the 'archetype' (Methods, Supplementary Information). novoSpaRc identified numerous distinct spatial archetypes (Fig. 4c, d, Extended Data Fig. 6). We compared representative genes of each spatial archetype with FISH images to visually assess the accuracy of the spatial reconstruction. Gene patterns that were expressed through the anterior–posterior or the dorsal–ventral axis were largely recapitulated: typical genes of the mesoderm (dorsal ectoderm), such as *twi* and *sna* (*zen* and *ush*), were colocalized ventrally (dorsally) (Fig. 4c, d, right, middle). novoSpaRc accurately captured localized spatial populations (Fig. 4c, d, left, Extended Data Fig. 6, archetype 5), whereas less-extensive spatial domains were reconstructed with varying degrees of accuracy (Extended Data Fig. 6). Note that within the de novo reconstruction, accurate localization entails global transformations, as described above (Supplementary Methods).

Before proceeding to more complex tissues, we reconstructed the zebrafish embryo dataset⁵ (Extended Data Fig. 7). Similar to the original seminal study, we mapped the cells onto the surface of a hemisphere consisting of 64 distinct locations. The resulting spatial expression patterns highly correlated to the experimentally verified ones; novoSpaRc reconstructed the zebrafish embryo by using only 15 marker genes (in contrast to the 47 genes that were previously required⁵) and the accuracy of the reconstruction increased with the number of marker genes (Extended Data Fig. 7, Methods). Furthermore—in contrast to previous reconstructions—no data imputation or other specialized preprocessing was necessary⁵.

novoSpaRc charts diverse complex tissues

To further demonstrate the applicability of novoSpaRc to complex tissues, diverse sequencing technologies and different organisms, we used it to reconstruct slices of mammalian brain cerebellum²³ (Fig. 5), the mammalian kidney²⁴ (Extended Data Fig. 8) and a dataset of hundreds of individual *Drosophila* embryos²² (Extended Data Fig. 9).

The adult mammalian brain is a well-studied, highly differentiated and complex tissue. To benchmark the capabilities of novoSpaRc in reconstructing complex tissues, we used mouse cerebellum slices from a recently developed spatial transcriptomics technology²³. The dataset of sagittal sections contained 46,376 locations, corresponding to a single cell or a few cells, with a median of 52 quantified transcripts per location. To provide enough information to novoSpaRc, we first coarse-grained the data by binning neighbouring locations. This resulted in 7,704 locations, with a median of 379 quantified transcripts per location (Methods, Fig. 5a). novoSpaRc successfully reconstructed the whole transcriptome, with a Pearson correlation coefficient of 0.5 over all 15,878 genes when using 15 marker genes and 0.94 when using 50 marker genes (Fig. 5b, Supplementary Methods). Spatial expression patterns emerged when using only a few markers. For example, spatial positions of Purkinje cells were revealed by reconstructing with only five marker genes (excluding all genes exhibiting an absolute Pearson correlation coefficient with *Pcp4* of 0.25 or higher). The signal improved markedly when more markers were included (Fig. 5c). The reconstructed cerebellum slices showed concordance with the original spatial gene expression for a large number of known cell-type marker genes (Fig. 5d). To illustrate the versatility of novoSpaRc, we further applied it to a coronal section of a brain cerebellum²³, with similar results (Fig. 5e).

Next, we used novoSpaRc to spatially reconstruct a single-cell dataset from whole kidney²⁴, which is a complex tissue with stereotypical organization. In the absence of a reference atlas of gene expression, the reconstruction was performed de novo. We focused on six major cell types of the kidney (Extended Data Fig. 8) and mapped the cells onto a two-dimensional target space. The de novo reconstruction recapitulated the urine flow within the kidney sub-compartments, as shown by the spatial gene expression of corresponding marker genes (Extended Data Fig. 8). We note that, as no prior information was required for this

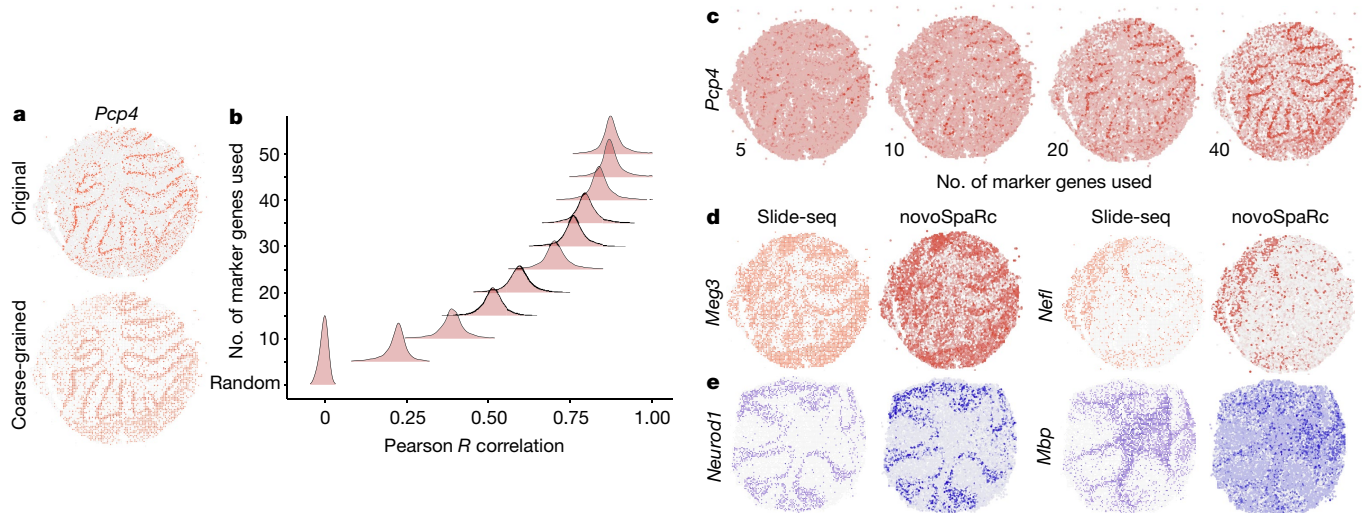


Fig. 5 | novoSpaRc reconstructs mouse cerebellum tissue. **a**, The original and the coarse-grained spatial expression of a marker of Purkinje cells (*Pcp4*) in a sagittal section of the cerebellum from direct spatial RNA sequencing²³. **b**, The overall Pearson correlation between original gene expression and gene expression predicted by novoSpaRc increases markedly as more marker genes are used. The correlation when using only five marker genes is substantially

higher than that of a random mapping of cells to locations. Density plots contain values for all 15,878 genes. **c**, The spatial gene expression of *Pcp4* is visible with only five marker genes and is enhanced as more markers are used for the reconstruction. **d**, Examples of original and predicted expression for neuronal marker genes. Reconstruction was performed with 35 marker genes. **e**, novoSpaRc accurately reconstructs a coronal section of the cerebellum²³.

reconstruction, this case demonstrates the applicability of novoSpaRc to a wide variety of medically relevant tissues.

Finally, to show that novoSpaRc can reconstruct not only a prototypical tissue but also individual samples, we used a dataset that captures

expression patterns in hundreds of individual *Drosophila* embryos²². In this dataset, the expression of four gap genes and four pair-rule genes was measured along the anterior–posterior axis for 101 and 177 embryos, respectively, providing a distribution over expression

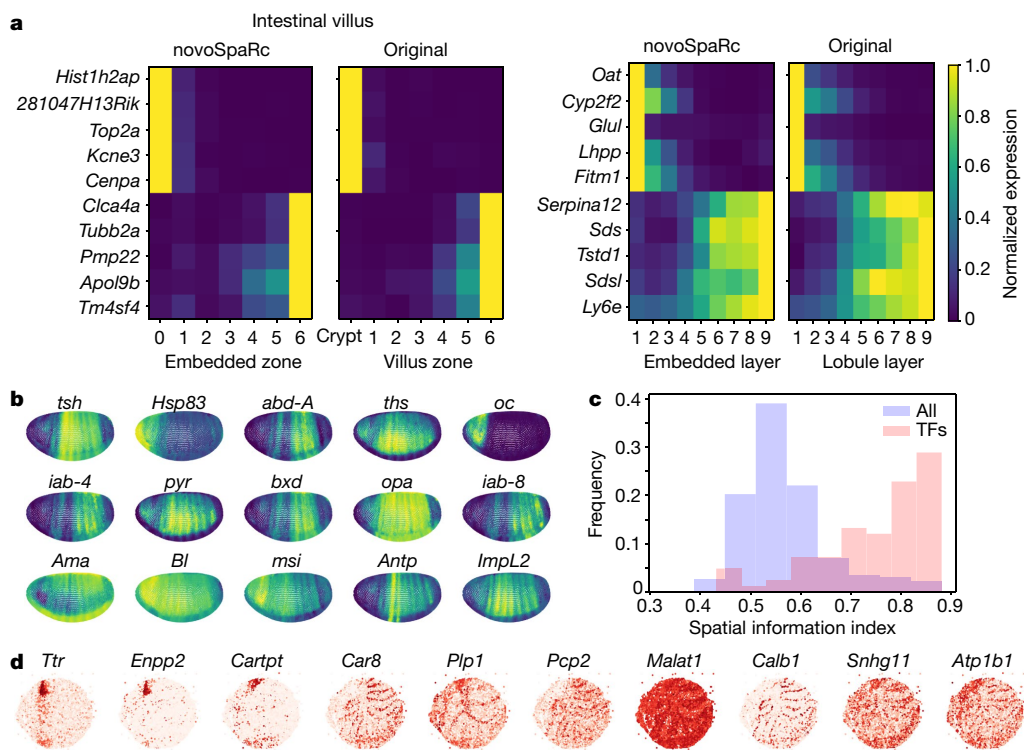


Fig. 6 | novoSpaRc identifies spatially informative genes. **a**, Identifying spatially informative genes in the mammalian intestine and liver. We identify de novo (that is, with no marker genes used) the most highly zoned genes along the crypt-to-villus axis in the intestine (left) and across the axis of a liver lobule (right). The prediction of novoSpaRc is compared against the original expression patterns. The expression level of each gene is normalized to its maximum value. **b–d**, Identifying spatially informative genes in the *Drosophila* embryo (reconstruction with the BDTNP marker genes) and a slice of the

mammalian cerebellum (reconstruction with 50 markers), using a measure of spatial autocorrelation. **b**, Expression patterns of the top 15 spatially informative genes in the *Drosophila* embryo. **c**, The spatial autocorrelation values (spatial information index) of the 84 transcription factors (TFs) chosen for the BDTNP dataset¹³ are among the highest values over all 8,924 genes of the fly embryo, demonstrating that they are identified to be highly spatially informative. **d**, Top 10 spatially informative genes (out of the top 1,000 variable genes) in a section of the cerebellum.

patterns. We used novoSpaRc to reconstruct the expression patterns of the gap and pair-rule genes for individual embryos. For a given embryo, novoSpaRc reconstruction using a reference atlas based on the gene expression within the same embryo consistently outperformed reconstruction using a reference atlas based on the averaged gene expression across all embryos in the dataset (Extended Data Fig. 9)—yet reached high correlation values for both (median Pearson correlation coefficients for reconstructing a fourth gene based on the three remaining genes were 0.99 (for expression within the same embryo) (0.95 for expression averaged across embryos) and 0.94 (0.77) for the gap and pair-rule genes, respectively).

We examined the effect of the interpolation between structural and marker gene information, and evaluated the performance of novoSpaRc by comparing it to available reconstruction methods that fully rely on a reference atlas (Seurat⁵ and DistMap¹²) (Extended Data Figs. 10, 11). novoSpaRc has several advantages when compared to the other existing methods and overall shows substantial benefits in reconstruction performance (Extended Data Fig. 10, Supplementary Discussion).

Identifying spatially informative genes

A novoSpaRc-based spatial reconstruction allows us to identify known and potentially new spatially informative genes directly from the single-cell sequencing data. For the intestine and liver datasets, we recovered highly zoned genes without a reference atlas (Methods, Supplementary Information), and found that the top inferred zoned genes were supported experimentally and/or computationally (Fig. 6a, Supplementary Tables 1, 2). Gene ontology enrichment analysis²⁵ further revealed that zonation-compatible biological processes enriched for different domains in the intestine and the liver were reconstructed by novoSpaRc (Supplementary Information). For the *Drosophila* single-cell dataset, we ranked all 8,924 genes according to their spatially informative rank (Methods, Fig. 6b, Supplementary Information), and found that transcription factors were (as known from classic genetics²⁶) among the most highly informative genes (Fig. 6c). In addition, novoSpaRc identified numerous long non-coding RNAs and transcription factors as being highly spatially informative, many of them already predicted in a previous study¹². Finally, we ranked all 15,878 genes in the cerebellum by their spatially informative rank (Methods, Fig. 6d, Supplementary Information), and found that well-known marker genes with a defined pattern of spatial expression are indeed among the highest-ranking spatially informative genes (Fig. 6d).

Discussion

Together, we have demonstrated here that one can spatially reconstruct diverse biological tissues on the basis of a simple hypothesis about how gene expression is organized in space—a structural correspondence between the distances of cells in expression space and in physical space—and that it can be used to extract spatially informative genes. Our current implementation is based on pairwise comparison of cells and locations. This requirement can be readily altered. In fact, it is compelling to hypothesize that within certain biological contexts, different cell types may require higher-order interactions or exhibit different principles of spatial organization. Furthermore, we stress that because of the availability of general mathematical results in optimal-transport theory, our framework is versatile and can support a variety of alternative ways to compare distances in expression and physical space by varying the optimization loss functions (Methods, Supplementary Methods). Such alternative schemes are not currently supported by novoSpaRc, but could be implemented.

Our data analyses and the success of the reconstructions by novoSpaRc suggest that we have identified a general principle for how gene expression is organized in tissue space (Supplementary Discussion). It will be interesting to find tissues for which this organization principle

is weak or not valid. However, this principle may be underestimated, as most of the single-cell data available are relatively shallow and noisy. Our data also suggest that many more genes than perhaps anticipated are involved in spatial features and functions (including physiology and pathophysiology) of tissue. We have demonstrated that we can systematically identify at least a subset of these genes directly from single-cell data. In the future, we will extend these analyses to identify genes that are predicted to functionally interact in space. Finally, our developed framework can be flexibly extended beyond spatial reconstruction. We are currently using it to recover different types of biological signals, such as temporal progression on short (for example, cell cycle) and long (for example, developmental) timescales.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1773-3>.

- Shapiro, E., Bizener, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618–630 (2013).
- Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34**, 1145–1160 (2016).
- Altschuler, S. J. & Wu, L. F. Cellular heterogeneity: do differences make a difference? *Cell* **141**, 559–563 (2010).
- Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58**, 610–620 (2015).
- 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).
- Achim, K. et al. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* **33**, 503–509 (2015).
- Halpern, K. B. et al. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature* **542**, 352–356 (2017).
- Durruthy-Durruthy, R. et al. Reconstruction of the mouse otocyst and early neuroblast lineage at single-cell resolution. *Cell* **157**, 964–978 (2014).
- Waldhaus, J., Durruthy-Durruthy, R. & Heller, S. Quantitative high-resolution cellular map of the organ of Corti. *Cell Rep.* **11**, 1385–1399 (2015).
- Moor, A. E., et al., Spatial reconstruction of single enterocytes uncovers broad zonation along the intestinal villus axis. *Cell* **175**, 1156–1167 (2018).
- Habib, N. et al. Div-Seq: Single-nucleus RNA-seq reveals dynamics of rare adult newborn neurons. *Science* **353**, 925–928 (2016).
- Karaiskos, N. et al. The *Drosophila* embryo at single-cell transcriptome resolution. *Science* **358**, 194–199 (2017).
- Berkeley *Drosophila* Transcription Network Project. <http://bdnptn.lbl.gov:8080/Fly-Net/>.
- Monge, G. Mémoire sur la théorie des déblais et des remblais. *Historie de l'Académie Royale des Sciences de Paris* **1781**, 666–704 (1781).
- Villani, C. *Topics in Optimal Transportation* (American Mathematical Society, 2003).
- Villani, C. *Optimal Transport: Old and New* Vol. **338** (Springer, 2008).
- Schiebinger, G. et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 928–943 (2019).
- Forrow, A. et al. *Statistical optimal transport via geodesic hubs*. Preprint at <https://arxiv.org/abs/1806.07348> (2018). If ref. 18 (preprint) has now been published in final peer-reviewed form, please update the reference details if appropriate.
- Mémoli, F., On the use of Gromov–Hausdorff distances for shape comparison. In *Eurographics Symposium on Point-Based Graphics* (eds Botsch, M. & Pajarola, R.) (Eurographics Association, 2007).
- Peyré, G., Cuturi, M. & Solomon, J. Gromov–Wasserstein averaging of kernel and distance matrices. In *Proc. 33rd International Conference on Machine Learning* (Journal of Machine Learning Research, 2016).
- Cuturi, M. Sinkhorn distances: lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26* (eds Burges, et al.) (NIPS, 2013).
- Petkova, M. D., Tkačik, G., Bialek, W., Wieschaus, E. F. & Gregor, T. Optimal decoding of cellular identities in a genetic network. *Cell* **176**, 844–855 (2019).
- Rodrigues, S. G. et al. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).
- Park, J. et al. Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science* **360**, 758–763 (2018).
- Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).
- Nüsslein-Volhard, C. & Wieschaus, E. Mutations affecting segment number and polarity in *Drosophila*. *Nature* **287**, 795–801 (1980).

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

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Data pre-processing

For the cases for which normalized data was not available or used by the authors, we adopted the standard library size normalization in log-space, for example, if d_{ij} represents the raw count for gene i in cell j , we normalized it as

$$d_{ij} \rightarrow d'_{ij} = \log_2 \left(10^5 \times \frac{d_{ij}}{\sum_k d_{kj}} + 1 \right).$$

Highly variable genes were identified by plotting the dispersion of a gene as a function of its mean and selecting the outliers above cut-off values (usually 0.125 for the mean and 1.5 for the dispersion).

In the Slide-seq datasets²³, we summed up the transcriptomes of neighbouring cells by rounding the coordinates of the physical locations to the next integer multiple of 50. This resulted in a total of 8,331 (9,890) cells for the sagittal (coronal) section of the cerebellum. Low-quality locations were further filtered out by requiring at least 50 genes per cell, resulting in a total of 7,704 (8,258) for the sagittal (coronal) section. Marker genes for the reconstruction were randomly selected from the set of 747 genes. As one of the means of benchmarking the different reconstructions was to visually assess the expression pattern of *Pcp4*, we ensured that no genes with a Pearson correlation of $|R| \geq 0.25$ with *Pcp4* were selected as marker genes.

Mathematical formulation of novoSpaRc

The procedure used by novoSpaRc includes several steps. We first compute the graph-based distance matrices for N single cells in expression space, $D^{\text{exp}} \in R^{N \times N}$, and for M locations, $D^{\text{phys}} \in R^{M \times M}$ (Extended Data Fig. 1a, Supplementary Methods). Then, optionally, if a reference atlas is available, we compute the matrix of disagreement, $D^{\text{exp,phys}} \in R^{N \times M}$, between each of the cells to each of the locations, on the basis of the inverse correlation between the partial expression profile for each location given by the reference atlas and the respective expression profile for each cell. Equipped with these measures of intra- and inter-dataset distances, we set out to find an optimal (probabilistic) assignment of each of the single cells to cellular physical locations.

We formulate this problem as an optimization problem within the generalized framework of optimal transport^{14–16}. Optimal transport is a mathematical framework that was first established in the eighteenth century by Gaspard Monge and was initially motivated by the question of the optimal (minimal cost) way to rearrange one pile of dirt into a different formation (the respective minimal cost is appropriately termed the ‘earth mover’s distance’). The framework evolved both theoretically and computationally^{15,16,21} and was extended to the correspondence between pairwise similarity measures via the Gromov–Wasserstein distance^{19,20}. Thus, in our context, it allows us to build on these results and tools to feasibly solve the cellular assignment problem.

We aim to find a probabilistic embedding, $T \in R_+^{N \times M}$, of N single cells to M locations that would minimize the discrepancy between the pairwise graph-based distances in expression space and in physical space, and—if a reference atlas is available—simultaneously minimize the discrepancy between its values across the tissue and the expression profiles of embedded single cells. For each cell i , the value of T_{ij} is the relative probability of embedding it to location j . These optimization requirements over T are formulated as follows. We measure the pairwise discrepancy of T for the expression and physical spaces using the Gromov–Wasserstein discrepancy¹⁹

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

where L is a loss function; specifically, we use the quadratic loss $L(a, b) = \frac{1}{2} |a - b|^2$. This term captures our preference to embed single cells such that their pairwise distance structure in expression space would resemble their pairwise distance structure in physical space. Intuitively, if expression profiles that correspond to cells i and k are embedded into cellular locations j and l , respectively, then the distance between i and k in expression space should correspond to the distance between j and l in physical space (for example, if i and k are close expression-wise they should be embedded into close locations, and vice versa). The discrepancy measure weighs these correspondences by the respective probability of the two embedding events.

To measure the match to existing prior knowledge, or an available reference atlas, we consider

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

This term represents the average discrepancy between cells and locations according to the reference atlas, weighted by T . Finally, we regularize T by favouring embeddings with higher entropy, where entropy is defined as

$$H(T) = - \sum_{i,j=1} T_{i,j} \log T_{i,j}$$

Intuitively, higher entropy implies more uncertainty in the mapping. Entropic regularization drives the solution away from arbitrary deterministic choices and was shown to be computationally efficient²¹.

Putting these together, we define the optimization problem for the optimal probabilistic embedding T^* :

$$T^* = \text{argmin}(1 - \alpha) D_1(T) + \alpha D_2(T) - \varepsilon H(T)$$

subject to

$$\sum_j T_{i,j} = p_i \quad \forall i \in \{1, \dots, N\}$$

$$\sum_i T_{i,j} = q_j \quad \forall j \in \{1, \dots, M\}$$

where 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 constraints reflect the fact that the transport plan T should be consistent with the marginal distributions $p \in \{p \in R_+^N; \sum_i p_i = 1\}$ and $q \in \{q \in R_+^M; \sum_j q_j = 1\}$, over the original input spaces of expression profiles and cellular locations, respectively.

These marginals can capture, for example, varying densities of single cells in the vicinity of different cellular grid locations, or the quality of different single-cell expression profiles (hence forcing low-quality single cells to have a smaller contribution to the reconstructed tissue-wide expression patterns). When such prior knowledge is lacking, p and q could be set to be uniform distributions.

We derive an efficient algorithm for this optimization problem, inspired by the combined results for entropically regularized optimal transport²¹ and mapping based on Gromov–Wasserstein distance between metric-measure spaces²⁰ (Supplementary Methods).

Then, given the original single-cell expression profiles, represented by a matrix $Y \in R^{N \times g}$ (for N single cells and g genes), and the inferred probabilistic embedding $T \in R_+^{N \times M}$ (for N single cells and M locations), we can derive a virtual in situ hybridization (vISH), $S = Y^T T \in R_+^{g \times M}$

(for g genes and M locations), which contains the gene expression values for every cellular location of the target space.

Note again that because our mapping is probabilistic, each of the cellular locations of the vISH does not correspond to a single cell in the original data. Rather, the vISH represents the expression patterns over an averaged, stereotypical tissue from which the single cells could have originated.

novoSpaRc algorithm

To spatially reconstruct gene expression, novoSpaRc performs the following steps:

1. Read the gene expression matrix.
 - 1a. Optional: select a random set of cells for the reconstruction;
 - 1b. Optional: select a small set of genes (for example, highly variable).
2. Construct the target space.
3. Set up the optimal-transport reconstruction.
 - 3a. Optional: use existing information of marker genes, if available.
4. Perform the spatial reconstruction including:
 - 4a. Assigning cells a probability distribution over the target space;
 - 4b. Deriving a vISH for all genes over the target space.

The novoSpaRc package, system requirements, installation guide and demo instructions are provided at <https://github.com/rajewsky-lab/novosparc>.

Generating in silico single-cell data for the BDTNP dataset

To test the performance of novoSpaRc with single-cell resolution ground truth, we generated an in silico single-cell dataset for the BDTNP data¹³. In that case we have access to expression profiles for different locations across the embryo. We effectively dissociate the embryo by taking these expression profiles to be the expression profiles of single cells in our in silico set, masking their true original locations, and use novoSpaRc to reconstruct the original embryo (which may be done at lower spatial resolution).

Identification of spatial archetypes

The identification of spatial archetypes is performed by clustering the spatial expression of a given set of genes. The gene expression is first clustered by hierarchical clustering at the vISH level, although in principle different clustering methods can be used. The number of archetypes is chosen by visually inspecting the resulting dendrogram. The expression values of each gene of the cluster are then averaged per location to produce the spatial archetype for that cluster. Representative genes for each cluster are identified by computing the Pearson correlation of each gene within the cluster against the spatial archetype. The derivation of the spatial archetypes strongly depends on the set of genes used. We observed that the set of highly variable genes generally resulted in sensible spatial archetypes. A list of genes that correspond to each archetype is provided in the Supplementary Information.

Identification of zoned genes

For tissues with one-dimensional symmetry, we produce a ranking of highly zoned genes, both according to the original spatial expression patterns (Extended Data Fig. 2c, d) and the reconstructed patterns (Fig. 6a).

The input is a spatial expression matrix (either original or reconstructed), specifying the expression level of each gene in each of the spatial zones. Then, to find a ranked list of genes that are highly zoned towards the first or last spatial zones (for example, crypt in the liver), we first select all genes (i) whose highest expression occurs in that respective zone; (ii) whose maximum expression value is in the top 1% of all genes; and (iii) that are statistically significantly zoned. To compute the zonation significance of individual genes, we used a non-parametric test based on the Kendall's tau coefficient. The Kendall's tau coefficient

is a measure for the correspondence between two ranked lists—in our case, the expression values of a given gene over consecutive spatial zones and the numbering of the zones. Finally, the remaining genes are ranked according to their centre of mass.

The lists of predicted zoned genes based on novoSpaRc's reconstruction for the mammalian intestine and liver are available in the Supplementary Information.

Gene ontology enrichment

We used GOrilla for gene ontology (GO) enrichment analysis²⁵, in which GO enrichment was computed on the basis of target and background lists of genes (Supplementary Methods). For both the target and background lists of genes, we selected genes that had a maximum expression value in the top 10% of all genes. The target lists for genes that were zoned towards the boundaries of the one-dimensional spatial axes (crypt and V6 in intestine; layers 1 and 9 in liver) were further filtered to contain only genes that are statistically significantly zoned, as described in 'Identification of zoned genes'. The background lists contained the corresponding complements of the target lists.

Identification of spatially informative genes

We use a spatial autocorrelation measure to rank genes as spatially informative. Specifically, we use Moran's I as a measure for global spatial autocorrelation. For each individual gene i , the Moran's I score for its spatial expression, y_i , over n cellular locations is:

$$I = \frac{n}{S_0} \frac{\sum_{i,j} z_i w_{i,j} z_j}{\sum_i z_i^2}$$

where $z_i = y_i - \bar{y}_i$, \bar{y}_i is the mean expression of gene i , $S_0 = \sum_{i,j} w_{i,j}$ and $w_{i,j}$ is a spatial weights matrix, which we base on a k -nearest neighbours graph for each cellular location ($k=8$). To calculate the Moran's I score and the respective P values for different genes, we used the implementation of PySAL, a Python spatial analysis library²⁷.

The Moran's I scores with their respective P values, based on novoSpaRc's reconstructions for all genes of the *Drosophila* embryo, zebrafish embryo and cerebellum, are available in the Supplementary Information.

Reporting summary

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

Data availability

The scRNA-seq datasets were acquired from the Gene Expression Omnibus (GEO) database with the following accession numbers: GSE99457 for the intestinal epithelium¹⁰, GSE84490 for the liver⁷, GSE95025 for the *Drosophila* embryo¹², GSE66688 for the zebrafish embryo⁵ and GSE107585 for the kidney²⁴. The cerebellum Slide-seq datasets²³ were acquired from the Broad Institute Single Cell Portal (https://portals.broadinstitute.org/single_cell/study/slide-seq-study). The individual *Drosophila* embryos dataset²² is available as a supplementary information file of the original manuscript²². The BDTNP dataset was downloaded directly from the BDTNP webpage¹³.

Code availability

A Python package for novoSpaRc, and the scripts for reconstructing selected tissues presented in the manuscript, are provided at <https://github.com/rajewsky-lab/novosparc>.

27. Rey, S. J. & Anselin, L. in *Handbook of Applied Spatial Analysis* (eds Fischer, M. & Getis, A.) 175–193 (Springer, 2010).
28. Tomancak, P. et al. Global analysis of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.* **8**, R145 (2007).

Article

Acknowledgements We thank A. Murray, A. Regev, T. Gregor, P. Rigollet, all members of our labs and many colleagues in the field for valuable comments and discussions. We thank L. Friedman for help with graphic design and illustration. This work was supported by the Israeli Science Foundation, through the I-CORE program (N.F.) and an Alexander von Humboldt Foundation Research Award (N.F.). N.K. was supported by grants DFG/GZ (Geschäftszeichen): RA 838/8-2 and DFG/GZ: KA 5006/1-1; and HGF Neurocure/GZ 0036-Phase 2-3. M.N. was supported by the James S. McDonnell Foundation, Schmidt Futures, the Israel Council for Higher Education and the John Harvard Distinguished Science Fellows Program within the FAS Division of Science of Harvard University. N.R. thanks Anna-Carina for useful discussions.

Author contributions N.R. conceived the structural correspondence assumption. N.K. and N.R. demonstrated the feasibility of such an assumption for spatial inference of toy models. M.N., N.K., N.F. and N.R. designed the research. M.N. developed the

optimal-transport-based spatial inference framework. M.N. and N.K. implemented the method and performed computational and data analyses. N.F. and N.R. supervised the study. All authors wrote the manuscript.

Competing interests The authors declare no competing interests.

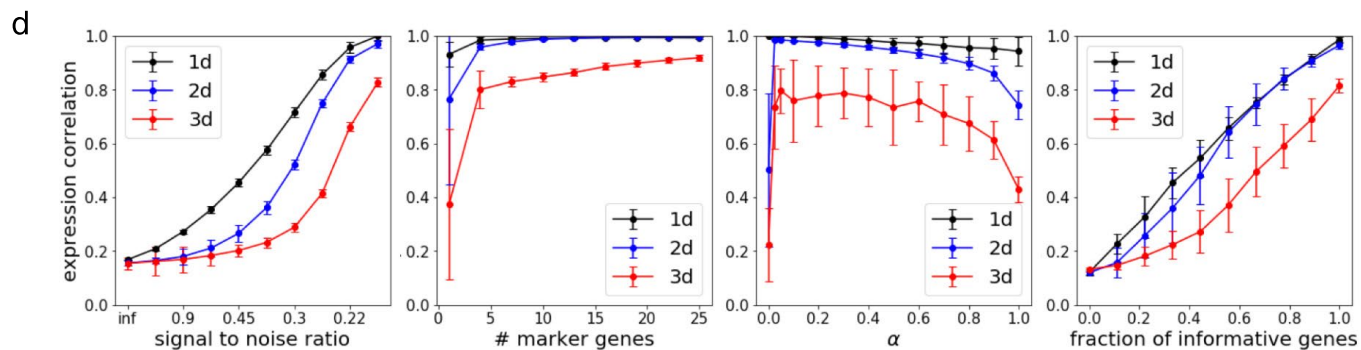
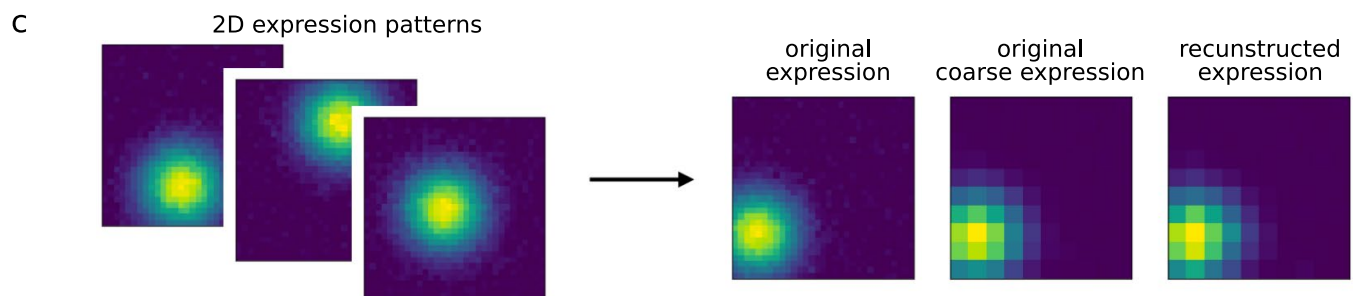
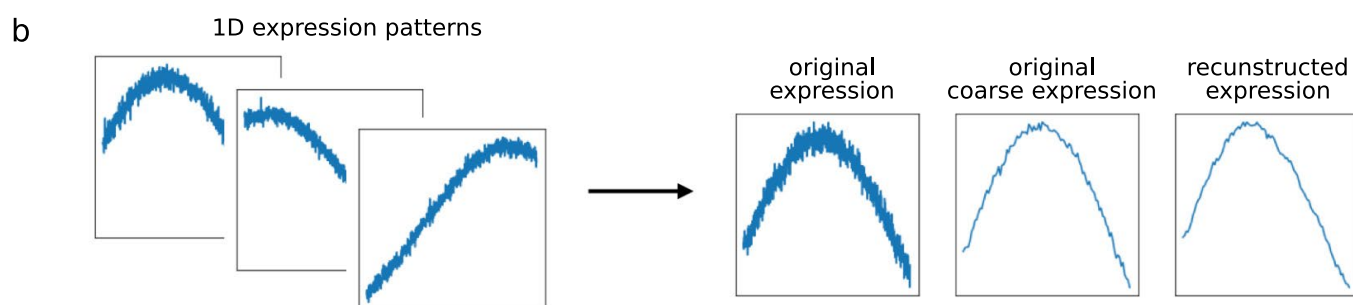
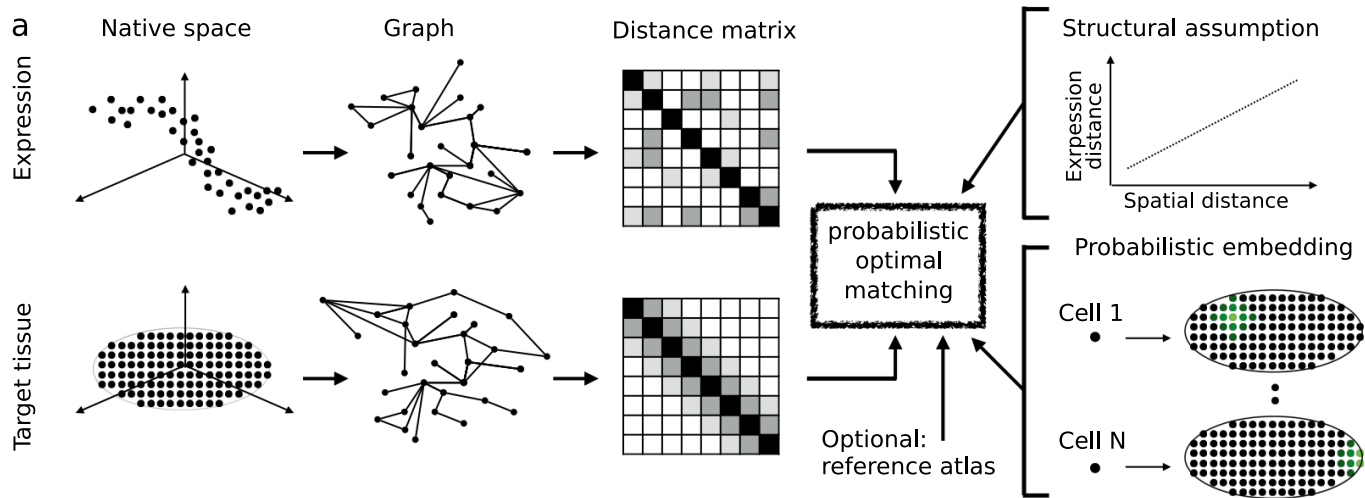
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1773-3>.

Correspondence and requests for materials should be addressed to N.F. or N.R.

Peer review information *Nature* thanks Eileen Furlong and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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

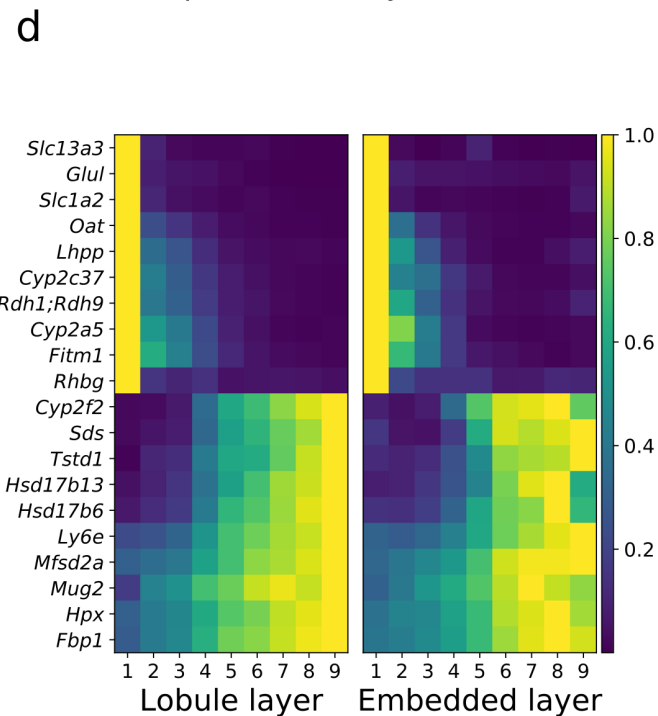
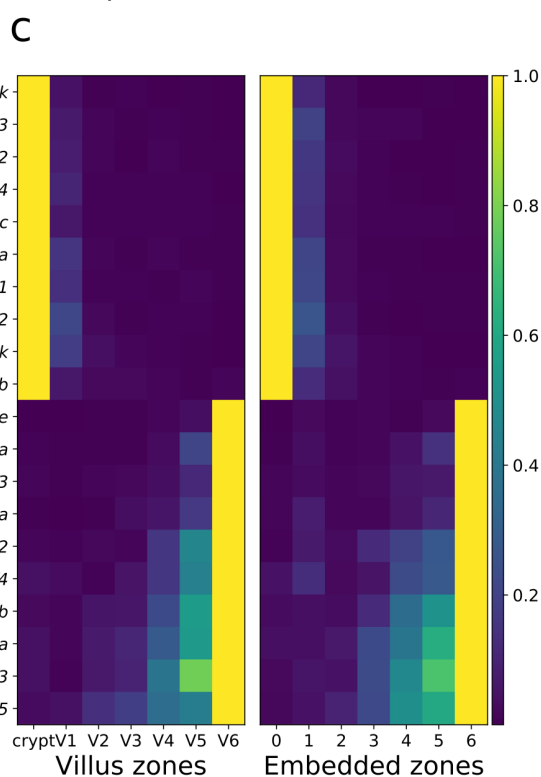
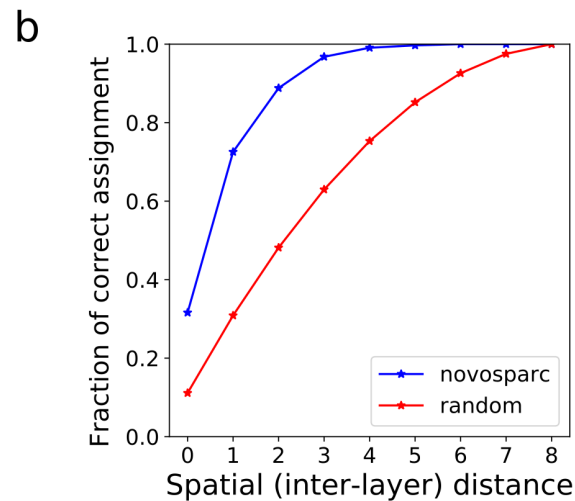
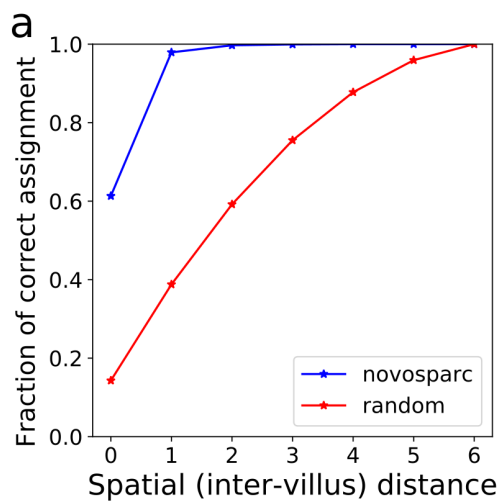


Extended Data Fig. 1 | See next page for caption.

Article

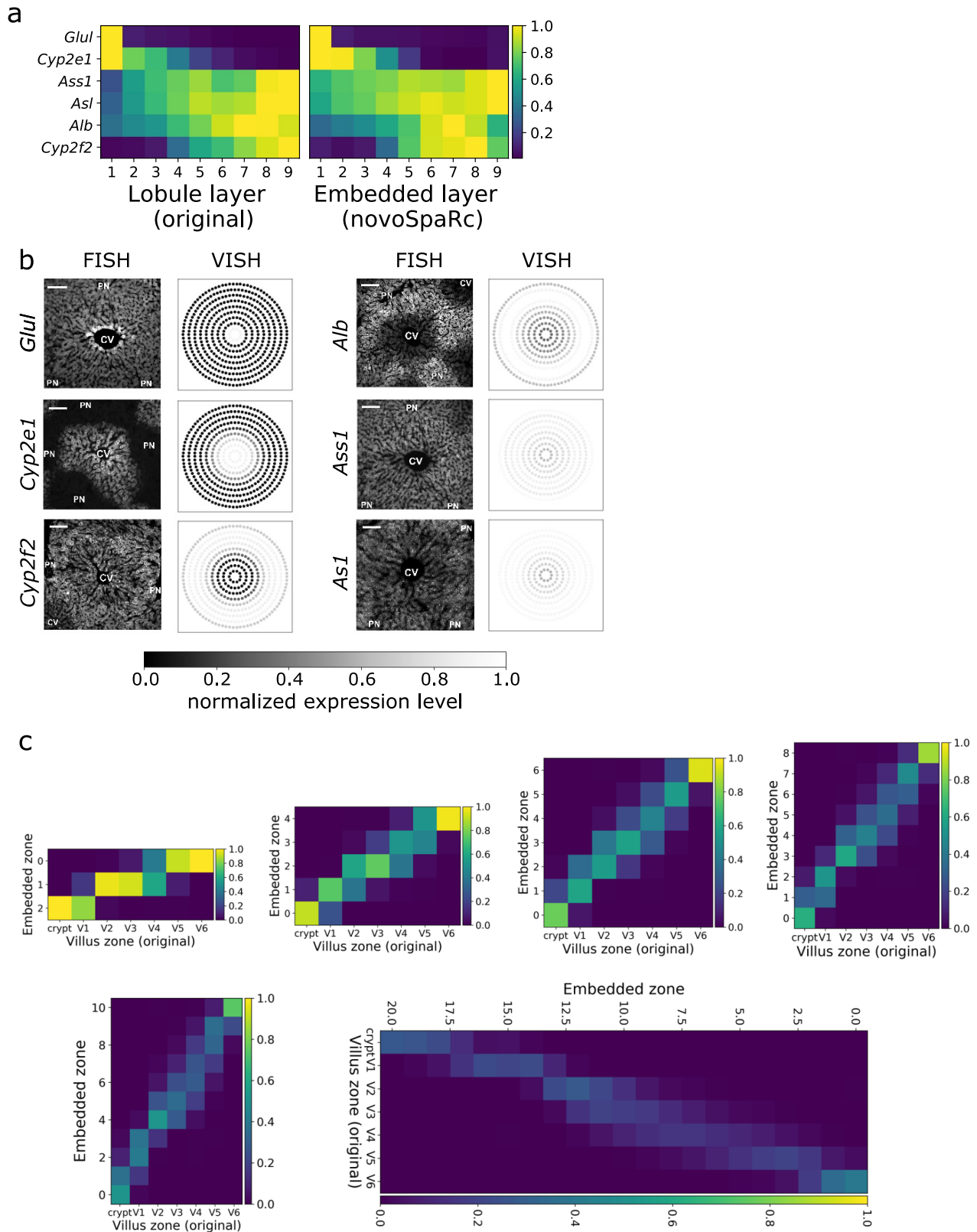
Extended Data Fig. 1 | Overview of probabilistic optimal matching using novoSpaRc and corresponding generative model. a, Based on the raw data of single cells in expression space and locations along a grid resembling the target tissue, graph structures are computed and distance matrices are derived from these graphs (Supplementary Methods). The two branches, and potentially a reference atlas, are aligned using novoSpaRc, under our structural correspondence assumption (distance in expression space on average monotonically increases with distance in physical space) and by using probabilistic embedding (Supplementary Methods). **b, c**, Left, visualization of noisy expression patterns for three random genes in models for 1-dimensional (1D) (**b**) and two-dimensional (2D) (**c**) tissues. Right, the original expression pattern for a representative gene, its coarse-grained representation

(decreased spatial resolution) and its reconstruction using novoSpaRc. **d**, The Pearson correlation of the reconstructed expression pattern data to the original synthetic expression data increases with increasing signal-to-noise ratio, with the number of marker genes and with the fraction of informative genes, and exhibits non-monotonic behaviour with the α parameter. We note that α is an interpolation parameter (defined in the Methods section 'Mathematical formulation of novoSpaRc') between using only a reference atlas ($\alpha = 1$) and using only structural information (driven by the structural correspondence assumption) ($\alpha = 0$). Results are averaged over 100 instantiations of the generative model; data are mean \pm s.d. The generative model and its default parameters are described in the Supplementary Methods.



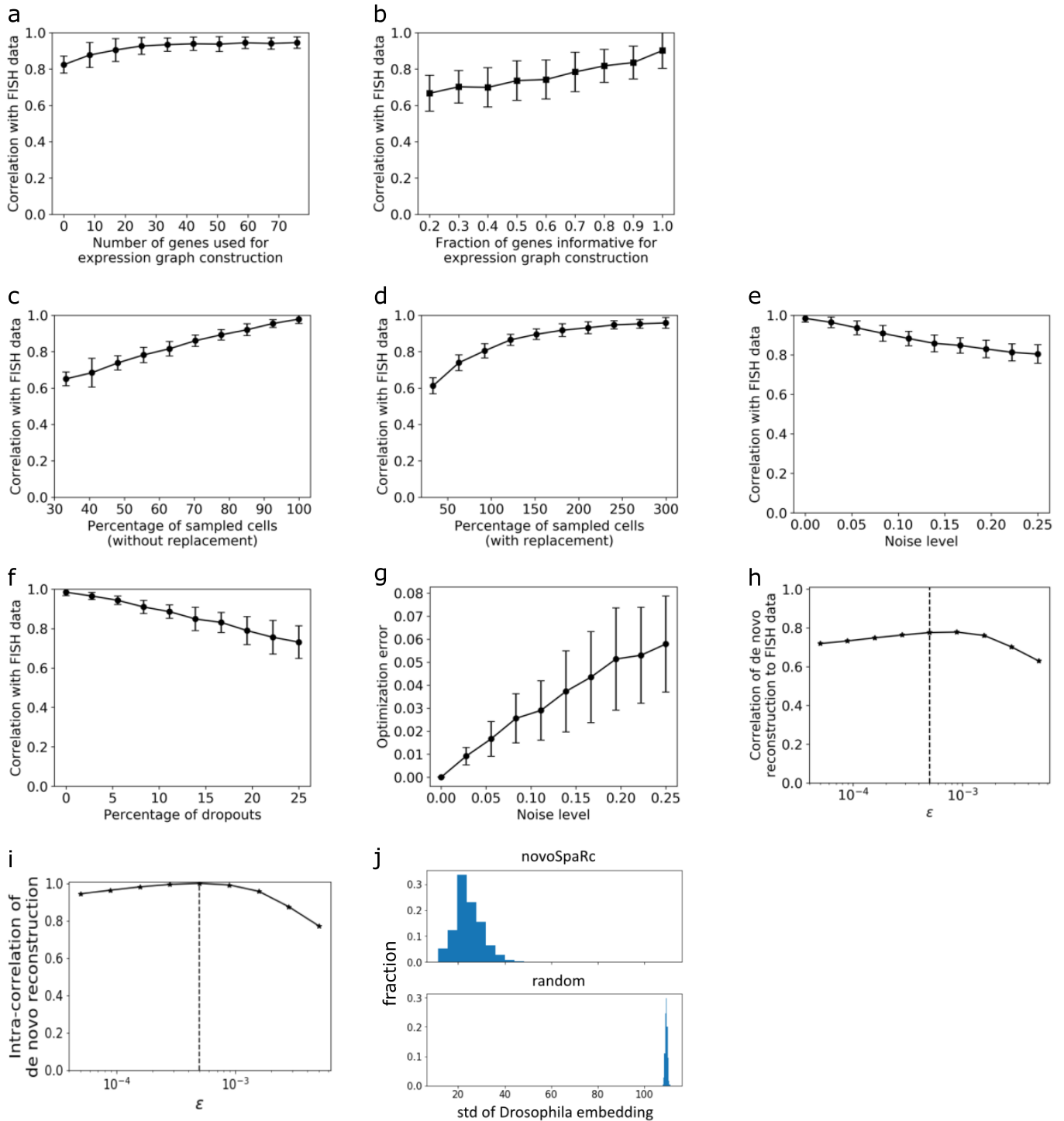
Extended Data Fig. 2 | Evaluation of novoSpaRc reconstruction of the intestinal epithelium and the liver lobule. a, b, The fraction of cells in the crypt-to-villus axis (a) and the liver lobule axis (b) that is correctly assigned to its corresponding original villus zone¹⁰ and original lobule layer⁷, or is assigned to a zone up to *d* zones away from the original zone (x axis), is substantially higher than that of random assignment. **c, d,** novoSpaRc reconstructs the spatial expression patterns of the top zonated genes in the intestinal

epithelium (c) (10 top zonated genes towards the crypt, and 10 top zonated genes towards V6) and in the liver lobule (d) (10 top zonated genes towards the central vein (CV), and 10 top zonated genes towards the portal node (PN)). 2810417H13Rik is also known as *Pclaf*. The selection of the top zonated genes is described in the Methods. The expression level of each gene in c and d is normalized to its maximum value.



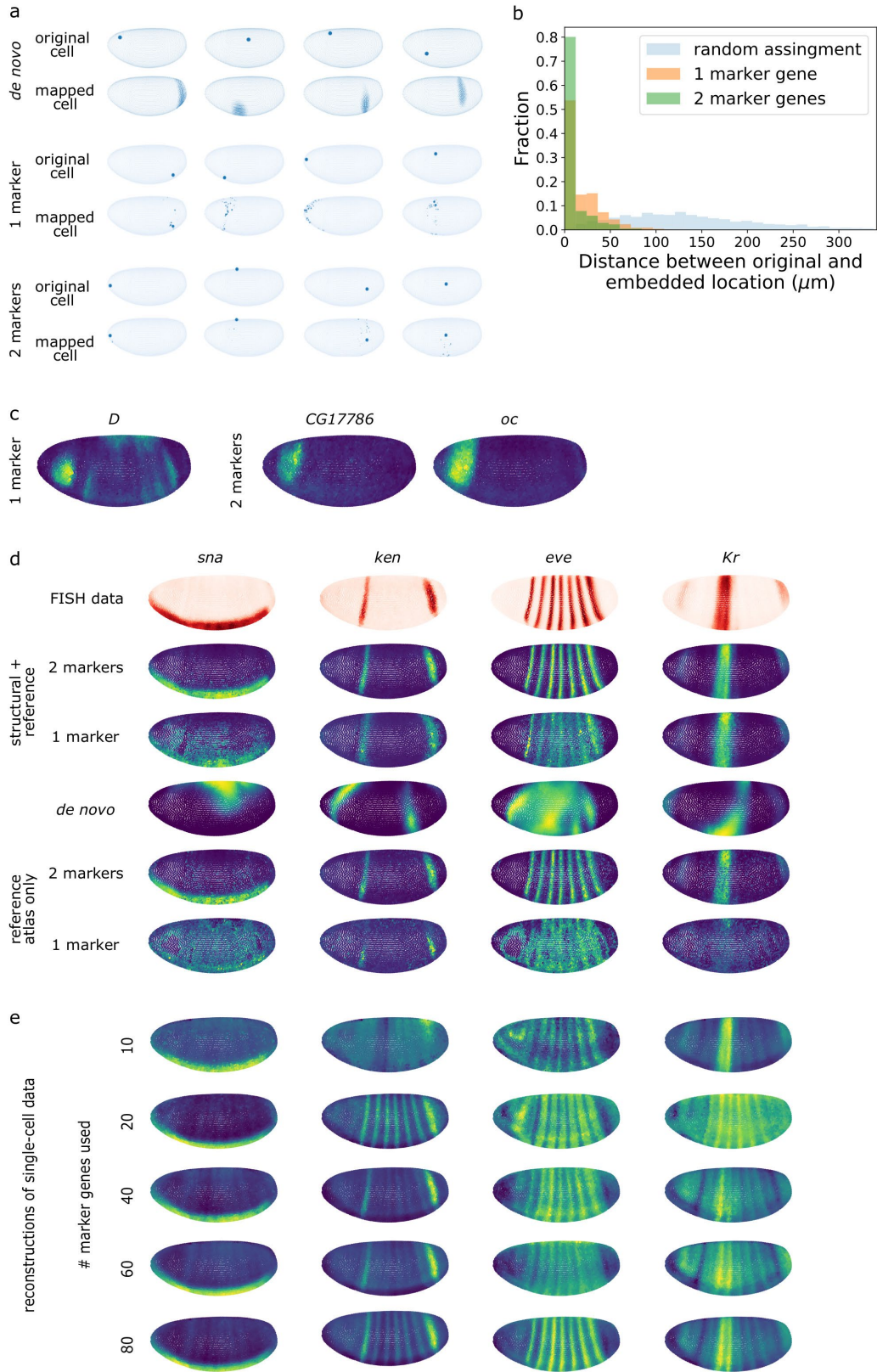
Extended Data Fig. 3 | novoSpaRc reconstruction of the intestinal epithelium and the liver lobule is robust and consistent with changing grid resolution. a, b. Examples of FISH expression patterns of six zoned genes across the liver lobules, comparing the reconstructed (de novo vISH data) expression patterns produced by novoSpaRc to the expression patterns reported in a previous study⁷ (a), and the original (FISH) data (adapted from the same study⁷) (b). The visualization in a is a heat map, which shows the expression values of each gene across the lobule layers. The visualization of the

reconstructed vISH data in b is intended to be comparable to the FISH images, and therefore the 1D reconstructed coordinates are projected onto a polar coordinate system (central vein–middle, portal node–outer circumference). c, The successful de novo reconstruction of the intestinal epithelium dataset⁹ is achieved for varying numbers of layers used for the target space (including both lower and higher numbers of layers compared with the original number (seven) of reference layers). The expression level of each gene is normalized to its maximum value.



Extended Data Fig. 4 | novoSpaRc reconstruction of the *Drosophila* embryo on the basis of the BDTNP dataset is robust and self-consistent. a, b, The Pearson correlation of the reconstructed expression patterns to the original FISH expression data¹² increases with the number of genes used to construct the structural cellular graph in expression space (a), and with the fraction of those genes that are spatially informative (b). Spatially non-informative genes in this case were simulated as random Gaussian variables with mean and s.d. comparable to that of the original set of genes. c–f, The Pearson correlation of the reconstructed expression patterns to the original FISH expression data¹² increases with the percentage of sampled single cells (without replacement) (c) and with the percentage of sampled single cells (with replacement) (d), and steadily decreases with noise level (e) and with the percentage of dropouts in the data (f). g, The mean value and variance of the optimization objective

function (which we aim to minimize) increase with noise level. The results in a–g are averaged over 100 random choices of two marker genes; data are mean \pm s.d. h, The Pearson correlation of the de novo reconstructed expression patterns to the original FISH data varies gradually with the entropic regularization parameter ϵ . i, The Pearson correlation of embedded de novo expression patterns of the BDTNP dataset¹² for different values of the entropic regularization parameter ϵ with the expression pattern for $\epsilon = 5 \times 10^{-5}$ (vertical dotted line). j, The spatial s.d. of embedded cells over the *Drosophila* embryo of the BDTNP dataset derived from de novo reconstruction by novoSpaRc is significantly lower than the s.d. derived from randomized embedding ($P < 10^{-200}$, two-sided Kolmogorov–Smirnov test). Histograms show results for all 3,039 cells.



Extended Data Fig. 5 | See next page for caption.

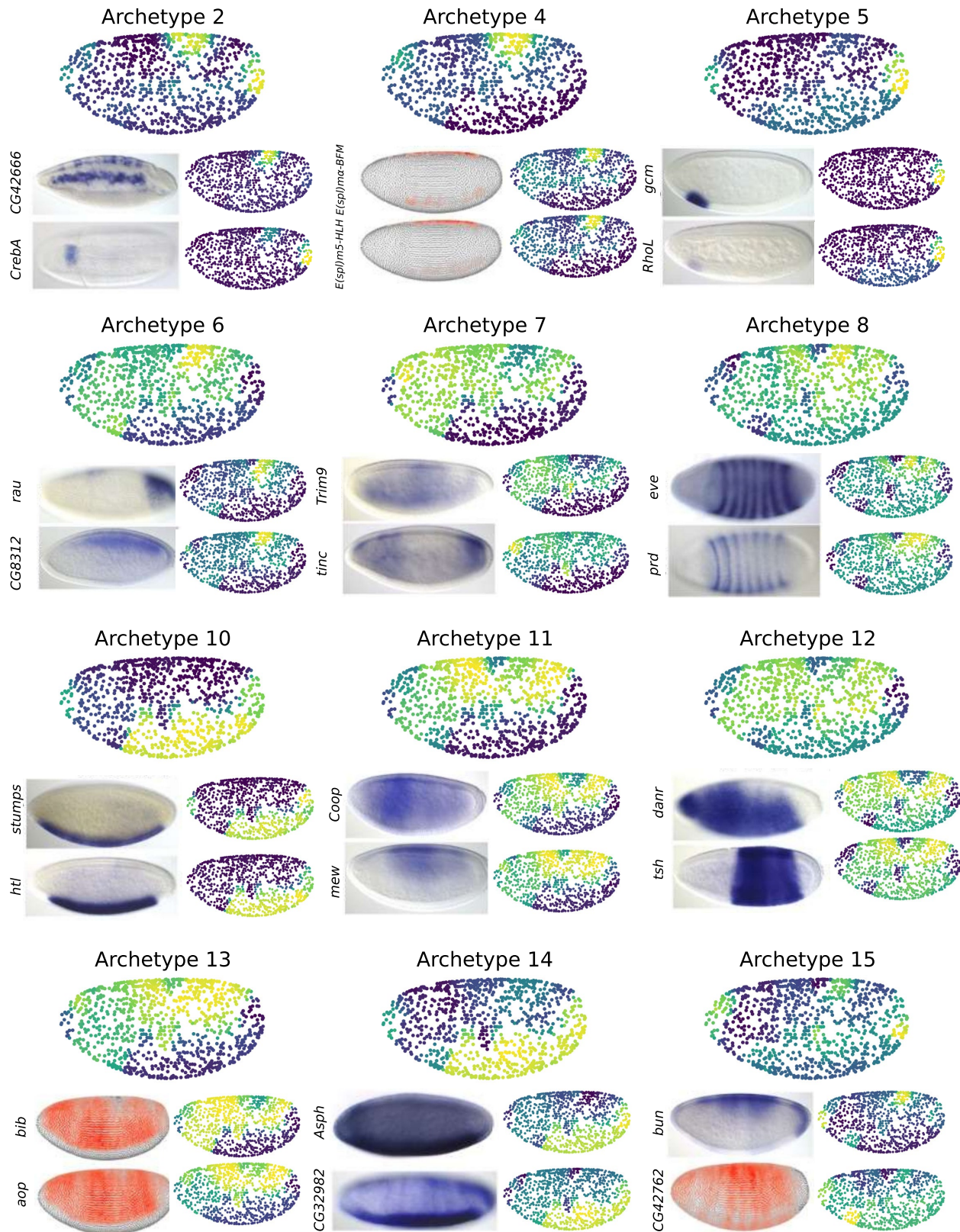
Extended Data Fig. 5 | novoSpaRc accurately reconstructs the *Drosophila* embryo on the basis of the BDTNP dataset and single-cell data. **a**, Examples of mapping probabilities of single cells produced by novoSpaRc for the *Drosophila* embryo, using the BDTNP dataset¹³. The predicted spatial positions of cells are distributed over relatively many locations when reconstruction is done de novo, and are more localized when marker genes are used.

b, Histogram of Euclidean distances between the original cellular location of single cells and the most likely location predicted by novoSpaRc using one and two marker genes, compared to a histogram for random spatial predictions.

c, The expression patterns of the two marker genes and one marker gene that were used for the results presented in **a**, **b** and in Fig. 3d, **e**. **d**, Visualization of

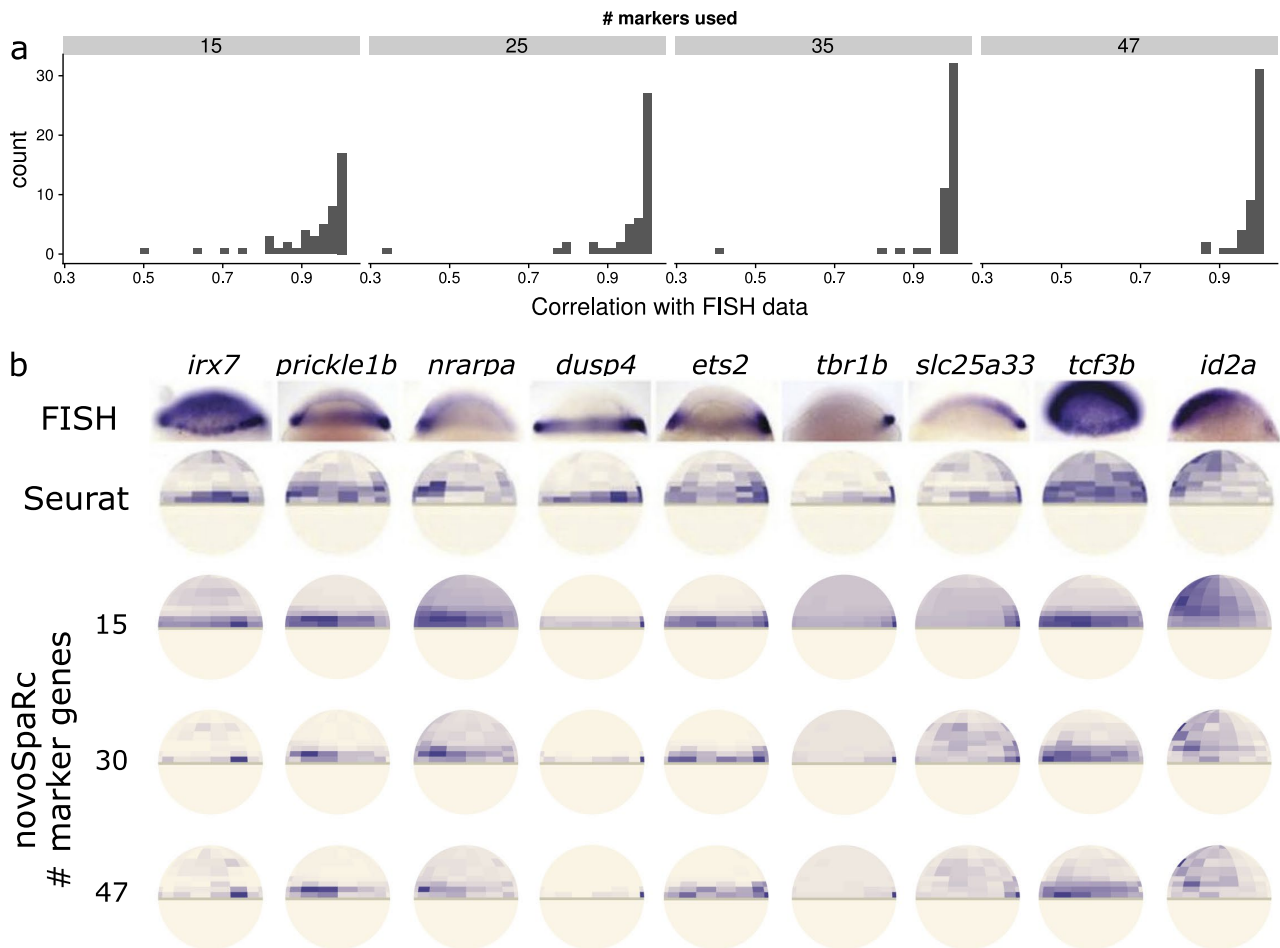
reconstruction results for four transcription factors. The original FISH data are compared to reconstruction by novoSpaRc that exploits both structural and marker gene information (using two marker genes and one marker gene), and reconstruction without any marker gene information (de novo).

Reconstruction that uses both structural and marker gene information (or a reference atlas) outperforms reconstruction that is based solely on a reference atlas. **e**, Visualization of novoSpaRc-based reconstruction results for the four transcription factors, based on single-cell data¹² that exploit both structural and marker gene information (using 10–80 marker genes). The results in **a–d** are based on the BDTNP dataset¹³, and the results in **e** are based on a single-cell dataset¹².



Extended Data Fig. 6 | novoSpaRc identifies spatially informative archetypes by using scRNA-seq data for the *Drosophila* embryo. The archetypes shown complement those of Fig. 4c, d. Preferred spatial positioning is denoted by colouring ranging from blue (low) to yellow (high). FISH images were taken from the BDGP database²⁸. For genes for which an image was not available, DVEX¹² was used instead. Two representative genes are

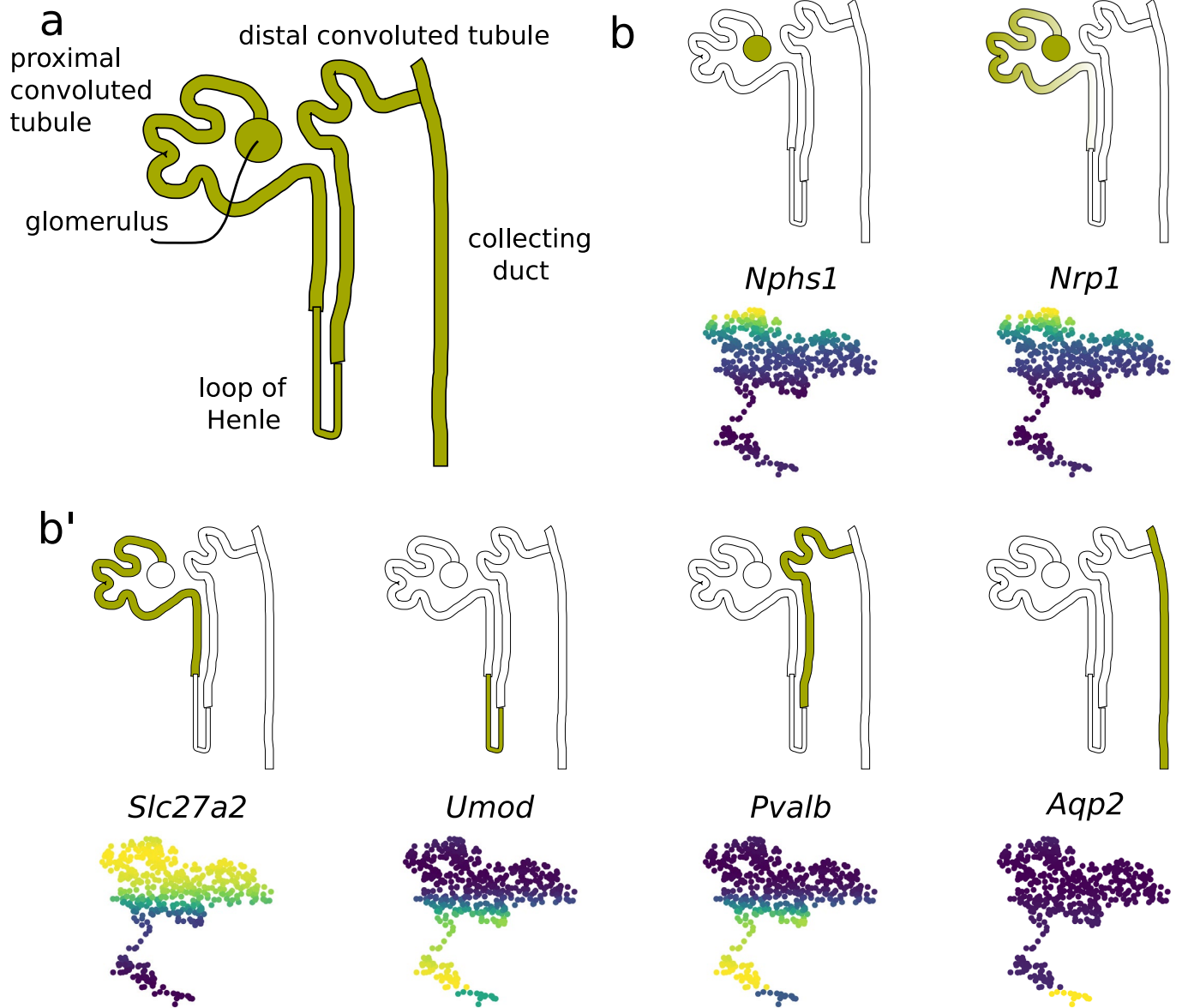
shown for each spatial archetype. novoSpaRc accurately groups genes expressed in a particular domain—for example, the subdomain of the mesoderm, which is characterized by the transcription factor *gcm* (Archetype 5)—whereas it does not capture the details of the fine expression patterns of pair-rule genes (Archetype 8). *CG42666* is also known as *prage*.



Extended Data Fig. 7 | novoSpaRc reconstructs the zebrafish embryo.
a, Histograms assessing the increase in the accuracy of novoSpaRc reconstruction (measured by the Pearson correlation with FISH data⁵) with increasing number of marker genes. **b**, novoSpaRc reconstructs patterns of gene expression in the zebrafish embryo on the basis of only 15 marker genes,

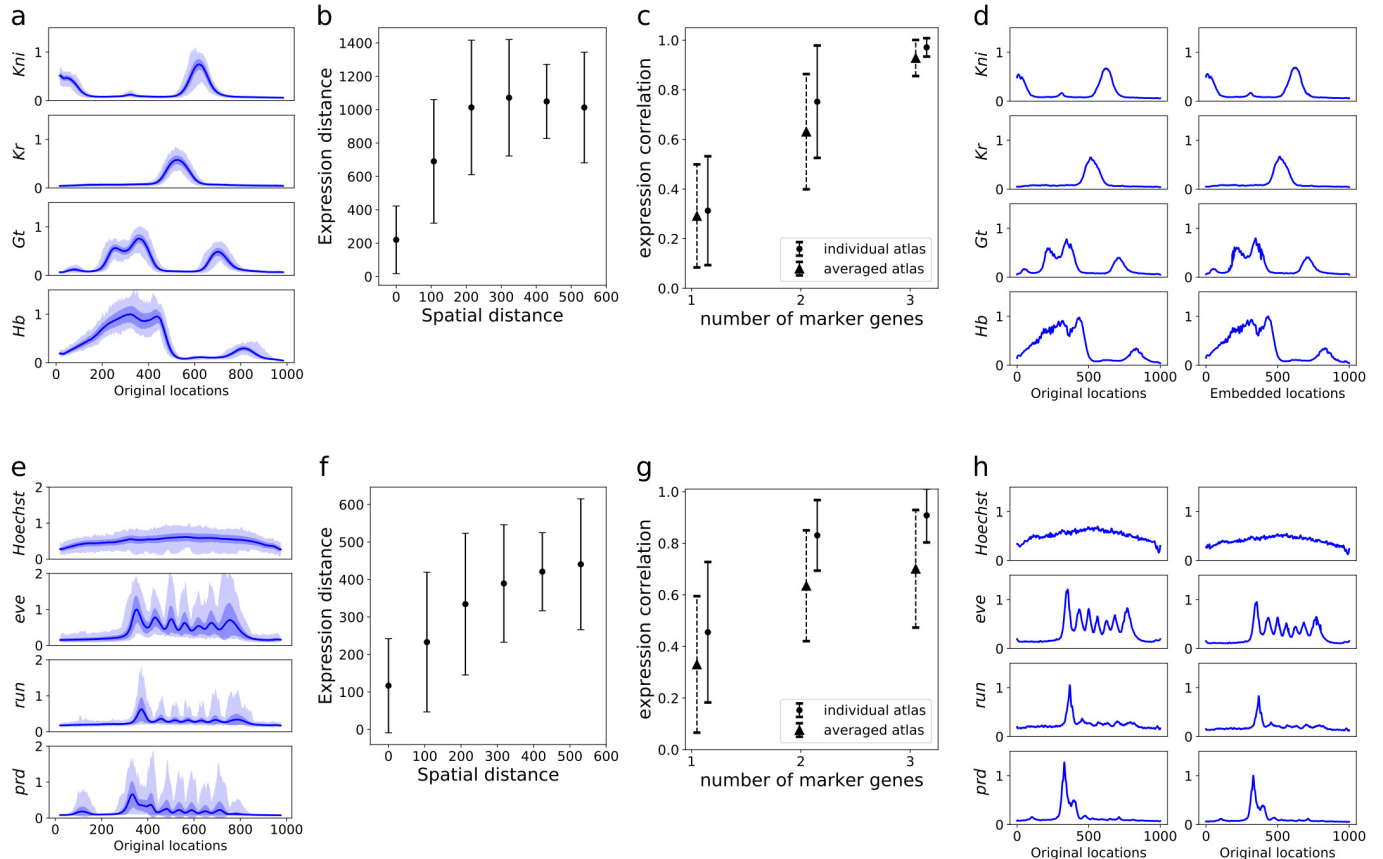
and the results improve as the number of marker genes increases. Top row, FISH data (reproduced from ref. ⁵); second row: Seurat predictions using 47 marker genes⁵; bottom three rows: novoSpaRc predictions using 15, 30 and 47 marker genes. The genes shown were not used in any of the reconstructions.

Article



Extended Data Fig. 8 | novoSpaRc reconstructs a whole-kidney dataset de novo. **a**, Sketch of the major cell types that are reconstructed with novoSpaRc. **b**, Representative marker genes for each of the cell types shown in **a**. Top rows depict a rough positioning for each cell type in yellow-green;

bottom rows show the gene expression predicted by novoSpaRc in the reconstructed tissue. *Nphs1*, podocytes; *Nrp1*, endothelial cells; *Slc27a2*, proximal tubule cells; *Umod*, loop of Henle; *Pvalb*, distal convoluted tubules; *Aqp2*, collecting duct cells. Expression ranges from low (blue) to high (yellow).

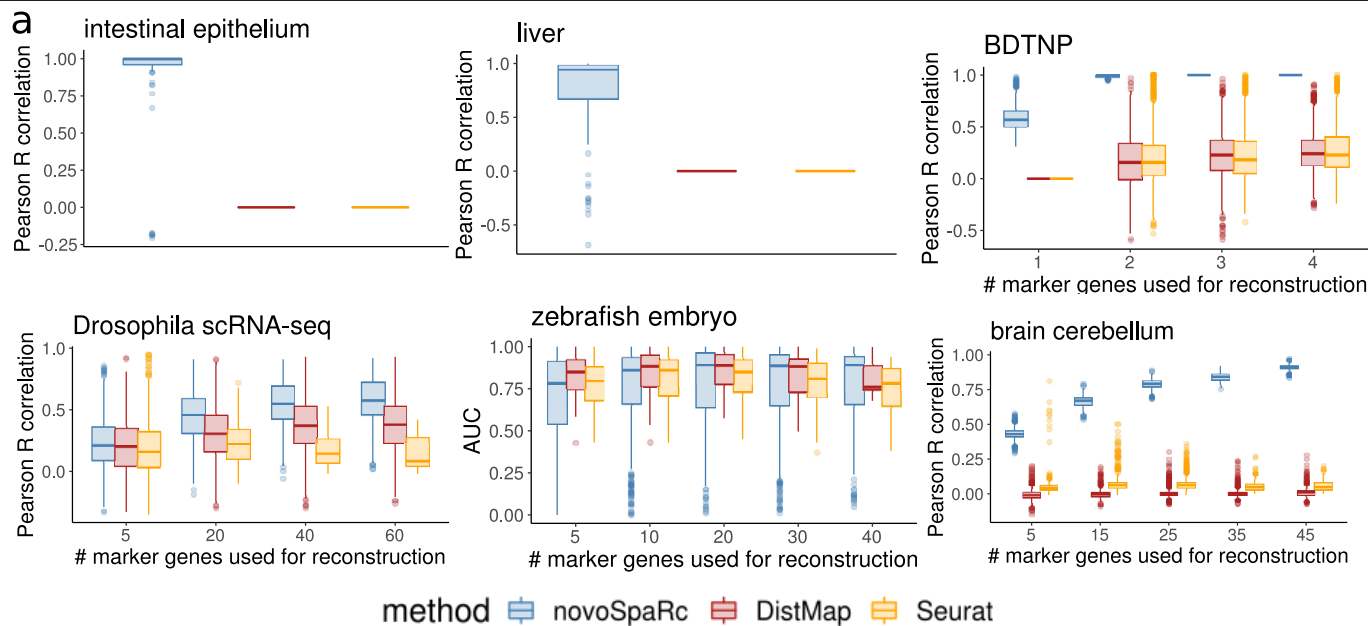


Extended Data Fig. 9 | NovoSpaRc reconstructs single *Drosophila* embryos.

a, e, The averaged original expression of four gap genes (**a**) and four pair-rule genes (**e**) is shown for 101 and 177 individual *Drosophila* embryos, respectively²². Solid line, mean; dark shadow, s.d.; light shadow, minimum and maximum values over all embryos. **b, f**, Demonstration of the monotonic relationship between cellular pairwise distances in expression and physical space, consistent with the structural correspondence assumption. Data are mean \pm s.d. **c, g**, The Pearson correlation increases with the number of marker genes used by novoSpaRc for the reconstruction of the remaining genes ($\alpha = 0.5$) for both gap genes (**c**) and pair-rule genes (**g**). Using a reference atlas that corresponds to the individual embryo being reconstructed ('individual

atlas') results in a consistently higher reconstruction quality than using an averaged reference atlas over all embryos ('averaged atlas'). Data are mean \pm s.d. **d, h**, Examples of the reconstruction of the expression patterns across a single random embryo, in which the reconstruction of each of the four genes is performed using the three complement genes as a reference, for both gap genes (**d**) and pair-rule genes (**h**). Note that the reconstructed expression patterns presented in **d, h** were computed while the corresponding gene in each case was not used for the reconstruction. The expression level of each gene in **a, d, e, h** is normalized to the maximum value over the mean expression of all embryos.

Article



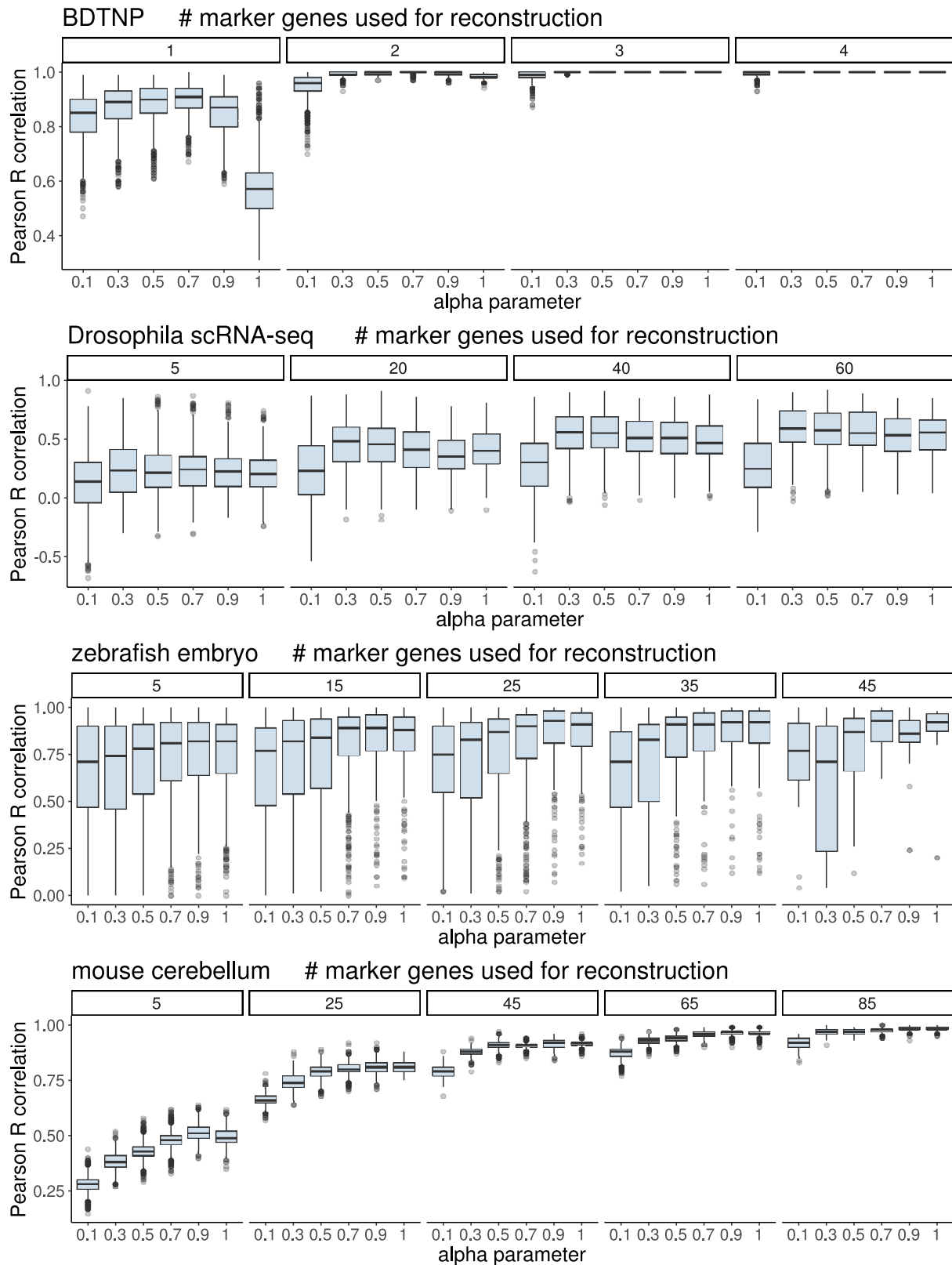
b

	Seurat	DistMap	novoSpaRc
Spatial mapping with reference atlas	✓	✓	✓
Reference atlas can have continuous values	X	X	✓
Spatial mapping <i>de novo</i>	X	X	✓
Does not require predetermined shape	✓	✓	X
Can exploit structural information	X	X	✓
Can use continuous expression data	X	X	✓
Can be applied to complex tissues	X	✓	✓
Does not require data imputation	X	✓	✓
Does not require a threshold	✓	X	✓

Extended Data Fig. 10 | Comparison of spatial reconstruction with novoSpaRc versus available methods that fully rely on a reference atlas.

a, The Pearson correlation of the predicted versus the original spatial gene expression is shown as a function of the top 100 highly variable genes for the intestinal epithelium and liver datasets, or the number of marker genes used for the reconstruction for the BDTNP dataset, the *Drosophila* and zebrafish embryos and the brain cerebellum (84, 84, 45 and 745 genes, respectively). For the 1D datasets, the reconstructions are done *de novo* (with no reference atlas) and the existing baseline methods are inapplicable. For the liver, the last lobule layer was removed from the analysis, as only five cells were associated with it. For the 2D datasets, correlations are computed only for genes that were not

used for the reconstructions. Note that for the *Drosophila* embryo novoSpaRc outperforms DistMap¹², and for the zebrafish embryo novoSpaRc performs comparably to or better than Seurat⁵—although those methods were developed and tailored for the *Drosophila* and zebrafish embryos, respectively, and the best-performing threshold was chosen for DistMap. For the box plots, the centre line is the median, box limits are the 0.25 and 0.75 quantiles and whiskers extend to ± 2.698 s.d. For the BDTNP dataset, the *Drosophila* and zebrafish embryos and the brain cerebellum, the results are shown for 100 random choices of marker genes. **b**, The intrinsic characteristics of novoSpaRc compared against Seurat⁵ and DistMap¹².



Extended Data Fig. 11 | Reconstruction quality varies with the α parameter. Reconstructions of the BDTNP dataset, the *Drosophila* and zebrafish embryos and the brain cerebellum, with varying numbers of marker genes used for the reconstruction and different values of the α parameter. The reconstruction quality is quantified by calculating Pearson correlations between the predicted and the original patterns of gene expression for all genes that were not used as markers for the reconstruction. The quality of the reconstruction decreases for $\alpha=1$ in the BDTNP and brain cerebellum cases, which corresponds to

reconstructing based only on reference marker genes, without taking the structural correspondence assumption into account. We note that α is an interpolation parameter (defined in the Methods section 'Mathematical formulation of novoSpaRc') between using only a reference atlas ($\alpha=1$) and using only structural information (driven by the structural correspondence assumption) ($\alpha=0$). For the box plots, the centre line is the median, box limits are the 0.25 and 0.75 quantiles and whiskers extend to ± 2.698 s.d. Results are shown for 100 random choices of marker genes.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

No software was used for data collection. All data shown in the manuscript is already publicly available.

Data analysis

We wrote custom software code which is available online on Github (distributed under the MIT License, version 0.2.2, <https://github.com/rajewsky-lab/novosparc>). The code is written in python and uses commonly used python libraries (numpy, matplotlib, sklearn, scipy, ot). To calculate spatial autocorrelation we used the implementation of PySAL (version 2.0.0), a Python spatial analysis library. We also used an implementation of the Gromov-Wasserstein transport method by Erwan Vautier (distributed under the MIT License).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

No datasets were generated during the current study. The single cell datasets analyzed for the current study were acquired from the GEO database with the following GEO accession numbers: GSE99457 for the intestinal epithelium, GSE84490 for the liver, GSE95025 for the Drosophila embryo, GSE66688 for the zebrafish embryo and GSE107585 for the kidney. The cerebellum Slide-seq datasets were acquired from the Broad Institute Single Cell Portal (https://portals.broadinstitute.org/single_cell/study/slide-seq-study). The individual Drosophila embryos dataset (Petkova, M.D., et al., Cell 2019) is available as Supplemental Information files of the original manuscript. The BDTNP dataset was downloaded directly from the BDTNP webpage (<http://bdtnp.lbl.gov>).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes for Fig. 3c & Ext. Data Figs. 2c 6a-g, 11b,c, 17,18 were based on 100 instantiations, and for Figs. 2b,f, 3b, 5b & Ext. Data Figs. 10b, 16b,c,f,g there was no subsampling of the data.
Data exclusions	No data were excluded
Replication	Experimental replication was not attempted and is not applicable to this study
Randomization	Since the single cell transcriptomes are unique and technically not reproducible, randomization was not applicable to the study
Blinding	No datasets were generated during the current study, and therefore blinding was not applicable.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging