Gene expression cartography

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

Massively multiplexed sequencing of RNA in individual cells is transforming basic and clinical life sciences[1-4]. In standard experiments, however, tissues must be first dissociated. Thus, crucial information about spatial relationships between cells, along with the tissue-wide expression patterns they confer, is lost. This poses a fundamental problem for elucidating collective function of tissues, developmental pathways, and mechanisms of cell-to-cell communication[5, 6]. Considerable efforts to overcome this challenge have been undertaken. However, experimental methods are either technically challenging, or have limited resolution or throughput[5, 7, 8]. Existing computational approaches predict spatial positions by comparing each sequenced cell, independently, to an imaging-derived spatial gene expression database for that tissue [9, 10]. However, these approaches rely on prior knowledge of spatial expression patterns which often does not exist, or is difficult to construct. Here, we explore a radically different idea. We postulate that cells in spatial proximity, overall, share more similar transcriptional profiles than cells farther apart. We validate this hypothesis for several complex biological systems. Consequently, we seek to find spatial arrangements of sequenced cells on tissue space which optimally preserve this principle. We show that this hard optimization problem can be cast as a generalized optimal transport problem for probabilistic embedding, for which we derived an efficient iterative algorithm. We successfully reconstruct the mammalian liver, intestinal epithelium, fly and zebrafish embryos, cerebellum sections and kidney. We then use the reconstructed tissues to infer spatially informative genes directly from single cell data. Our results demonstrate that we have identified a spatial expression organization principle in animal tissues which can be used to infer meaningful spatial position probabilities for individual cells. Our framework ("novoSpaRc") is flexible, can naturally incorporate prior spatial information, is scalable to large number of cells and compatible with any single-cell technology. We envision that novoSpaRc can be valuable in collaborative efforts to characterize various tissues[11, 12], and that additional or generalized principles underlying spatial organization of gene expression can be formulated and tested using our approach.

Single-cell transcriptome sequencing (scRNA-seq) has revolutionized our understanding of the rich heterogeneous cellular populations that compose tissues, the dynamics of developmental processes, and the underlying regulatory mechanisms that control cellular function[1-4]. However, to understand how

[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

single cells orchestrate multi-cellular 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 since tissues must commonly be dissociated into single cells prior to scRNA-seq. Thus, the original spatial context and relationships between cells are lost. Two seminal papers tackled this problem computationally[9, 10], the key idea being to use a reference atlas of informative marker genes as a guide to assign spatial coordinates to sequenced cells. This scheme was successfully employed in various tissues[13-17], including the complete early *Drosophila* embryo[18]. However, such methodologies heavily rely 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 unique 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 the single cell data itself.

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 normal puzzle, here we typically do not know the image that we want to reconstruct. While 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 which are specifically expressed in spatially contiguous territories and thus consistent with only a small subset of all possible arrangements. Thus, we set out to identify simple, testable assumptions which govern how gene expression is organized in space, and to subsequently find the arrangements of cells that best respects those assumptions.

Here, we specifically explore the assumption that cells which are physically close tend to share similar transcription profiles, and vice versa (Extended Data Fig. 1, Supplementary Note). Biologically, this phenotype can result from multiple mechanisms, such as gradients of oxygen, morphogens and nutrients, trajectory of cell maturation, and communication between neighboring cells. We stress that this is an assumption about overall gene expression across the entire tissue – not about individual genes and not about all physically close cells (Supplementary Note). Here, we show that on average, the distance between cells in expression space indeed increases with their physical distance, for diverse tissues in matured organisms or whole embryos in early development. Thus, to predict spatial locations of sequenced cells, we seek to find a map of sequenced cells to tissue space (“cartography”) such that overall *structural correspondence* is preserved, meaning that cells have similar 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 *Drosophila* embryo during late stage 5 of development [19]) or approximated by a grid (Supplementary Note). The distances are computed for each pair of cells across graphs constructed over the two spaces (Extended Data Fig. 1, Supplementary Note). Then, novoSpaRc optimally aligns 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 available (Methods, Supplementary Note). For both biologically- and computationally-motivated reasons, we seek a probabilistic mapping which assigns each cell a distribution over locations on the physical space (Supplementary Note). We formulate this as a generalized optimal transport problem[20-22], which has been proven to be increasingly valuable for diverse fields, including biology[23, 24], and renders the reconstruction task feasible for large datasets. Specifically, we formulate an interpolation between entropically regularized Gromov-Wasserstein[25, 26] and optimal transport[27] objectives, serving to satisfy the structural correspondence assumption between gene expression space and physical space, and to match available prior knowledge, respectively (Methods). We show this optimization problem can be efficiently solved using projected gradient descent, reduced to iterations of linear optimal transport sub-problems (Supplementary Note).

To systematically assess novoSpaRc's performance, we employed a simple generative model of spatial gene expression (Methods). As expected, reconstruction quality gradually increased with decreasing tissue dimensions, increasing signal to noise ratio of the expression levels, increasing number of marker genes used as a reference atlas, and increasing fraction of spatially informative genes (Methods, Extended Data Fig. 2). In addition, reconstruction quality peaked when combining both structural (driven by the structural correspondence assumption) and atlas-based (marker gene) information (Extended Data Fig. 2). This conclusion was further supported by the reconstruction results for the BDTNP and brain cerebellum datasets discussed below.

Focusing on real single-cell datasets, we first *de novo* reconstructed tissues with inherent symmetries which render them effectively 1-dimensional, such as the mammalian intestinal epithelium[16] and the liver lobules[13]. Schematic figures of the reconstruction process are shown in Figs. 2a and 2e respectively. For both tissues, cells were previously classified into distinct zones, or layers, based on robust marker gene information (7 zones for the intestinal tissue[16], 9 layers for the liver[13]). We found that the average pairwise distances between cells in expression space increased monotonically with the pairwise distances in physical 1-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, Extended Data Figs. 3,4, Methods). Median Pearson correlation of reconstructed expression patterns to original patterns for the top 100 variable genes was 0.99 and 0.94 for intestine and liver, respectively (Methods). The fraction of cells correctly assigned up to one layer away from their original layer was 0.98 and 0.73 for intestine and liver, respectively (Methods, Extended Data Fig. 3). novoSpaRc captured spatial expression patterns of the top zonated genes (Methods, Extended Data Fig. 3) and spatial division of labor within the intestine epithelium (Fig. 2d), as well as within the layers of the liver lobules (Fig. 2h, Extended Data Figs. 3,4), where cells in different tissue layers perform different tasks and exhibit different expression profiles. For the intestinal epithelium data, varying the grid resolution to include either less or more embedded zones did not seem to compromise the quality of the reconstructed expression patterns (Extended Data Fig. 5) and shows the potential for increased resolution of single cell embedding relative to atlas-based embedding. We recovered the observed ordering of the peaks of expression along the intestinal villi of groups of genes that play important roles in the absorption and transportation of different nutrient groups, including apolipoproteins cholesterol, peptides, carbohydrates and amino acids (Fig. 2d, Supplementary Note). Similarly, spatial expression patterns of genes in the liver exhibiting pericentral, periportal or non-monotonic profiles were correctly identified (Fig. 2h, Extended Data Fig. 3).

Next, we focused on spatially reconstructing the well-studied *Drosophila* embryo, as a more challenging, higher dimensional tissue. At late stage 5, the fly embryo consists of ~6,000 cells. It has been previously suggested [28] that at early stages of the fly development, the expression levels of gap genes can be optimally decoded into positional information. The expression levels of 84 transcription factors were registered using fluorescence *in situ* hybridization (FISH) for each of the cells in a highly quantitative manner by the Berkeley Drosophila Transcription Network Project (BDTNP)[19].

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 informational marker genes (Fig. 3a). Similar to the 1D datasets, we found a monotonically increasing relationship between the cell-cell pairwise distance in expression space and in physical space (Fig. 3b), confirming that the data adheres to our structural correspondence assumption.

The reconstructed spatial gene expression patterns highly correlated with the original ones (Fig. 3c, Methods). We found that employing novoSpaRc using both structural and marker gene information outperformed the reconstruction based on only the latter, and performance was saturated at 2 marker genes (Fig. 3c). As expected, reconstruction quality increased with the number of genes used to provide structural information in expression space, and with the fraction of spatially-informative genes (Methods, Extended Data Fig. 6). The majority of spatial patterns were recapitulated faithfully, even when only a single marker gene was used (Fig. 3d). We observed that novoSpaRc reconstructed the patterns robustly and independently of the marker genes used (Fig. 3c). In addition, novoSpaRc identified the physical neighborhoods that single cells originated from when used *de novo* (up to inherent symmetries, Supplementary Note), and pinpointed their true locations ($p<0.05$ compared to random assignment) when a handful of marker genes were used (Extended Data Fig. 7).

We examined the expression patterns of four transcription factors spanning the dorsal-ventral and anterior-posterior axes in detail (Fig. 3e). Reconstruction quality improved when employing the structural correspondence assumption (Extended Data Fig. 8). The *de novo* reconstruction correctly identified both axes of the embryo. The reconstructed portrait was remarkably similar to the original one (Fig. 3e, Extended Data Fig. 9). Generally, since *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, translations) relative to the original configuration along the two major axes of the embryo (Supplementary Note). Consequently, the resulting gene expression patterns might be shifted or flipped relative to the expected ones. However, there are features of a faithful reconstruction we can test for, such that the reconstruction would be robust to small changes in the optimization parameters (Extended Data Fig. 10) and that the embedding of single cells onto the embryo would be relatively localized, as we would expect for a biologically-meaningful embedding (Fig. 3f). This means that the distribution over locations that each single cell is assigned should be localized, and indeed, the mean standard deviation of that distribution for all single cells is significantly lower than that of a randomized embedding (Extended Data Fig. 10). Furthermore, we demonstrated that novoSpaRc's results, as measured by correlation to observed imaging data and optimization error, were robust to optimization parameters and diverse noise sources, including partially sampling cells, additive expression noise, and dropouts (Extended Data. Fig. 6).

As an intermediate step bridging the BDTNP dataset and a raw scRNA-seq dataset, we applied novoSpaRc to spatially reconstruct the *in silico* virtual *Drosophila* embryo[18] (Methods), quantifying the expression of ~8,000 genes in each of the single cells (Extended Data Fig. 11). novoSpaRc successfully reconstructed the virtual embryo, with the accuracy increasing with the number of marker genes used for reconstruction (Extended Data Fig. 11, Methods). To assess the performance of reconstruction in cases where no ground truth expression patterns are available, we show that intra-correlation between virtual embryos reconstructed by using different sets of marker genes reflected successful reconstruction and increased with the number of marker genes used (Extended Data Fig. 11).

We next employed novoSpaRc to reconstruct the stage 6 *Drosophila* embryo by using a scRNA-seq dataset[18] (Fig. 4a). In that work, 84 marker genes were required for reconstruction that distributed 1,297 single cells over 3,039 embryonic locations. Since novoSpaRc naturally exhibits a probabilistic mapping, we reasoned that the above dataset is a good candidate for testing its efficacy. When using both structural information and the reference atlas, the accuracy of reconstruction by novoSpaRc increased with the number of marker genes, reaching high correlation (Pearson correlation coefficient: 0.74) with the FISH data (Fig. 4b, Extended Data Fig. 12, Methods). The *de novo*, atlas-free reconstruction by novoSpaRc accurately separated the major post-gastrulation spatial domains (mesoderm, neurogenic ectoderm, 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, termed *archetype* (Methods, Supplementary File). novoSpaRc identified numerous distinct spatial archetypes (Fig. 4c,d, Extended Data Fig. 13). We compared representative genes of each spatial archetype with FISH images to visually assess the accuracy of the spatial reconstruction. Gene patterns expressed through the anterior-posterior or the dorsal-ventral axis were largely recapitulated: typical mesoderm genes, such as *twi* and *sna*, were co-localized ventrally (Fig. 4c,d, right), while typical dorsal ectoderm genes, such as *zen* and *ush,* were co-localized dorsally (Fig. 4c,d, middle). novoSpaRc accurately captured localized spatial populations (Fig. 4c,d, left, Extended Data Fig. 13, archetype 5), while less extensive spatial domains were reconstructed with diverse degrees of accuracy (Extended Data Fig. 13). Note that within the *de novo* reconstruction, accurate localization entails global transformations as described above. This is mostly evident for archetype 5 (Extended Data. Fig. 13, see also Supplementary Note).

Before proceeding to more complex tissues, we reconstructed the zebrafish embryo dataset [9] (Fig. 4e). Similar to the original seminal study, we mapped the cells onto the surface of a hemisphere constituting of 64 distinct locations. The resulting spatial expression patterns were highly correlated to the experimentally verified ones and novoSpaRc reconstructed the zebrafish embryo by using only 15 marker genes, in contrast to the 47 genes previously required[9] (Extended Data Fig. 14, Methods). The accuracy of the reconstruction increased with the number of marker genes (Extended Data Fig. 14). Furthermore, no data imputation or other specialized preprocessing was necessary as before[9].

To further showcase the applicability of novoSpaRc to complex tissues, diverse sequencing technologies and different organisms, we used it to reconstruct slices of brain cerebellum [29] (Fig. 5), the mammalian kidney [30] (Extended Data Fig. 15), and a dataset of hundreds of individual *Drosophila* embryos [31] (Extended Data Fig. 16).

The adult mammalian brain is a well-studied, highly differentiated and complex tissue. To benchmark novoSpaRc's capabilities in reconstructing complex tissues, we used murine cerebellum slices from a recently developed spatial transcriptomics technology [29]. The sagittal section dataset contained 46,376 locations with a median of 52 quantified transcripts per location. To ensure that enough information is available to novoSpaRc, we first coarse-grained the data by binning neighboring locations. This resulted after quality filtering in 7,704 locations with a median of 379 quantified transcripts (Methods, Fig. 5a). novoSpaRc successfully reconstructed the whole transcriptome, with the Pearson correlation over all 15,878 genes equal to 0.5 when using only 15 marker genes and increasing to 0.94 when using 50 marker genes (Fig. 5b, Methods). Spatial expression patterns start to emerge when using only a handful of marker genes. For example, spatial positions of Purkinje cells were revealed by reconstructing with only 5 marker genes (excluding all genes exhibiting a Pearson correlation with *Pcp4* of 0.25 or higher) and the signal improved dramatically by including more markers (Fig. 5c). The reconstructed cerebellum slices illustrated great 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 cerebellum, also published in [29], with similarly successful results (Fig. 5e).

Next, we used novoSpaRc to spatially reconstruct a single-cell dataset from whole-kidney [30], which is a complex tissue with stereotypical organization. As no reference atlas of gene expression was available in this case, the reconstruction was performed *de novo*. We focused on six major cell types within the kidney (Extended Data Fig. 15) and mapped the cells onto a 2-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. 15). We note that, since no prior information was required for this reconstruction, this case demonstrates the applicability of novoSpaRc to a wide variety of medically-relevant tissues.

Finally, to show that novoSpaRc can reconstruct individual samples and not only a prototypical tissue, we used a dataset that captures expression patterns in hundreds of individual *Drosophila* embryos [31]. In that case, 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 patterns. novoSpaRc was able to predict expression patterns based on a limited reference atlas (Extended Data Fig. 16). 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. 16), yet reached high correlation values for both (median Pearson correlation for reconstructing a fourth gene based on the three remaining genes were 0.99 (0.95) 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 (Extended Data Fig. 17), as well as extensively benchmarked novoSpaRc's performance when comparing to available reconstruction methods that fully rely on a reference atlas (Seurat[9] and DistMap[18]). novoSpaRc possesses several advantages when compared to the other existing methods (Extended Table 1, Methods) and shows overall substantial benefits in reconstruction performance (Extended Data Fig. 18).

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 zonated genes without a reference atlas (Methods, Supplementary File), and found that the top inferred zonated genes were indeed supported experimentally and/or computationally (Fig. 6a,b, Extended Tables 2, 3). Gene ontology (GO) enrichment analysis [32] further revealed zonation-compatible biological processes enriched for different domains in the intestine and the liver, reconstructed by novoSpaRc (Methods, Supplementary Note, Supplementary Files). For the *Drosophila* single cell dataset we ranked all 8924 genes according to their spatially informative rank (Methods, Fig. 6c, Supplementary File), and found that transcription factors were, as known from classic genetics [33], among the most highly informative genes (Fig. 6c). In addition, novoSpaRc identifies numerous lncRNAs and TFs as being spatially highly informative, many of them having been already predicted in [18]. Finally, we ranked all 15,878 genes in the cerebellum by their spatially informative rank (Methods, Fig. 6d, Supplementary File), and found that well-known marker genes with defined spatial expression pattern are indeed among the highest ranking spatially informative genes (Fig. 6d).

Taken together, we have demonstrated here that novoSpaRc can spatially reconstruct a diverse number of biological tissues, based on a simple hypothesis about how gene expression is organized in space - a structural correspondence between distances of cells in expression space and in physical space, and 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 conjure that within certain biological contexts, different cell types may require higher-order interactions or exhibit different spatial organization principles. In this context, it is important to stress that because of the availability of general mathematical results in optimal transport theory, our framework is versatile and can support a large variety of alternative ways to compare distances in expression and physical space by varying the optimization loss functions (Methods, Supplementary Note). Such alternative schemes are currently not supported by novoSpaRc, but can be implemented.

Our data analyses and the success of the reconstructions by novoSpaRc suggest that we have identified a general organization principle for how gene expression is organized in tissue space. It will be interesting to find tissues in which this organization principle is weak or not valid. However, we are almost certainly underestimating the strength of the structural correspondence principle 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 control of spatial tissue features and functions. We believe that we have demonstrated that we can systematically identify at least a subset of these genes directly from the single-cell data. In the future, we will extend these analyses to identify genes predicted to functionally interact in space. Finally, our developed framework can be flexibly extended beyond spatial reconstruction. We are currently utilizing it to recover different types of biological signals such as temporal progression on short (e.g. cell cycle) and long (e.g. developmental) scales.

Code availability A python package for novoSpaRc, as well as scripts reconstructing selected tissues presented in the manuscript, are provided at https://github.com/rajewsky-lab/novosparc.

Acknowledgements We thank all members of our labs and many colleagues in the field for valuable comments. This work was supported by the Israeli Science Foundation, through the I-CORE program (NF) and an Alexander von Humboldt Foundation Research Award (NF). NK was supported by the DFG Leibniz prize (NR) and grants DFG RA 838/8-2, HGF ExNet-0036 and DFG KA 5006/1-1. MN was supported by the James S. McDonnell Foundation, Schmidt Futures, Israel Council for Higher Education, and the John Harvard Distinguished Science Fellows Program within the FAS Division of Science of Harvard University.

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

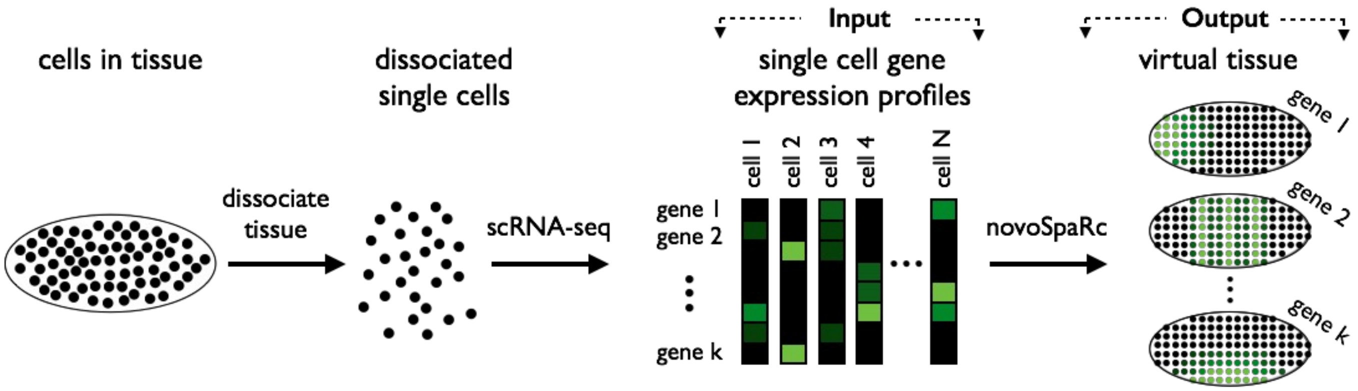


Figure 1 | Overview of novoSpaRc. A matrix containing single-cell transcriptome profiles, sequenced from dissociated cells, is the main input for novoSpaRc. The output is a virtual tissue of chosen shape which can be queried for the expression of all genes quantified in the data.

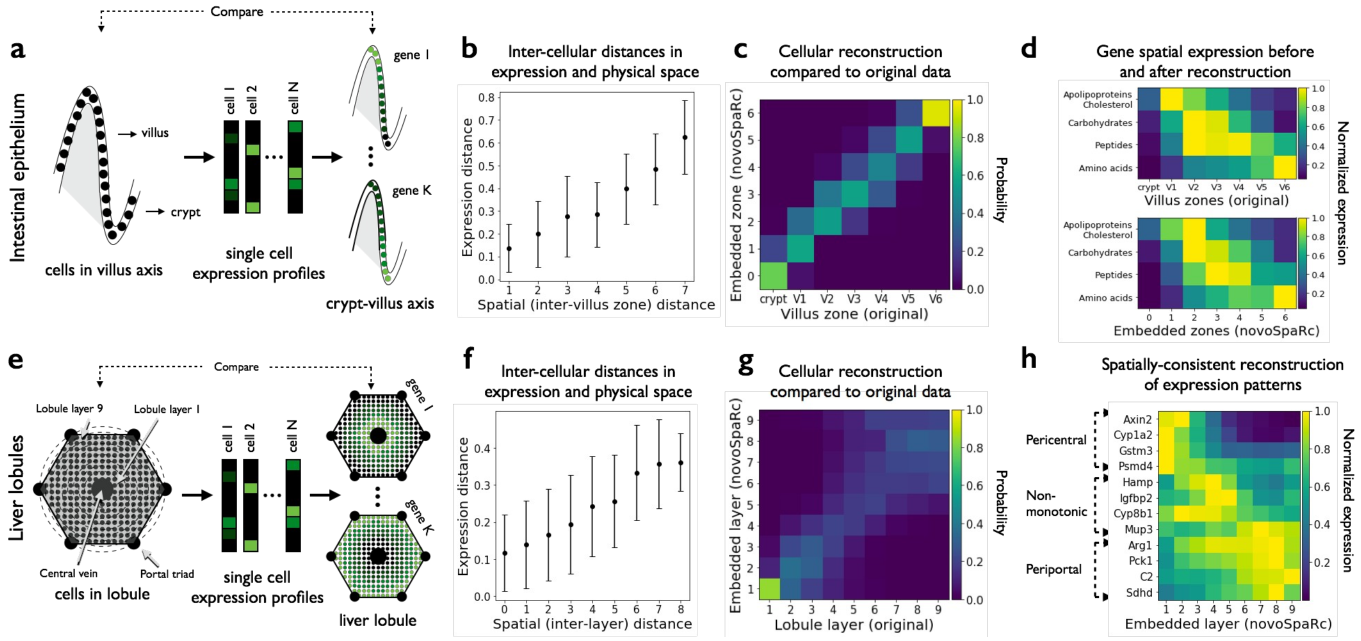


Figure 2 | novoSpaRc successfully reconstructs complex tissues with effective 1D structure *de novo*. **a**, **e**, The reconstruction scheme for the mammalian intestinal epithelium and liver lobules respectively. **b**, **f**, Demonstration of the monotonic relationship between cellular pairwise distances in expression and physical space. Center point, mean; error bars, SD. **c**, **g**, novoSpaRc infers the original spatial context of single cells with high accuracy. Heatmaps show the inferred distribution over embedded layers (rows) for the cells in each of the original layers (columns). **d**, novoSpaRc captures the spatial division of labor of averaged expression of genes that play a role in the absorption of different nutrient classes in the intestine. **h**, novoSpaRc captures spatial expression patterns (pericentral, periportal and non-monotonic) at single-cell resolution in the liver. The expression level of each gene in both (d) and (h) is normalized to its maximum value.

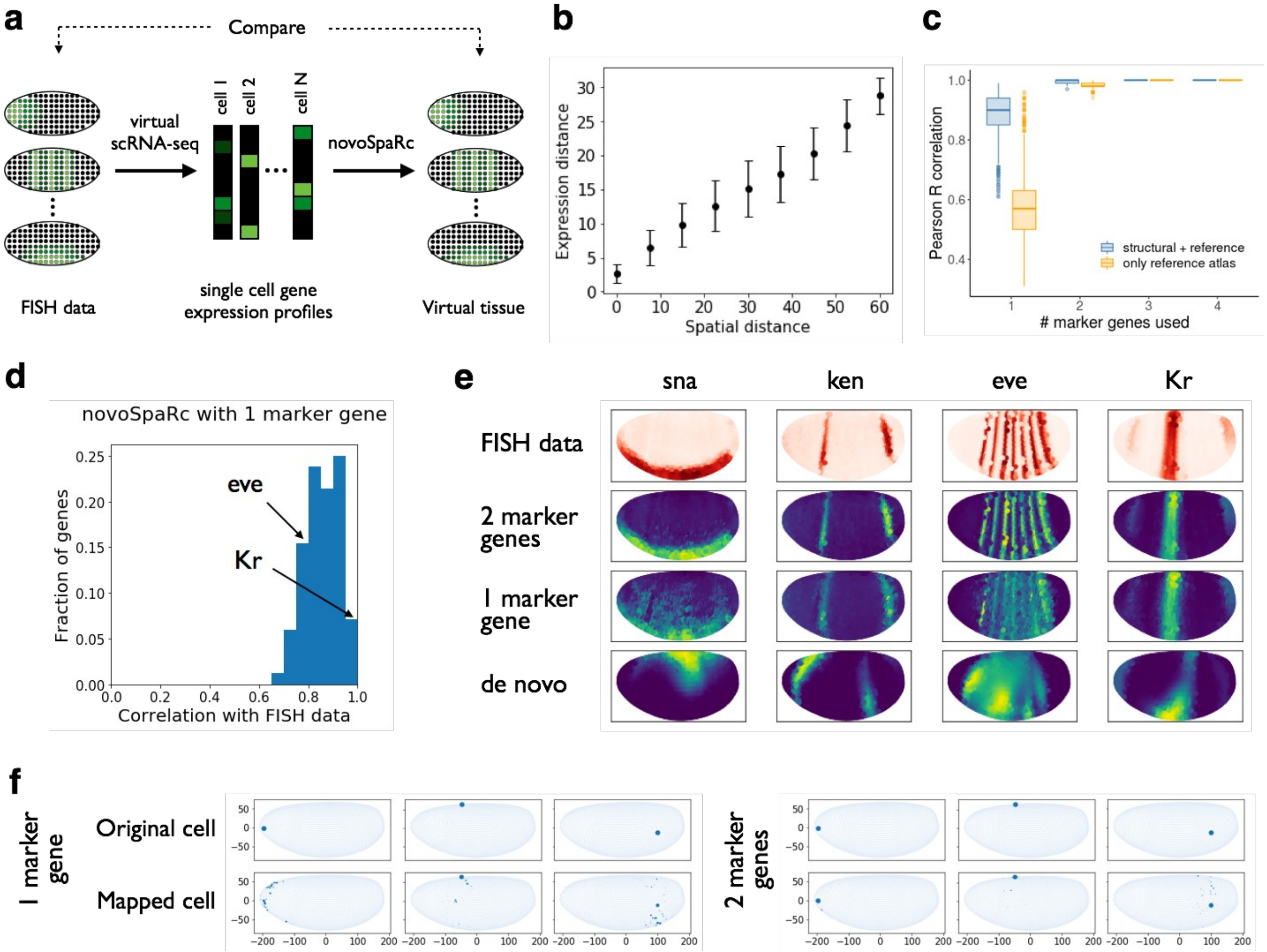


Figure 3 | novoSpaRc accurately reconstructs the *Drosophila* embryo based on the BDTNP dataset[19]. a, FISH data is used to create virtual scRNA-seq data, which novoSpaRc then inputs to reconstruct a virtual embryo. **b**, Demonstration of the structural correspondence hypothesis. Pairwise cellular distances in expression space increase monotonically with distances in physical space. Center point, mean; error bars, SD **c**, novoSpaRc spatially reconstructs the *Drosophila* embryo with only a handful of marker genes. The quality of reconstruction (as measured by Pearson correlation with FISH data) increases with the number of marker genes and saturates at perfect reconstruction at 2 marker genes, when using both structural information (driven by the structural correspondence assumption) and marker gene information (black line, 'structural + reference'). This outperforms reconstruction that relies only on marker gene information (dotted line, 'only reference atlas'). Results are averaged for 100 different marker gene combinations. Center line: median; whiskers: +/-2.698SD. **d**, Distribution of gene-specific coefficients of correlation with the FISH data, from an instance of novoSpaRc reconstruction using 1 marker gene. Lower correlation values correspond to finer expression patterns. **e**, Visualization of reconstruction results for 4 transcription factors. The original FISH data (first row) is compared to reconstruction by novoSpaRc that exploits both structural and marker gene information (using 2 and 1 marker genes) and reconstruction without any marker gene information (*de novo*). **f**, The original locations of three cells are compared to their respective reconstructed locations by novoSpaRc (using 2 and marker genes). The expression patterns of the 2 and 1 marker genes used for the results shown in panels d-f are shown in Extended Data Fig. 7c.

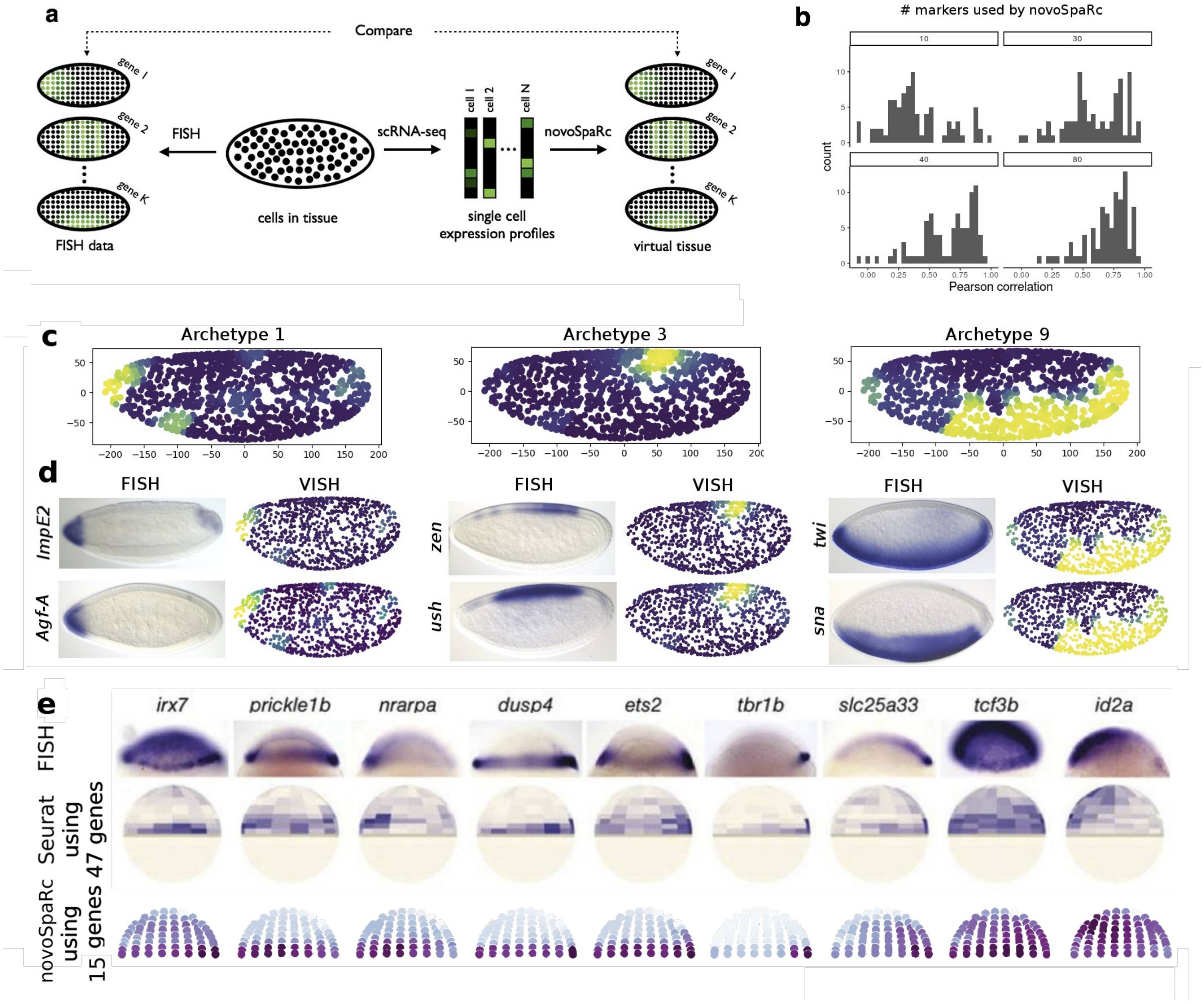


Figure 4 | novoSpaRc identifies spatial archetypes in the *Drosophila* embryo by using scRNA-seq data and successfully reconstructs the zebrafish embryo. **a**, Schematic overview. The expression patterns as reconstructed by novoSpaRc are compared with the BDTNP expression values. **b**, Reconstruction 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 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 novoSpaRc predictions (right columns). **e**, novoSpaRc reconstructs gene expression patterns in the zebrafish embryo by using only 15 marker genes, and the results improve with increasing number of marker genes (Extended Data Fig. 14). Genes shown were not used in any reconstruction. Top row: FISH data[9]; second row: Seurat predictions by using 47 marker genes[9]; third row: novoSpaRc predictions by using 15 marker genes.

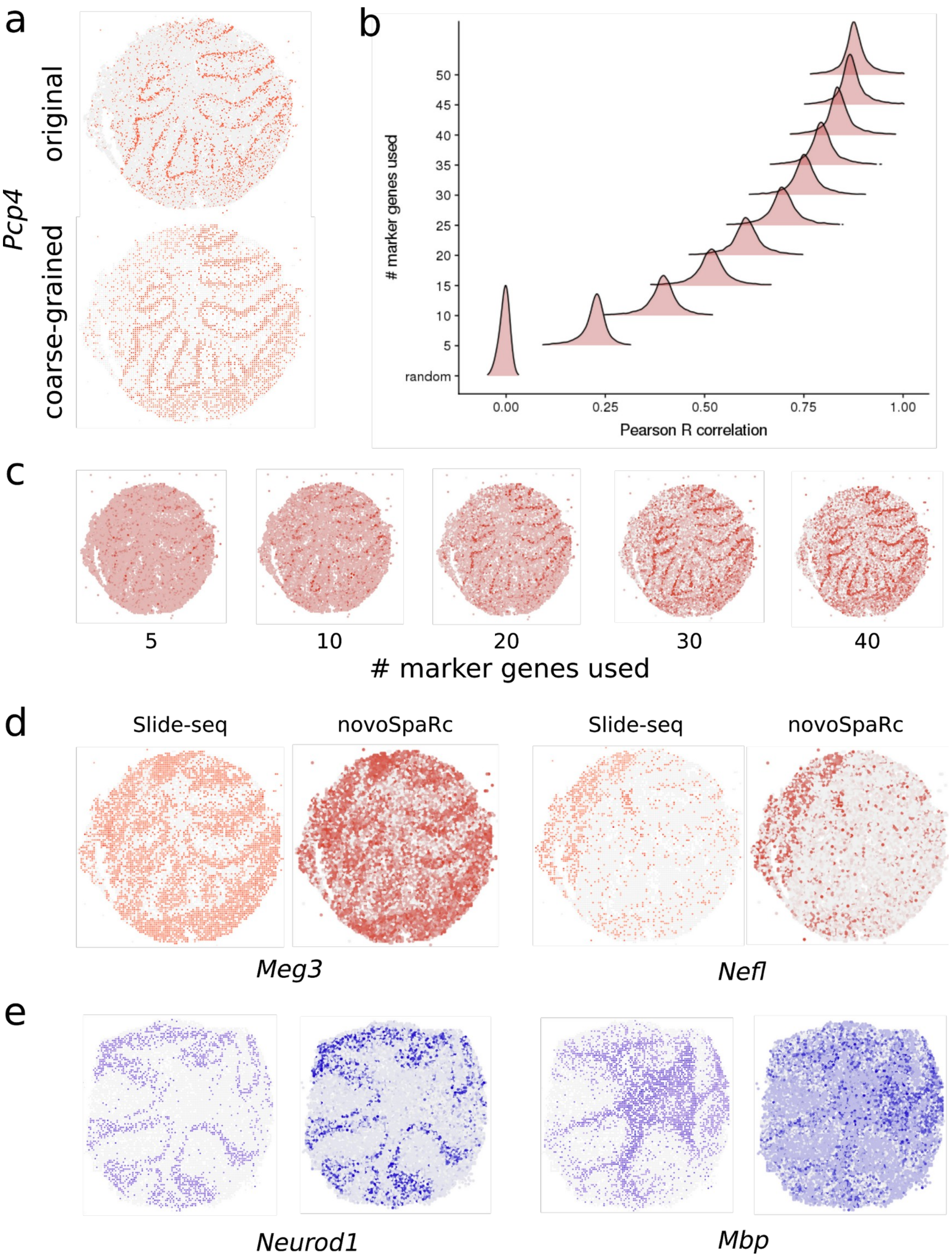


Figure 5 | novoSpaRc reconstructs mouse cerebellum tissue. **a,** The original and the coarse-grained spatial expression of a Purkinje cells marker (*Pcp4*) in a sagittal cerebellum section from direct spatial RNA sequencing [29]. **b,** The overall Pearson correlation between original and novoSpaRc predicted gene expression increases drastically by using more marker genes. With only 5 marker genes, the correlation is already 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* signal is visible with only 5 marker genes and increases as more markers are included 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 cerebellum section stemming from [29].

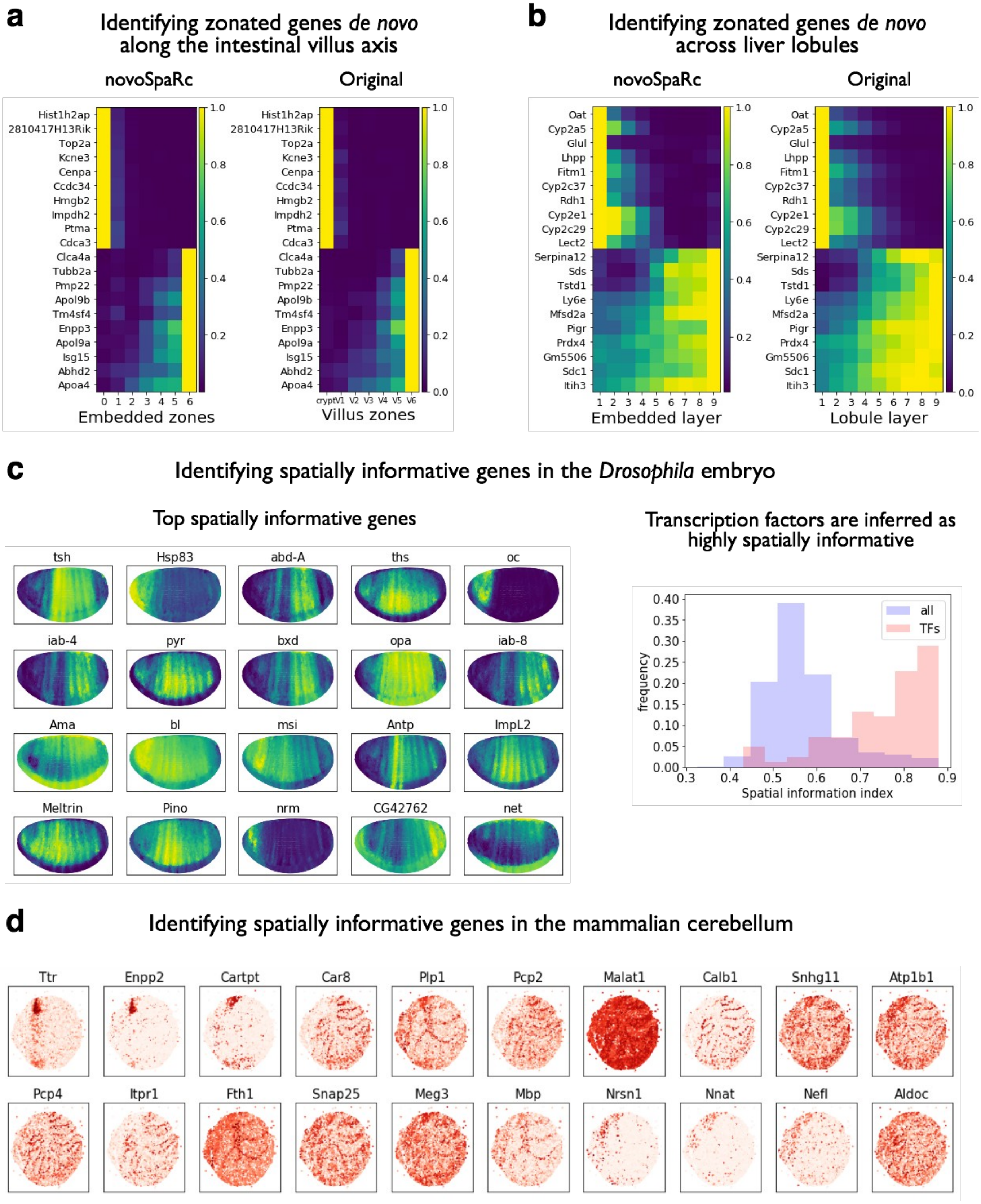

a
Identifying zonated genes *de novo* along the intestinal villus axis
novoSpaRc
Original
Hist1h2ap
2810417H13Rik
Top2a
Kcne3
Cenpa
Ccdc34
Hmgb2
Impdh2
Ptma
Cdca3
Clca4a
Tubb2a
Pmp22
Apol9b
Tm4sf4
Enpp3
Apol9a
Isg15
Abhd2
Apoa4
0 1 2 3 4 5 6
Embedded zones
cryptV1 V2 V3 V4 V5 V6
Villus zones
1.0
0.8
0.6
0.4
0.2
0.0
b
Identifying zonated genes *de novo* across liver lobules
novoSpaRc
Original
Oat
Cyp2a5
Glul
Lhpp
Fitm1
Cyp2c37
Rdh1
Cyp2e1
Cyp2c29
Lect2
Serpina12
Sds
Tstd1
Ly6e
Mfsd2a
Pigr
Prdx4
Gm5506
Sdc1
Itih3
1 2 3 4 5 6 7 8 9
Embedded layer
Lobule layer
c
Identifying spatially informative genes in the *Drosophila* embryo
Top spatially informative genes
tsh
Hsp83
abd-A
ths
oc
iab-4
pyr
bxd
opa
iab-8
Ama
bl
msi
Antp
ImpL2
Meltrin
Pino
nrm
CG42762
net
Transcription factors are inferred as highly spatially informative
all
TFs
frequency
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.3 0.4 0.5 0.6 0.7 0.8 0.9
Spatial information index
d
Identifying spatially informative genes in the mammalian cerebellum
Ttr
Enpp2
Cartpt
Car8
Plp1
Pcp2
Malat1
Calb1
Snhg11
Atp1b1
Pcp4
Itpr1
Fth1
Snap25
Meg3
Mbp
Nrsn1
Nnat
Nefl
Aldoc

Figure 6 | Utilizing novoSpaRc to identify spatially informative genes. (a, b), Identifying spatially informative genes in the mammalian intestine and liver (Methods). We identify *de novo* (no marker genes used) the most highly zonated genes along the crypt-to-villus axis in the intestine (**a**) and across the liver lobule axis, where novoSpaRc's spatial reconstruction of these genes is shown on the left and their respective original expression patterns are shown on the right. The expression level of each gene in both (a) and (b) is normalized to its maximum value. **(c, 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 (Methods). **c**, Expression patterns of the top 20 spatially informative genes in the *Drosophila* embryo (left). The spatial autocorrelation values of the 84 transcription factors chosen for the BDTNP dataset [19] are among the highest values over all 8924 genes of the fly embryo, demonstrating that they are identified to be highly spatially informative. **d**, Top 20 spatially informative genes (out of top 1000 variable genes) in a slice of the cerebellum. Four out of the five marker genes in Fig. 5 (*Pcp4, Meg3, Mbp,* and *Nefl*), which are patterned neuronal markers, are among the top spatially informative genes. For the fly (c), 0.25 fraction of the genes identified as the top 20 spatially informative genes (left) were used as marker genes. More generally, only 0.19 fraction of the genes in the top 100 spatially informative genes were used as marker genes. For the cerebellum (d), none of the genes identified as the top 20 spatially informative genes were used as marker genes.

Methods

Data acquisition and pre-processing. The single cell RNA-seq datasets were acquired from the GEO database with the following GEO accession numbers: GSE99457 for the intestinal epithelium [16], GSE84490 for the liver [13], GSE95025 for the *Drosophila* embryo [18], GSE66688 for the zebrafish embryo [9] and GSE107585 for the kidney [30]. The cerebellum Slide-seq datasets [29] were acquired from the Broad Institute Single Cell Portal (https://portals.broadinstitute.org/single_ cell/study/slide-seq-study). The individual *Drosophila* embryos dataset [31] is available as Supplemental Information file of the original manuscript. The BDTNP dataset was downloaded directly from the BDTNP webpage [19]. For the cases where normalized data was not available or used by the authors, we adopted the standard library size normalization in log-space, e.g. if d_{ij} represents the raw count for gene i in cell j, we normalized it as

$$d_{ij} \to d_{ij}^{'} = lo\,g_2\left(10^5 \times \frac{d_{ij}}{\sum_k d_{kj}} + 1\right).$$

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

In the Slide-seq datasets [29], we summed up the transcriptomes of neighboring 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 sagital (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 sagital (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 having at least a Pearson correlation of R>=0.25 with *Pcp4* were selected as marker genes.

Mathematical formulation of novoSpaRc. novoSpaRc's procedure includes several steps. We first compute the graph-based distance matrices for single cells in expression space, $D_{\square}^{\exp} \in R_{\square}^{N \times N}$, and for locations, $D_{\square}^{phys} \in R_{\square}^{M \times M}$ (Extended Data Fig. 1, Supplementary Note). Then, optionally, if a reference atlas is available, we compute the matrix of disagreement, $D_{\square}^{\exp, phys} \in R_{\square}^{N \times M}$, between each of the cells to each of the locations, based on 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[20-22]. Optimal transport is a mathematical framework that was first established in the eighteenth century by Gaspard Monge and was initially motivated by a question of the optimal (minimal cost) way to rearrange one pile of dirt into a different formation (the respective minimal cost is appropriately termed earth mover's distance). The framework evolved both theoretically and computationally [21, 22, 27] and drew extensions to correspondence between pairwise similarity measures via the Gromov-Wasserstein distance[25, 26]. Thus, in our context, it allows us to build upon these results and tools to feasibly solve the cellular assignment problem.

We would like to find a probabilistic embedding, $T^{\square}\in R_{+\text{¿}^{N\times M}\text{¿}}$, of N single cells to M locations, which 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 $T_{i,j}$ is the relative probability of embedding it to location j. These optimization requirements over $T^{\square}$ are formulated as follows. We measure the pairwise discrepancy of T for the expression and physical spaces using the Gromov-Wasserstein discrepancy[25]

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

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 corresponding 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 (e.g. 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 use the measure

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

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

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

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 efficient[27].

Putting these together, we define the optimization problem for the optimal probabilistic embedding $T^{\text{¿}}$:

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

$$subject\text{¿}$$

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

$$\sum_i T_{i,j}=q_i\forall\, j\in\{1,\ldots,M\}$$

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 $p\in\{p\in R_{+\text{¿}^N;\sum_i p_i=1\}\text{¿}}$

and $q \in \{q \in R_{+¿^{M}}; \sum_i q_i = 1\},¿$ 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 should be set to be uniform distributions.

We derive an efficient algorithm for this optimization problem inspired by the combined results for entropically regularized optimal transport[27] and Gromov-Wasserstein distance-based mapping between metric-measure spaces[26] (Supplementary Note).

Then, given the original single cell expression profiles, represented by a matrix $A \in R^{N\times g}$ (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), $S = A^T T \in R_{+¿^{g\times M}¿}$ (for g genes and M locations), which contains the gene expression values for every cellular location of the target space.

Note again that since 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 that the single cells could have originated from.

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 (e.g. highly variable).
2. Construct the target space.
3. Setup 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. derive a virtual *in situ* hybridization (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.

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. 3a,c). (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.

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, $E(x \vee \mu_g, \sigma_g) \propto e^{\frac{-(x-\mu_g)^2}{2\sigma_g^2}}$, 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, $E(x \vee \mu_g, \Sigma_g) \propto e^{\frac{-1}{2}(x-\mu_g)^T \Sigma_g^{-1}(x-\mu_g)}$, where μ_g is the mean vector (sampled uniformly across the 2d or 3d grid), and Σ_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 $\sigma_{expression} = 0.1$.

Generating *in silico* single cell data for BDTNP and virtual embryo datasets. To test the performance of novoSpaRc with single-cell resolution ground truth, we generated *in silico* single cell datasets for two cases: the BDTNP data [19] and the virtual *Drosophila* embryo data [18]. In both cases we have access to expression profiles for different locations across the embryo. We effectively dissociate the embryos 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 embryos (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 corresponding to each archetype is provided as a Supplementary File.

Identification of zonated genes. For tissues with 1d symmetry, we produce a ranking of highly zonated genes, both according to the original spatial expression patterns (Extended Data Figs. 3b,e) and the reconstructed patterns (Fig. 6a,b).

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 zonated towards the first or last spatial zones (e.g. 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, (iii) and that are statistically significantly zonated. 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 a list of the zones numbering. Finally, the remaining genes are ranked according to their center of mass.

The list of predicted zonated genes based on novoSpaRc's reconstruction for the mammalian intestine and liver are available as Supplementary Files.

Gene ontology enrichment. We used GOrilla for GO enrichment analysis[32], where GO enrichment was computed based on target and background lists of genes (Supplementary Note). For both the target and background lists of genes we selected genes whose maximum expression value is in the top 10% of all genes. The target lists for genes zonated towards the boundaries of the 1d spatial axes (crypt and V6 in intestine, layer 1 and 9 in liver) were further filtered to contain only genes that are statistically significantly zonated, as described in the 'identification of zonated genes' Methods subsection. 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, Moran's I measure 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 $z_i = y_i - \overline{y}_i$, $\overline{y}_i$ 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 neighbors graph for each cellular location (k=8). To calculate Moran's I measure and respective p-values for different genes, we used the implementation of PySAL, a Python spatial analysis library [34].

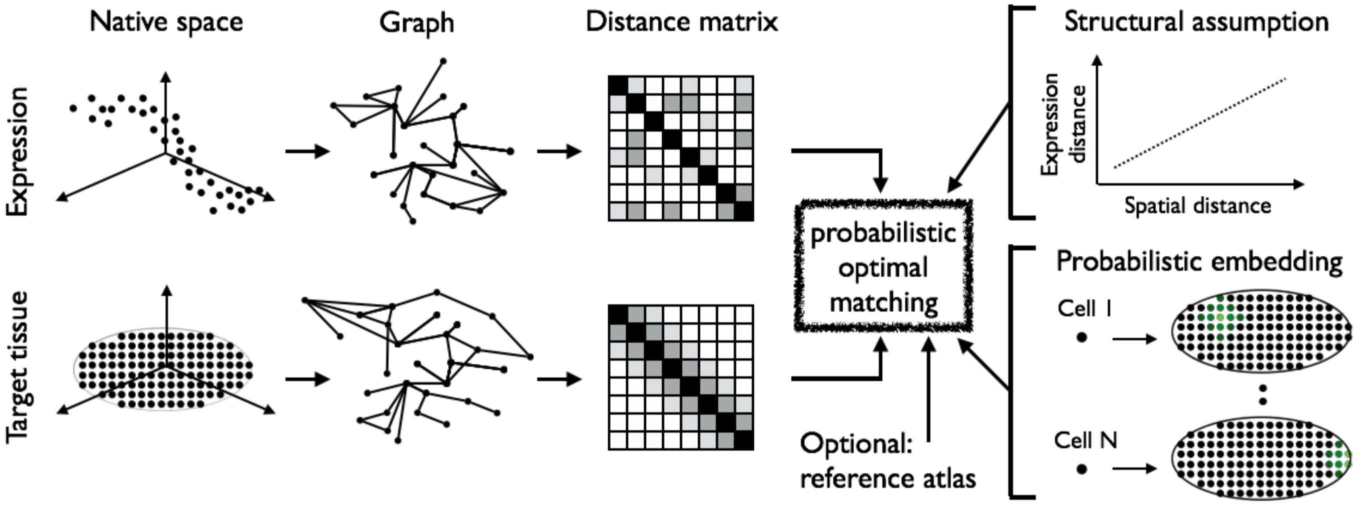
We acknowledge the use of Moran's I measure for single cell analysis by [35] and the Monocle 3 tutorial by the Trapnell Lab.

The Moran's I scores, with their respective p-values, based on novoSpaRc's reconstruction for all genes of the *Drosophila* embryo, zebrafish embryo, and cerebellum are available as Supplementary Files.

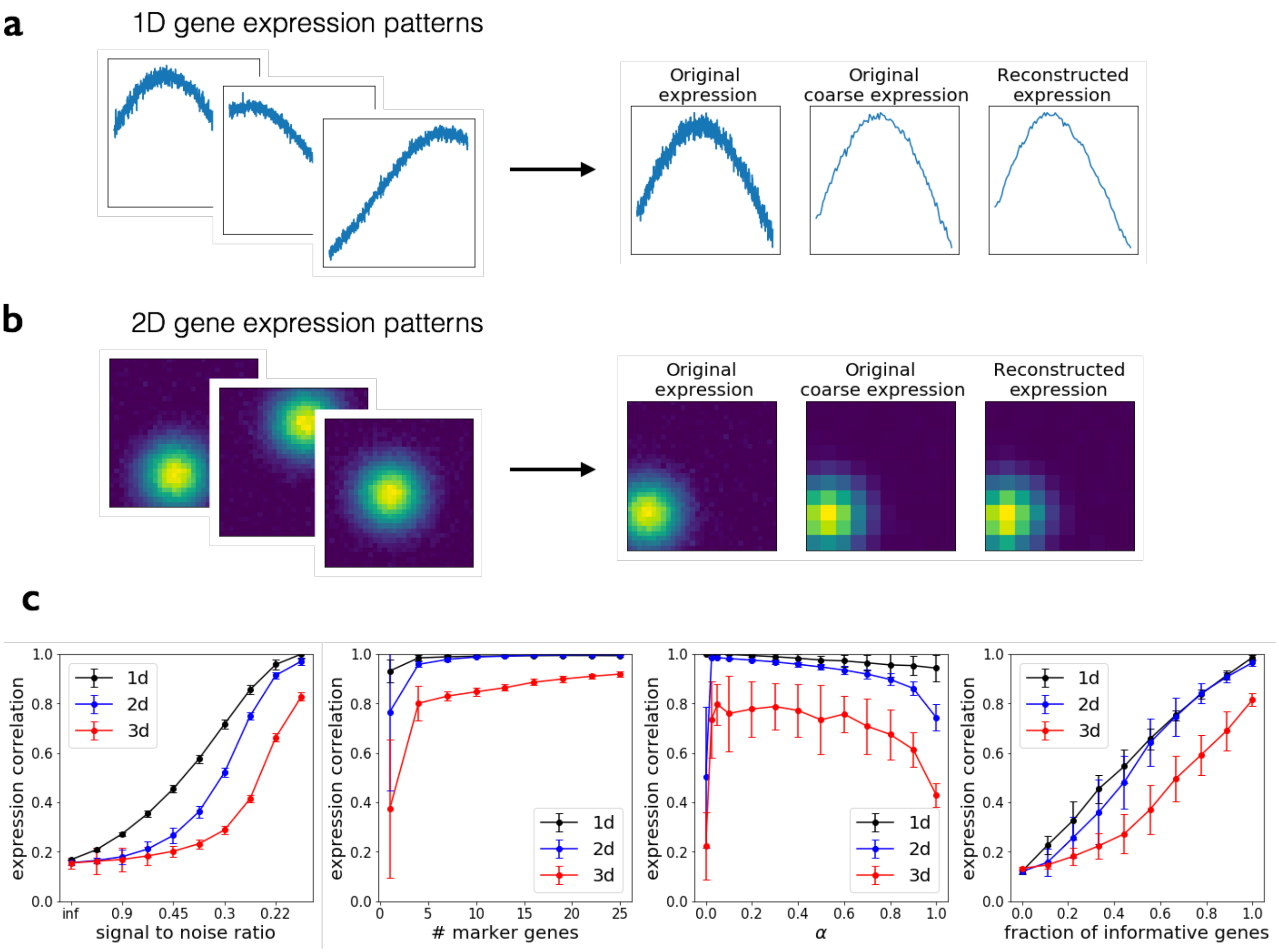
novoSpaRc's advantages, limitations, and direct comparison to existing reconstruction methods. novoSpaRc offers several features which cannot be exploited as a whole by existing methods: (a) it enables incorporation and interpolation of both structural information (such as the structural correspondence 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 [9] and DistMap [18]. A comparison of the intrinsic characteristics of the three approaches is provided in Extended Table 1. The reconstruction results for the BDTNP data [19], as well as scRNA-seq data of the *Drosophila* [18] and zebrafish embryos [9] and the cerebellum [29] using the three different approaches is shown in Extended Data Fig. 18. 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.

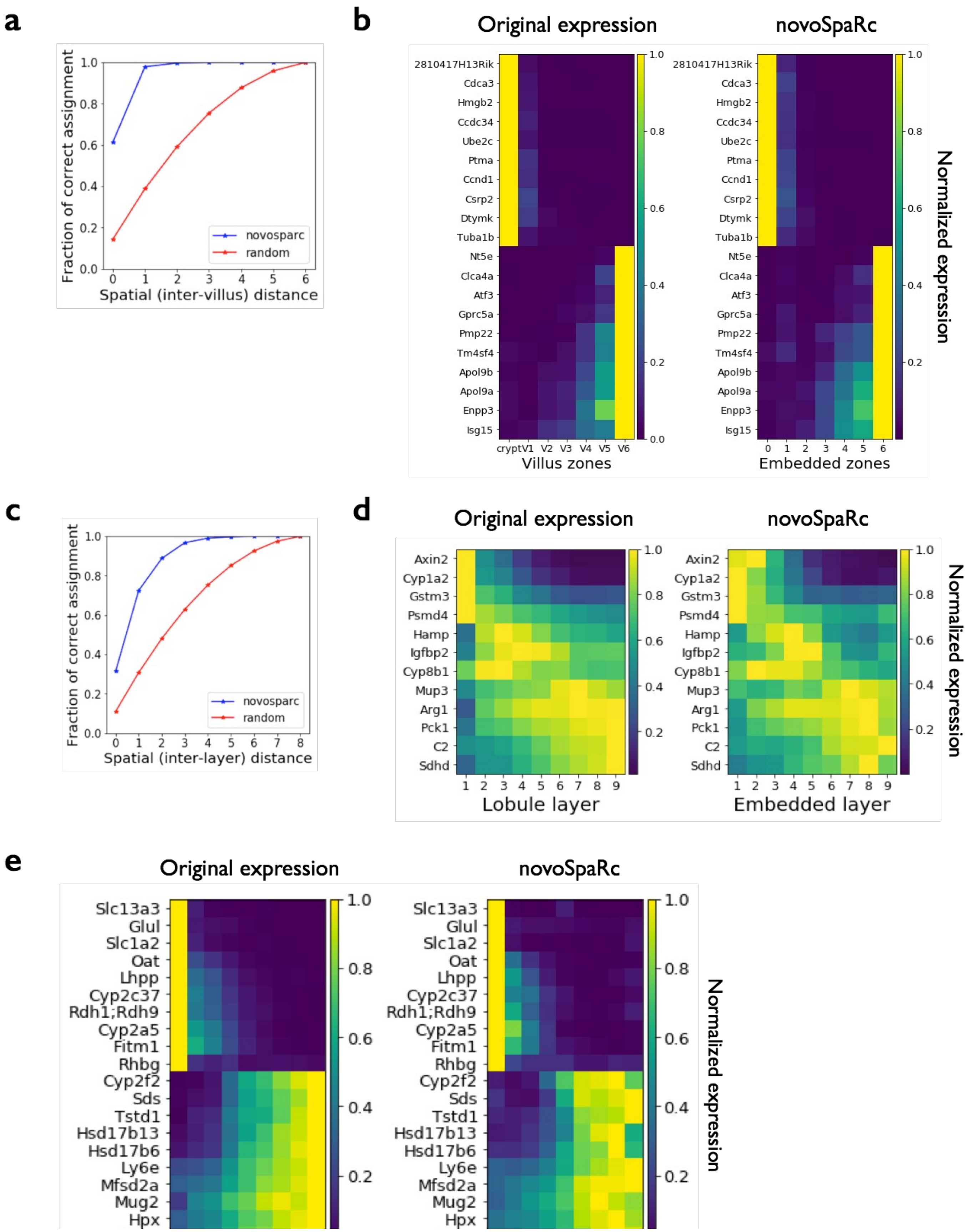
It is important, however, to keep in mind novoSpaRc's limitations. novoSpaRc works by embedding the single cells into a predefined shape, and so does not allow to learn a latent representation of the data that was not used as input. In addition, as mentioned in the main text, *do novo* reconstruction 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 (Supplementary Note). 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. However, 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.



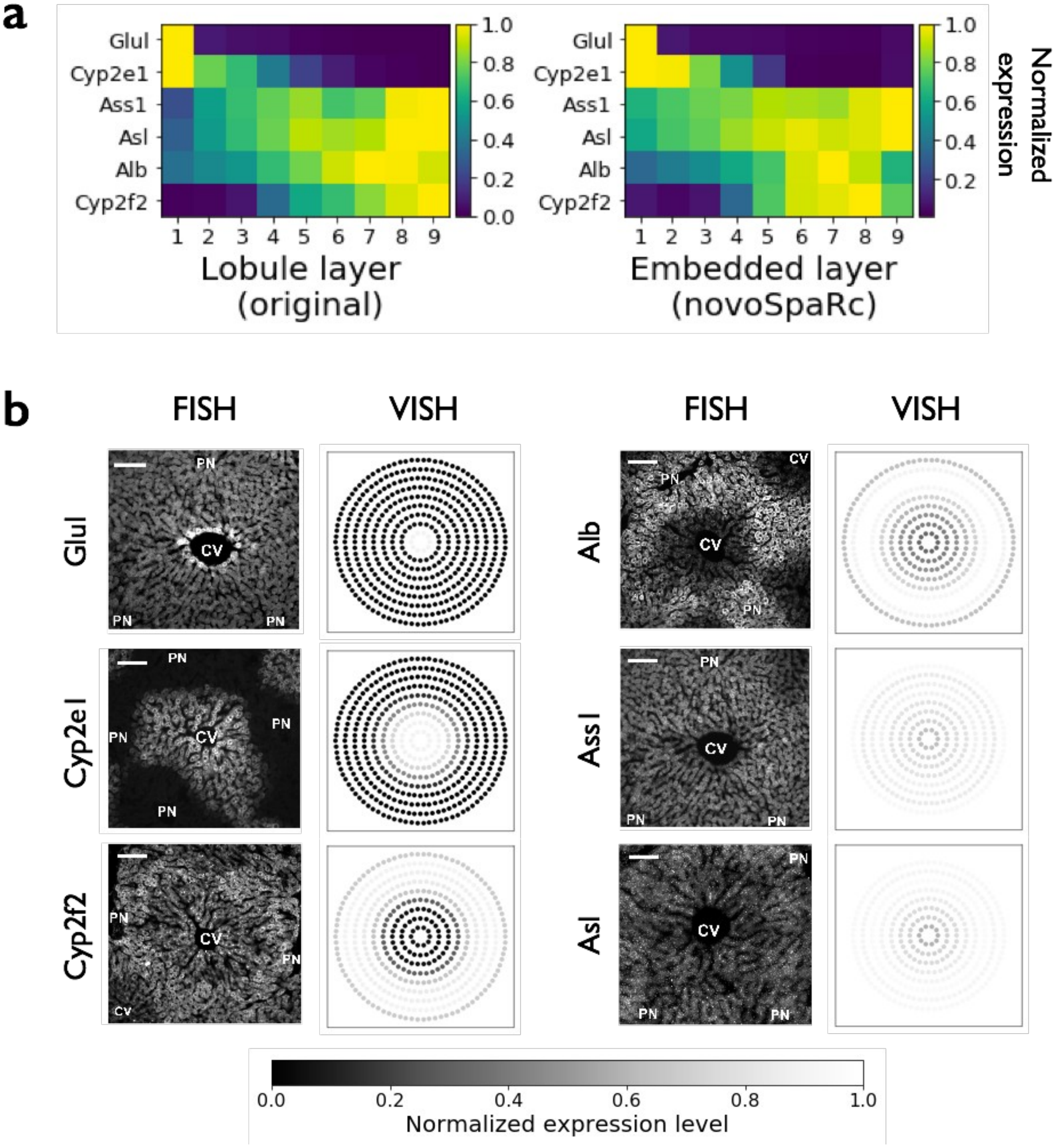
Extended Data Figure 1 | Overview of probabilistic optimal matching using novoSpaRc. Based on the raw data of single cells in expression space and locations along a grid resembling the target tissue, graph structures are computed, and distance matrices are derived from these graphs (Supplementary Note). The two branches, and potentially a reference atlas, are aligned using novoSpaRc, under our structural correspondence assumption (distance in expression space is on average monotonically increasing with distance in physical space) and by using probabilistic embedding (Supplementary Note).



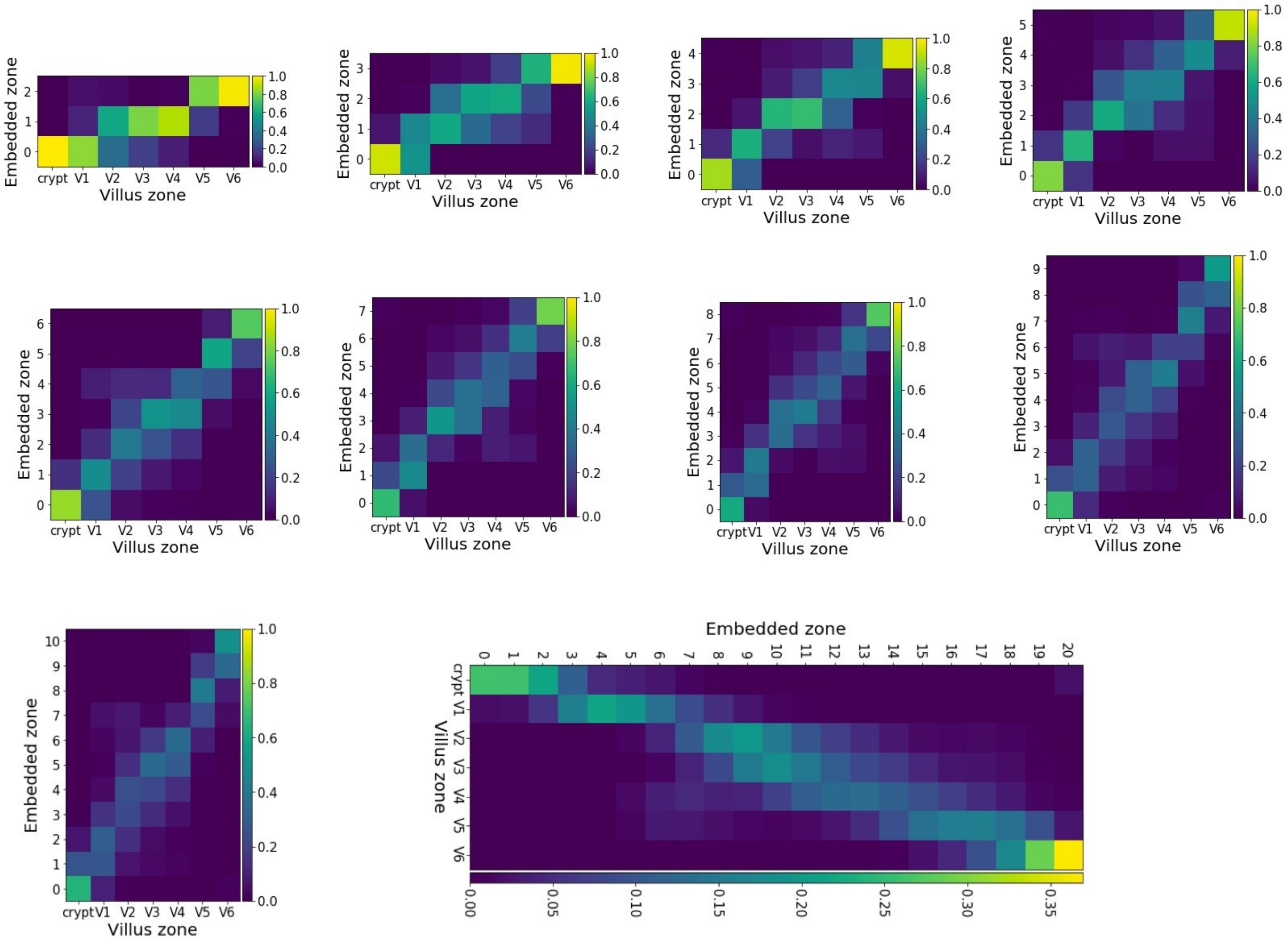
Extended Data Figure 2 | Generative model for spatial gene expression. The generative model and its default parameters are described in Methods. (**a, b**) Visualization of noisy expression patterns for three random genes in models for a 1d (**a**) and 2d (**b**) tissues are shown on the left. On the right we show the original expression pattern for a representative gene, its coarse-grained representation (decreased spatial resolution), and its reconstruction using novoSpaRc. **c,** Pearson correlation of the reconstructed expression patterns 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 behavior with the alpha parameter. We note that alpha is an interpolation parameter (defined in the section 'Mathematical formulation of novoSpaRc' in Methods), between using only a reference atlas (α = 1) and using both structural information (driven by the structural correspondence assumption) and a reference atlas. Results are averaged over 100 instantiations of the generative model, where center point, mean; error bars, SD.


a
Fraction of correct assignment
1.0
0.8
0.6
0.4
0.2
0.0
0 1 2 3 4 5 6
Spatial (inter-villus) distance
novosparc
random
b
Original expression
novoSpaRc
2810417H13Rik
Cdca3
Hmgb2
Ccdc34
Ube2c
Ptma
Ccnd1
Csrp2
Dtymk
Tuba1b
Nt5e
Clca4a
Atf3
Gprc5a
Pmp22
Tm4sf4
Apol9b
Apol9a
Enpp3
Isg15
crypt V1 V2 V3 V4 V5 V6
Villus zones
0 1 2 3 4 5 6
Embedded zones
1.0
0.8
0.6
0.4
0.2
0.0
Normalized expression
c
Fraction of correct assignment
1.0
0.8
0.6
0.4
0.2
0.0
0 1 2 3 4 5 6 7 8
Spatial (inter-layer) distance
novosparc
random
d
Original expression
novoSpaRc
Axin2
Cyp1a2
Gstm3
Psmd4
Hamp
Igfbp2
Cyp8b1
Mup3
Arg1
Pck1
C2
Sdhd
1 2 3 4 5 6 7 8 9
Lobule layer
1 2 3 4 5 6 7 8 9
Embedded layer
1.0
0.8
0.6
0.4
0.2
Normalized expression
e
Original expression
novoSpaRc
Slc13a3
Glul
Slc1a2
Oat
Lhpp
Cyp2c37
Rdh1;Rdh9
Cyp2a5
Fitm1
Rhbg
Cyp2f2
Sds
Tstd1
Hsd17b13
Hsd17b6
Ly6e
Mfsd2a
Mug2
Hpx
1.0
0.8
0.6
0.4
0.2
Normalized expression

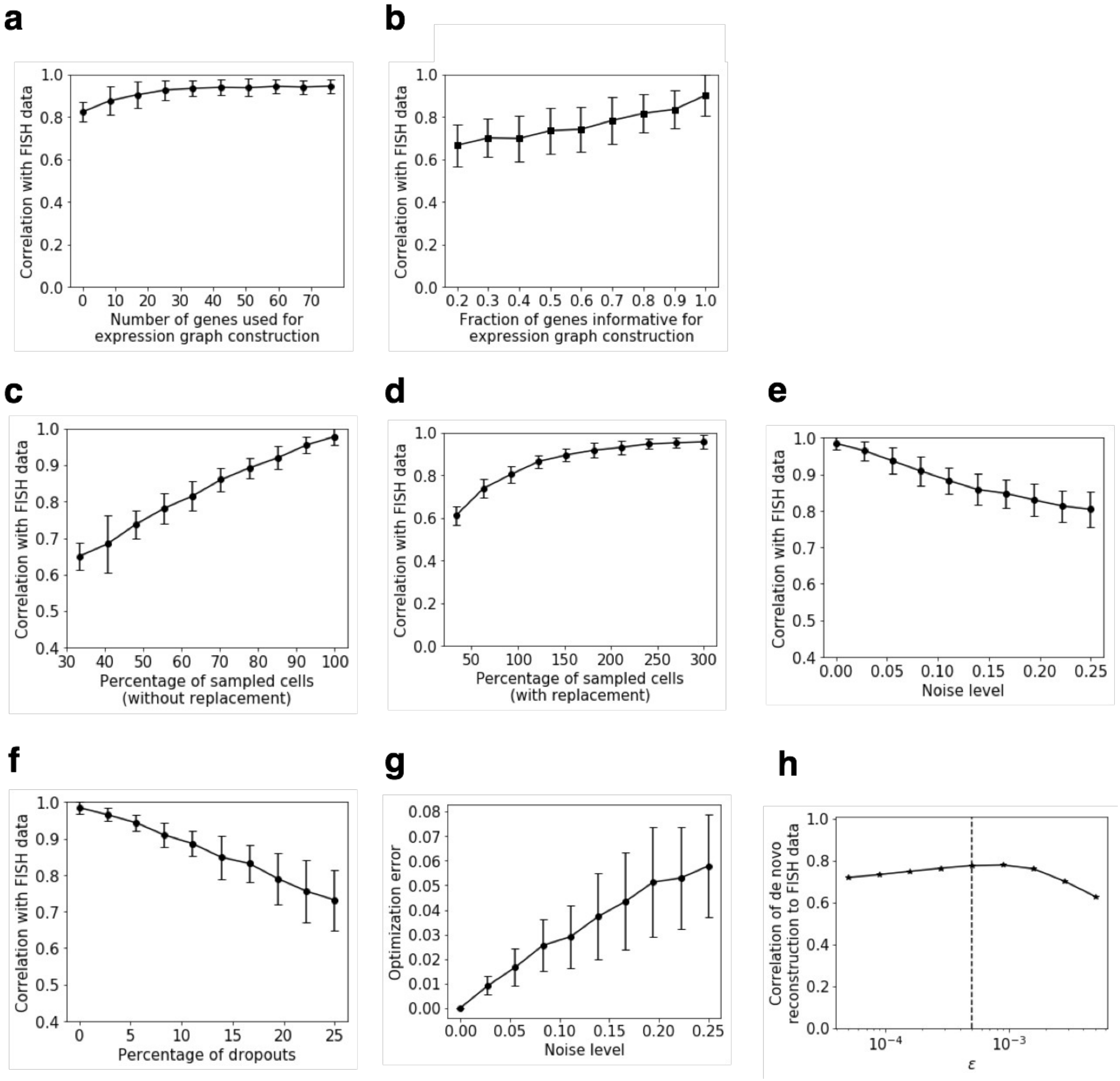
Extended Data Figure 3 | Evaluation of novoSpaRc reconstruction of the intestinal epithelium and the liver lobule. a, The fraction of cells in the crypt-to-villus axis (y-axis) that is correctly assigned to its corresponding original villus zone[16], or is assigned to a zone up to a d-zones-away from the original zone (x-axis), is substantially higher than that of random assignment. **b,** novoSpaRc successfully reconstructs the spatial expression patterns of the top zonated genes (10 top zonated genes towards the crypt, and 10 top zonated genes towards V6). **c,** The fraction of cells in the liver lobule axis (y-axis) that is correctly assigned to its corresponding original lobule layer[13], or is assigned to a layer up to a d-layers-away from the original layer (x-axis), is substantially higher than that of random assignment. novoSpaRc successfully reconstructs the spatial expression patterns of **d,** a group of pericentral, periportal and non-monotonic genes (complementing Fig. 2h) and **e,** the top zonated genes (10 top zonated genes towards the CV, and 10 top zonated genes towards PV). Selection of top zonated genes is described in Methods. The expression level of each gene in (b,d,e) is normalized to its maximum value.



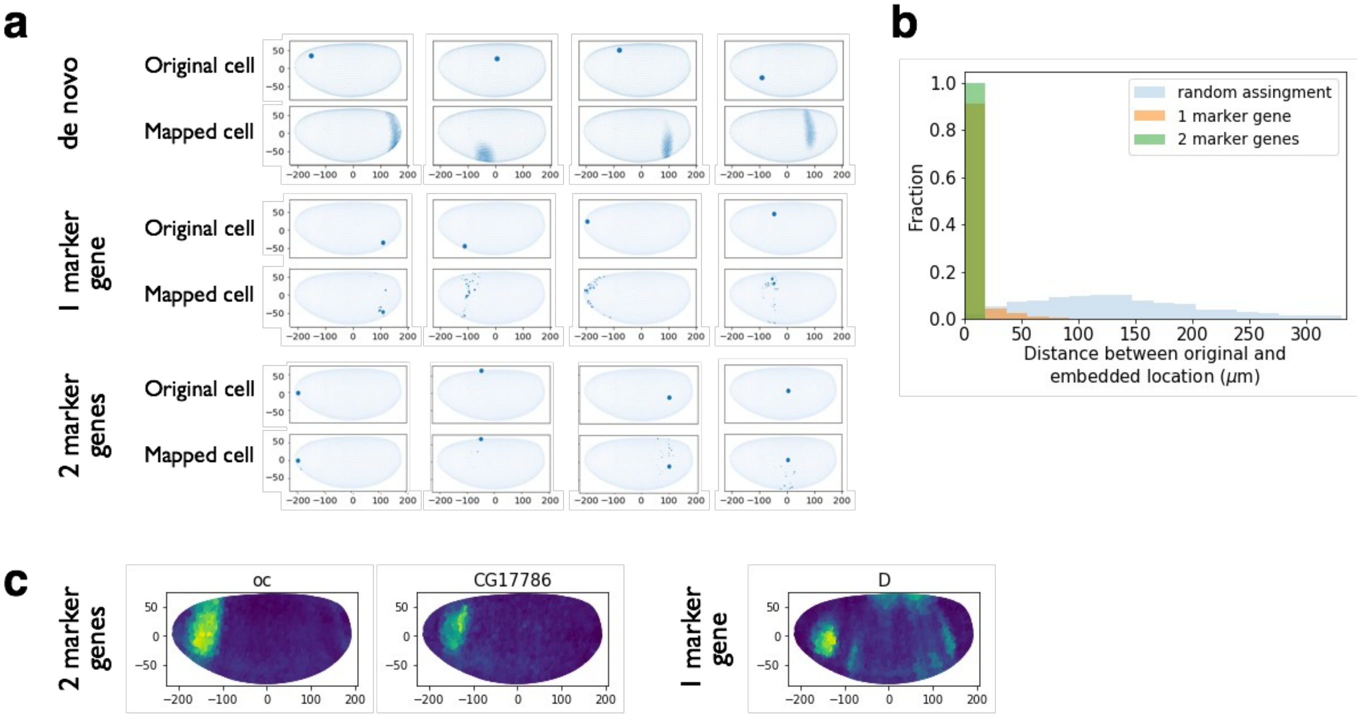
Extended Data Figure 4 | novoSpaRc reconstructs the mammalian liver *de novo*. Examples of FISH expression patterns of six zonated genes across the liver lobules, comparing the reconstructed (*de novo* vISH data) expression patterns by novoSpaRc to (a) the expression patterns reported in [13], and (b) the original (FISH) data (adapted from [13]). The visualization in (a) is a heatmap, showing the expression values of each gene across the lobule layers. The visualization of the reconstructed vISH data in (b) is meant to be comparable to the FISH images, and therefore the 1d reconstructed coordinates are projected onto a polar coordinate system (CV – middle, PV – outer circumference). The expression level of each gene in both (a) and (b) is normalized to its maximum value.



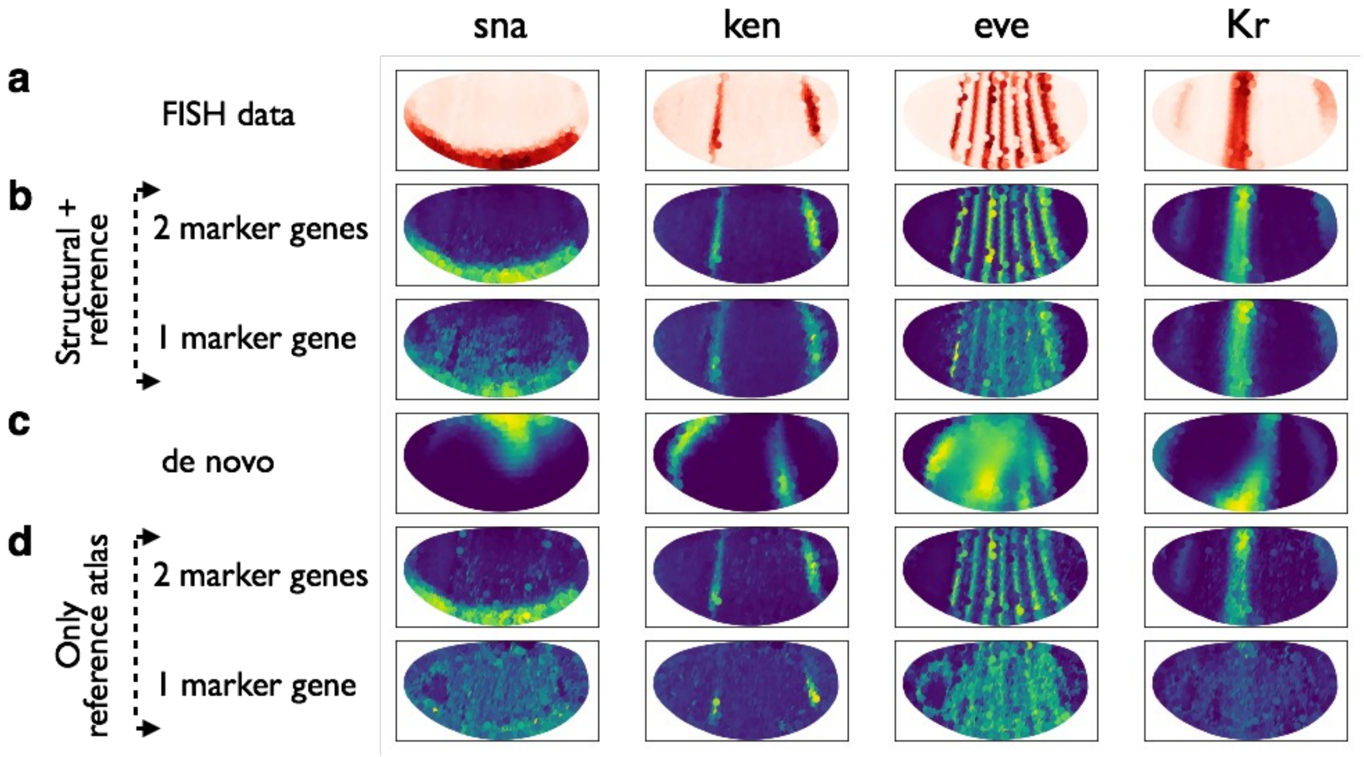
Extended Data Figure 5 | novoSpaRc reconstruction of the intestinal epithelium is robust and consistent with changing grid resolution. The successful *de novo* reconstruction of the intestinal epithelium dataset[16] is achieved for varying numbers of layers used for the target space (including both lower and higher number of layers compared with the seven original number of reference layers). The expression level of each gene is normalized to its maximum value.



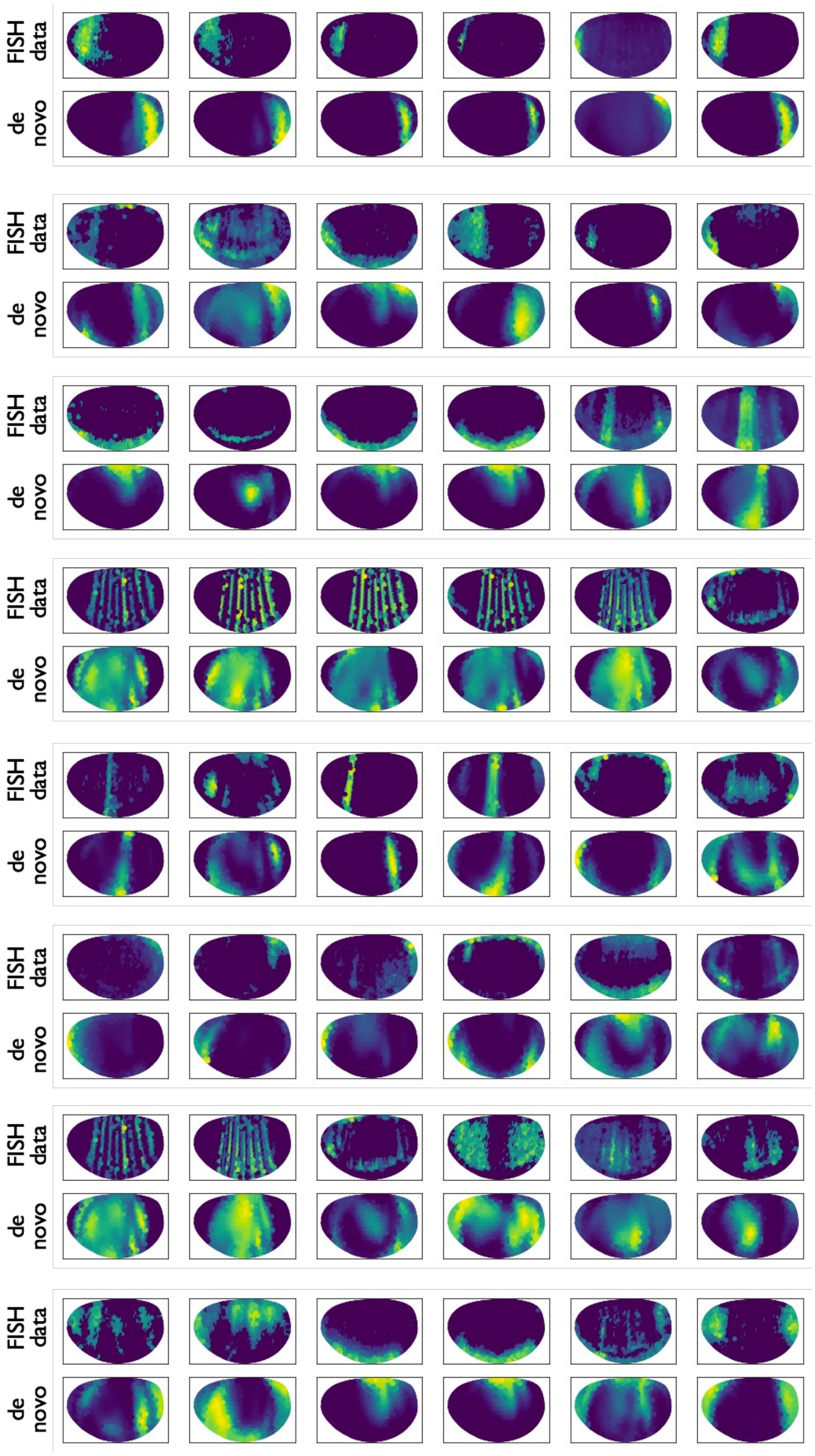
Extended Data Figure 6 | novoSpaRc reconstruction of the Drosophila embryo based on the BDTNP dataset is robust. a, Pearson correlation of the reconstructed expression patterns to the original FISH expression data [19] increases with the number of genes used to construct the structural cellular graph in expression space, and **b**, with the fraction of those genes that are spatially-informative, where spatially non-informative genes in this case were simulated as random Gaussian variables with mean and standard deviation comparable to that of the original gene set. Pearson correlation of the reconstructed expression patterns to the original FISH expression data [19] increases with the percentage of sampled single cells (without (**c**) and with (**d**) replacement), and steadily decreases with noise level (**e**) and percentage of dropouts in the data (**f**). **g,** The mean value and variance of the optimization objective function (which we aim to minimize) increases with noise level. **h**, Pearson correlation of the *de novo* reconstructed expression patterns to the original FISH data varies gradually with the entropic regularization parameter ϵ. Results for subplots (a-g) are averaged over 100 random choices of 2 marker genes, where center point, mean; error bars, SD.



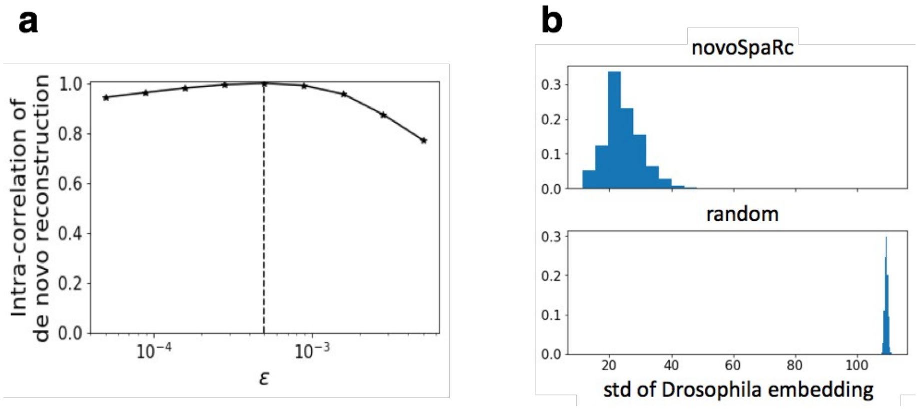
Extended Data Figure 7 | novoSpaRc predicts spatial positions of individual cells. a, Examples of mapping probabilities of single cells produced by novoSpaRc for the *Drosophila* embryo, based on the BDTNP dataset[19]. The predicted spatial positions of cells are distributed in a localized fashion over relatively many locations when reconstruction is done *de novo* (top panel), and are sharply localized when marker genes are used (1 and 2 marker genes, middle and bottom panel). **b**, Histogram of Euclidean distances between the original cellular location of single cells and the most likely location predicted by novoSpaRc using 1 and 2 marker genes, and compared to a histogram for random predictions. **c,** The expression patterns of the 2 and 1 marker genes used for the results presented in panels a,b**.**



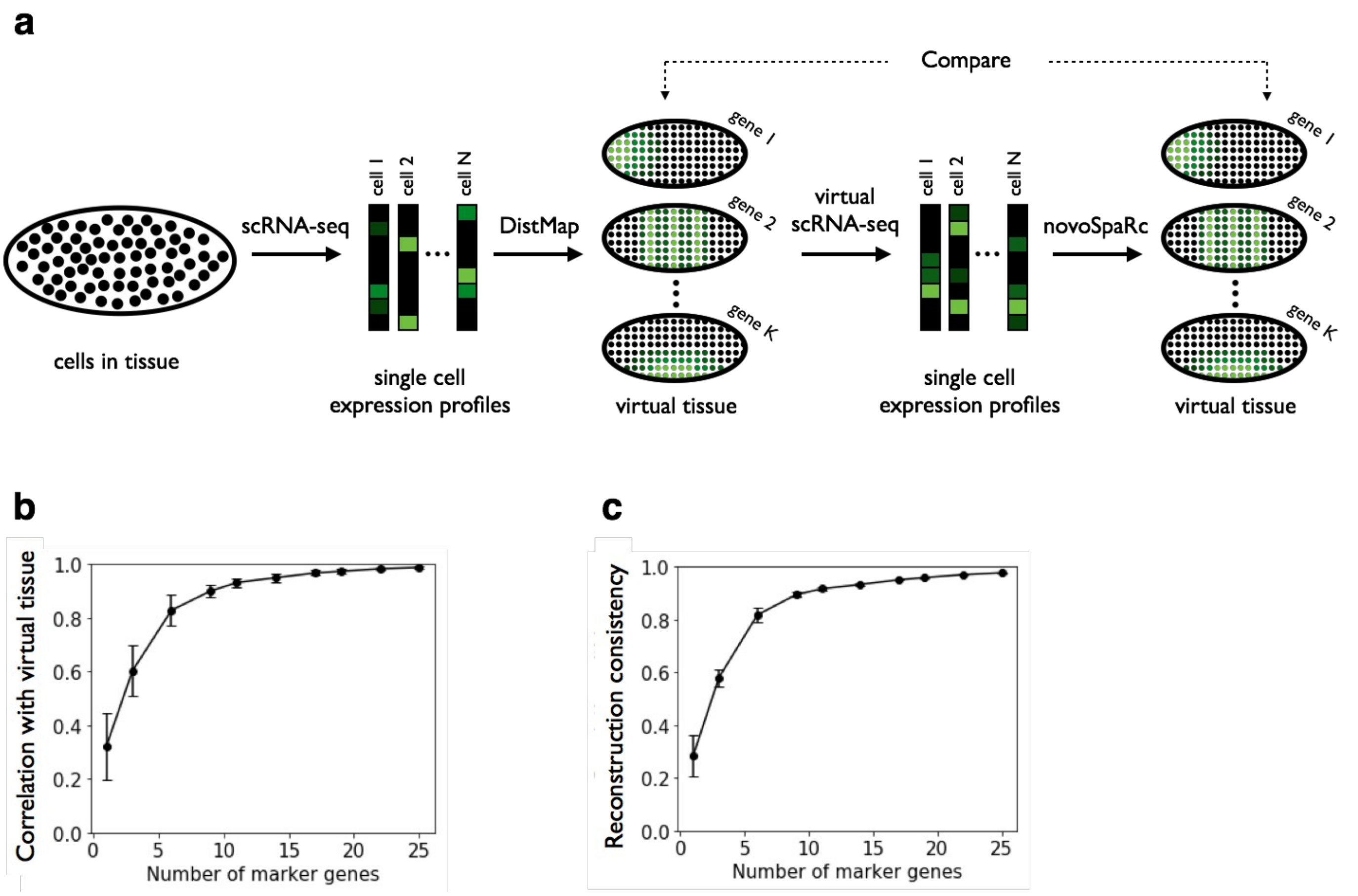
Extended Data Figure 8 | novoSpaRc accurately reconstructs the Drosophila embryo based on the BDTNP dataset[19]. Visualization of reconstruction results for 4 transcription factors. The original FISH data (**a**) is compared to reconstruction by novoSpaRc that exploits both structural and marker gene information (using 2 and 1 marker genes, **b**) and reconstruction without any marker gene information (*de novo*, **c**). Reconstruction using both structural and marker gene information (or a reference atlas) outperforms reconstruction based solely on a reference atlas (**d**).



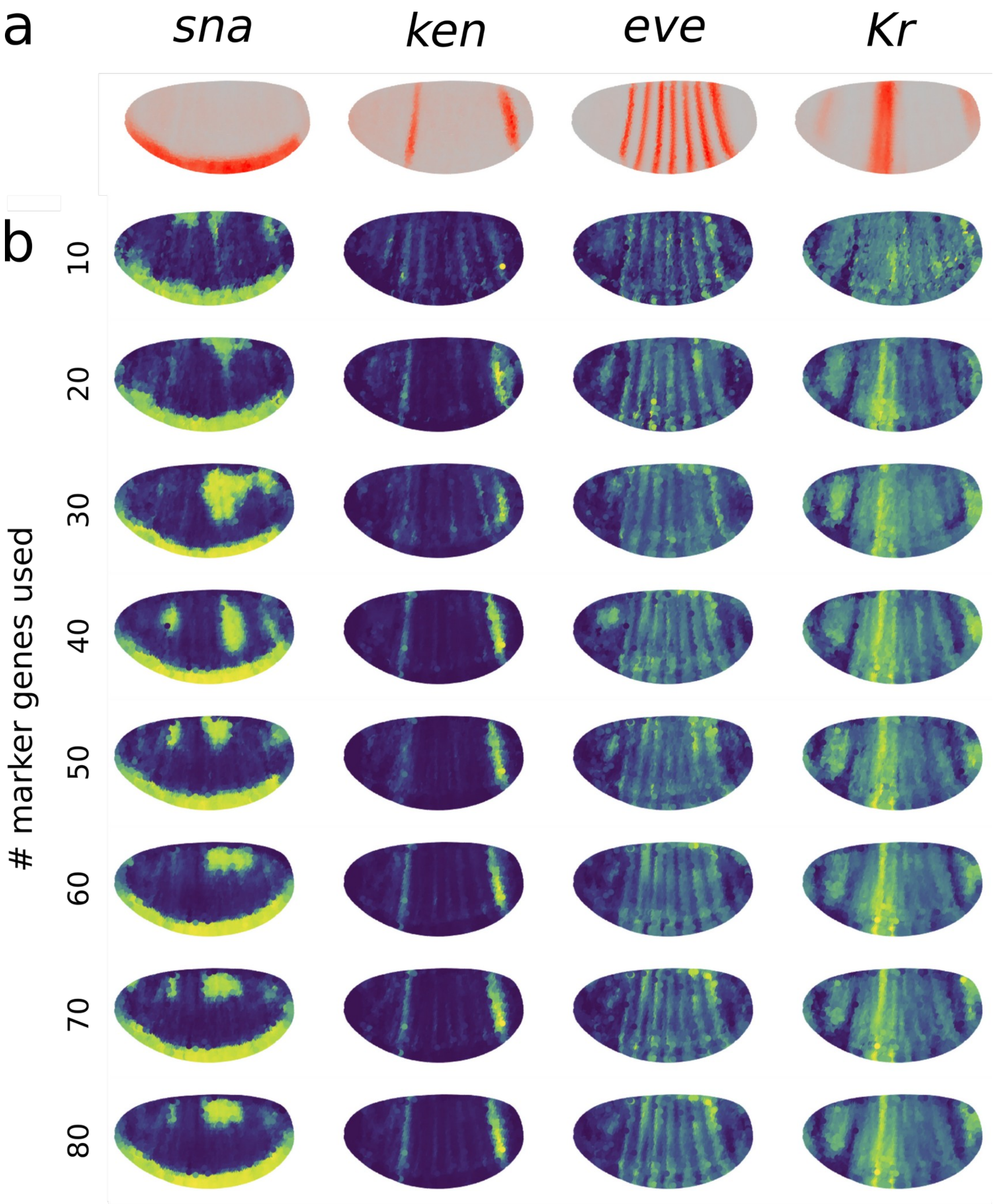
Extended Data Figure 9 | novoSpaRc reconstructs the *Drosophila* embryo *de novo*. Reconstruction based on the BDTNP dataset[19]. Examples of marker gene expression patterns across the embryo comparing the original (FISH) data and the reconstructed (*de novo*) data.



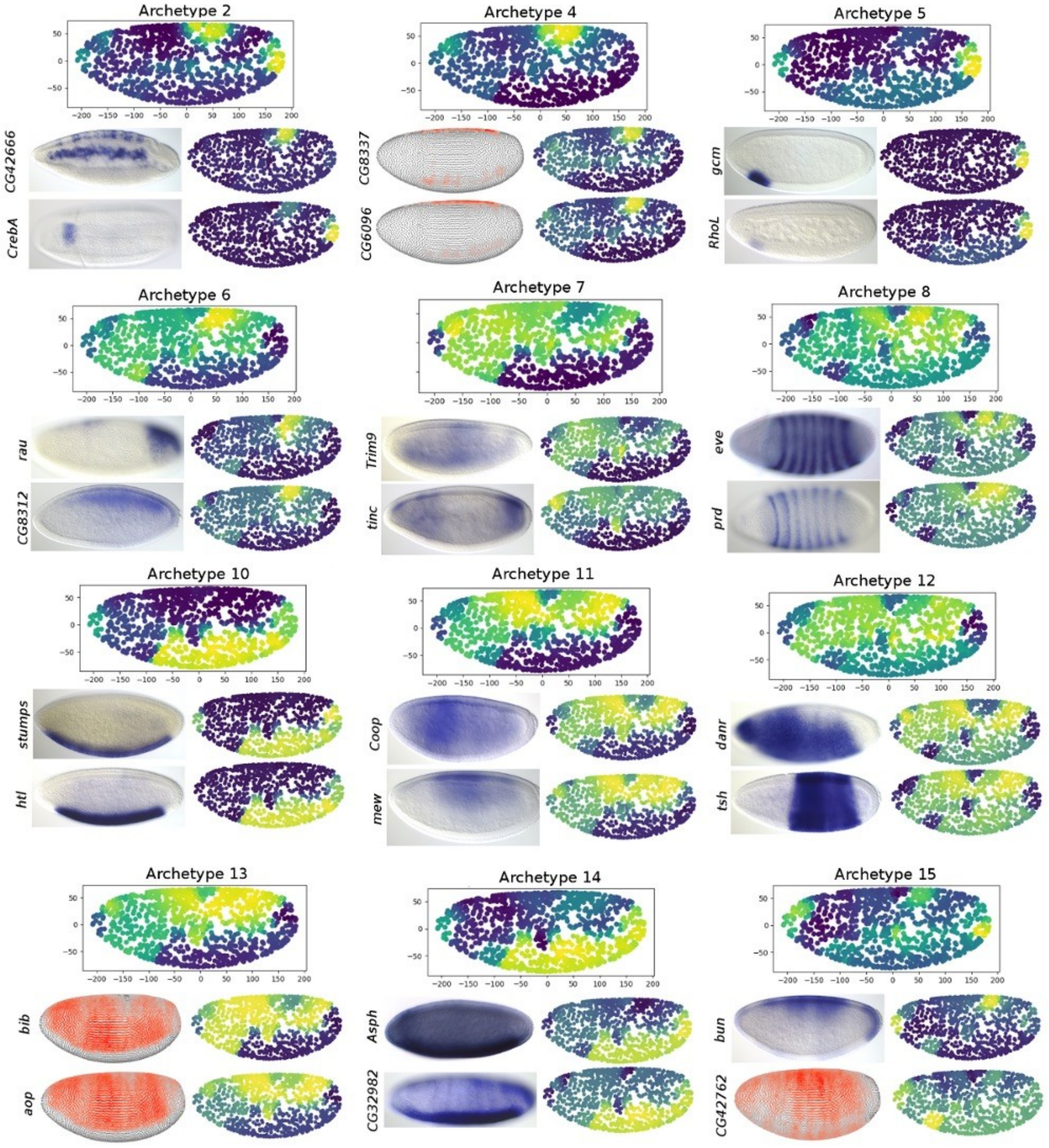
Extended Data Figure 10 | Self-consistency analysis of *de novo* reconstruction with novoSpaRc. a, Pearson correlation of embedded *de novo* expression patterns of the BDTNP dataset [19] for different values of the entropic regularization parameter (ϵ) with the expression pattern for $\epsilon = 5 * 10^{-5}$ (vertical dotted line). **b,** The spatial standard deviation of embedded cells over the *Drosophila* embryo of the BDTNP dataset via *de novo* novoSpaRc is statistically significantly lower than the standard deviation of randomized embedding (two-sample K-S p < 10^{-200}).



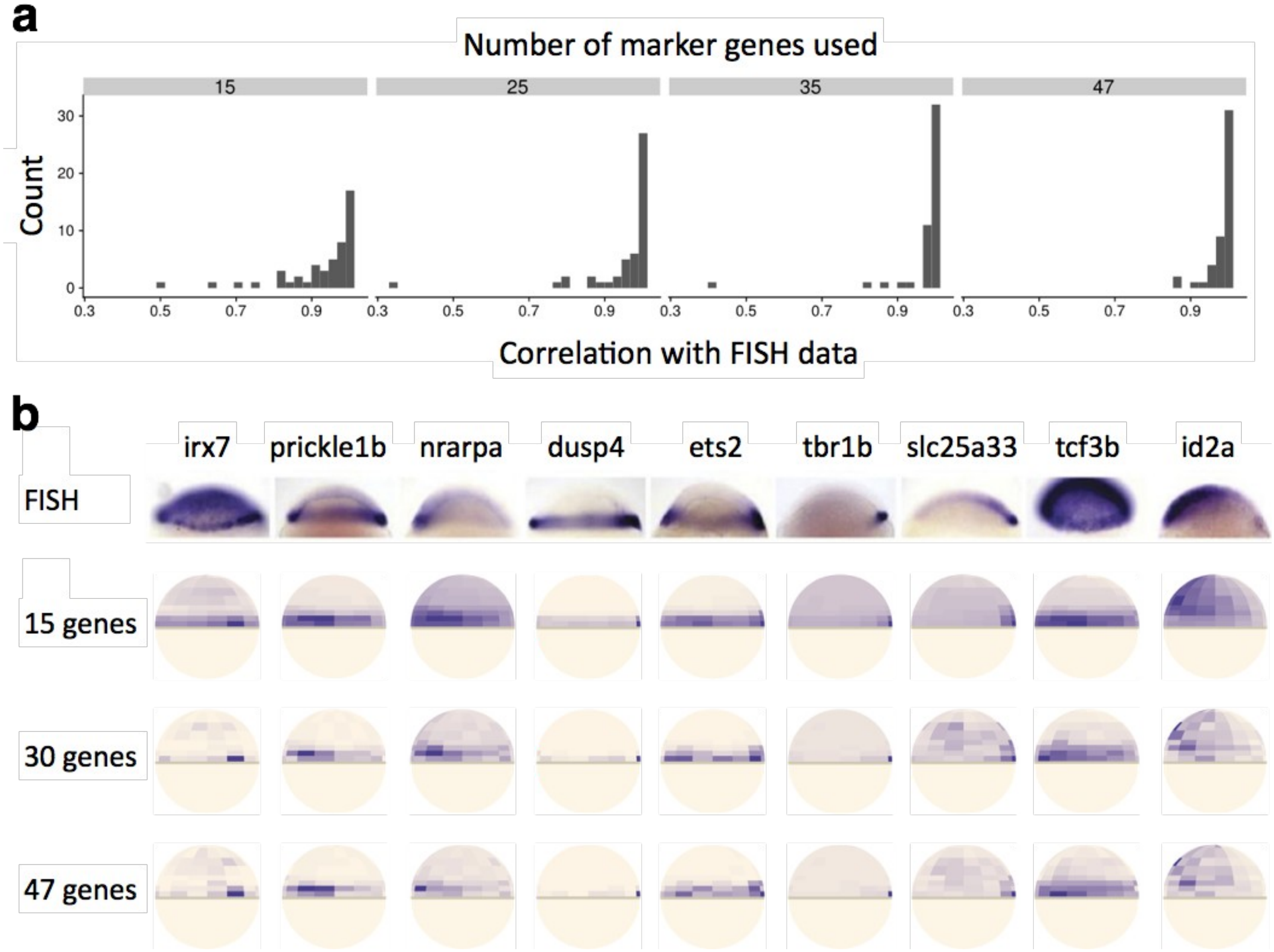
Extended Data Figure 11 | novoSpaRc reconstructs the Drosophila virtual embryo. **a**, Overview of the process of the spatial reconstruction of the Drosophila virtual embryo. **b**, The Pearson correlation of the reconstructed expression patterns of the virtual embryo with the corresponding original data increases with the number of marker genes used for the reconstruction. **c**, The self-consistency of reconstruction of the virtual embryo increases with the number of marker genes. The consistency score was calculated as the average pairwise Pearson correlation within reconstructed expression patterns for different sets of marker genes. Results are averaged over 100 random choices of 4 marker genes. For subplots b,c: center point, mean; error bars, SD.



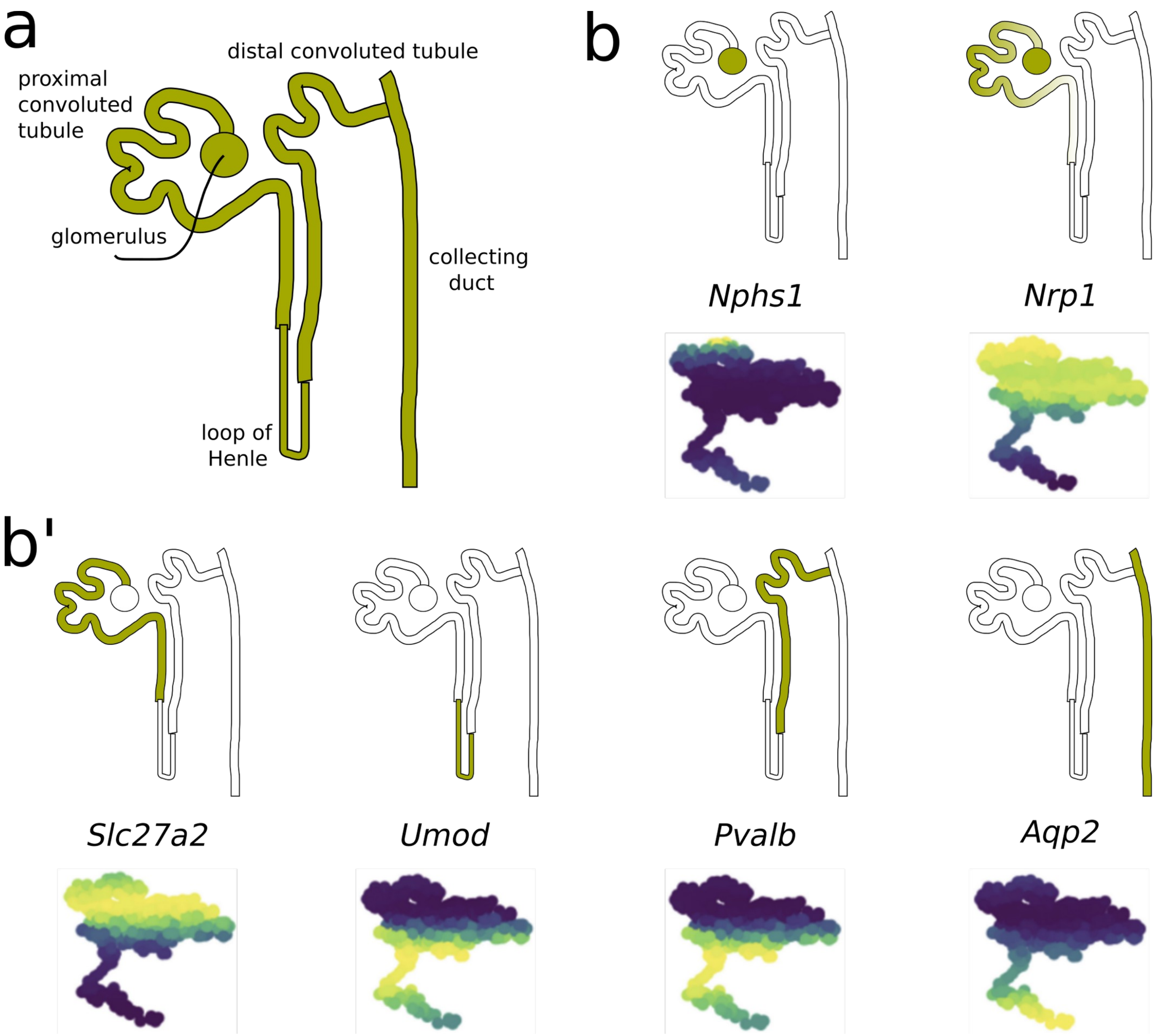
Extended Data Figure 12 | novoSpaRc accurately reconstructs the Drosophila embryo based on single cell data. Original spatial expression (a) compared to visualization of novoSpaRc–based reconstruction results (b) for 4 transcription factors that exploits both structural and marker gene information (using 10-80 marker genes).



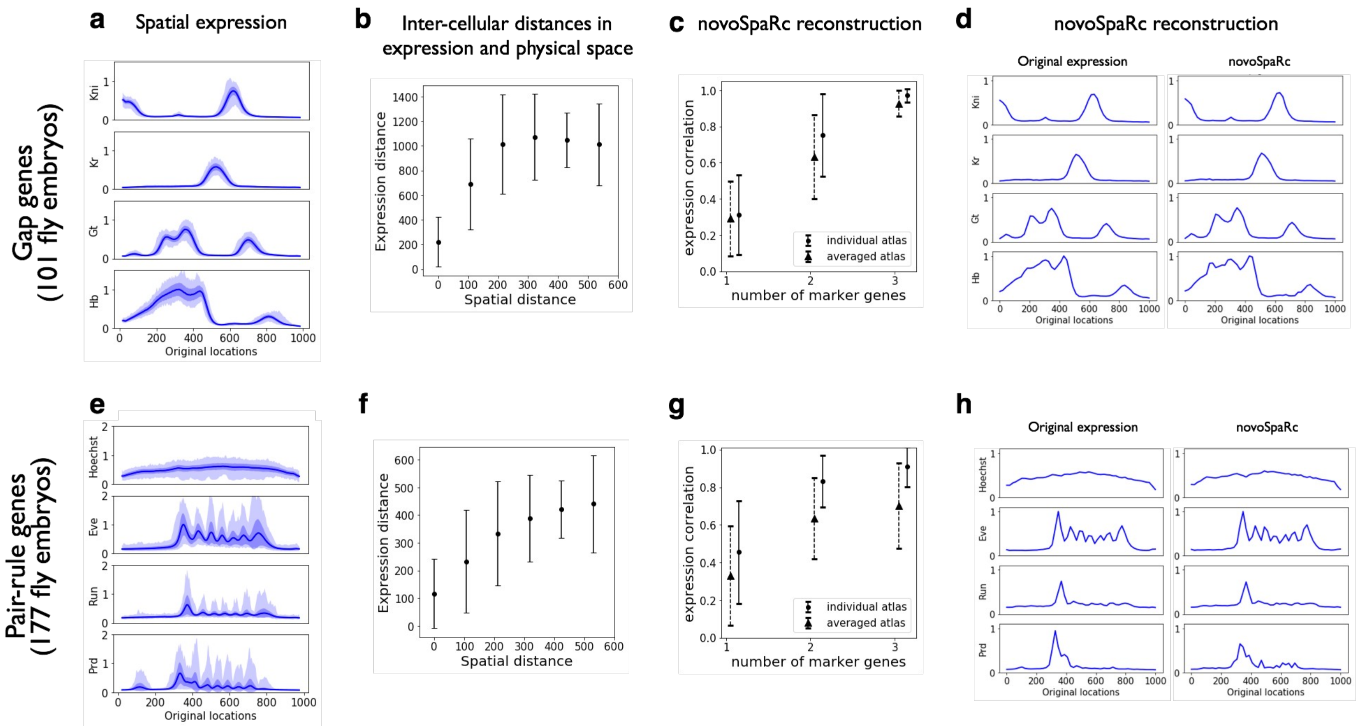
Extended Data Figure 13 | novoSpaRc identifies spatially informative archetypes by using scRNA-seq data for the *Drosophila* embryo. The archetypes shown complement those of Fig. 4c,d in the main text. Preferred spatial positioning is denoted by coloring ranging from blue (low) to yellow (high). FISH images were taken from the BDGP database[36]. For genes for which an image was not available, DVEX*[18]* was used instead. Two representative genes are shown for each spatial archetype. novoSpaRc accurately groups genes expressed in a particular domain, such as the subdomain of the mesoderm characterized by the transcription factor *gcm* (Archetype 5), while it does not capture the details of the fine expression patterns of pair-rule genes (Archetype 8).



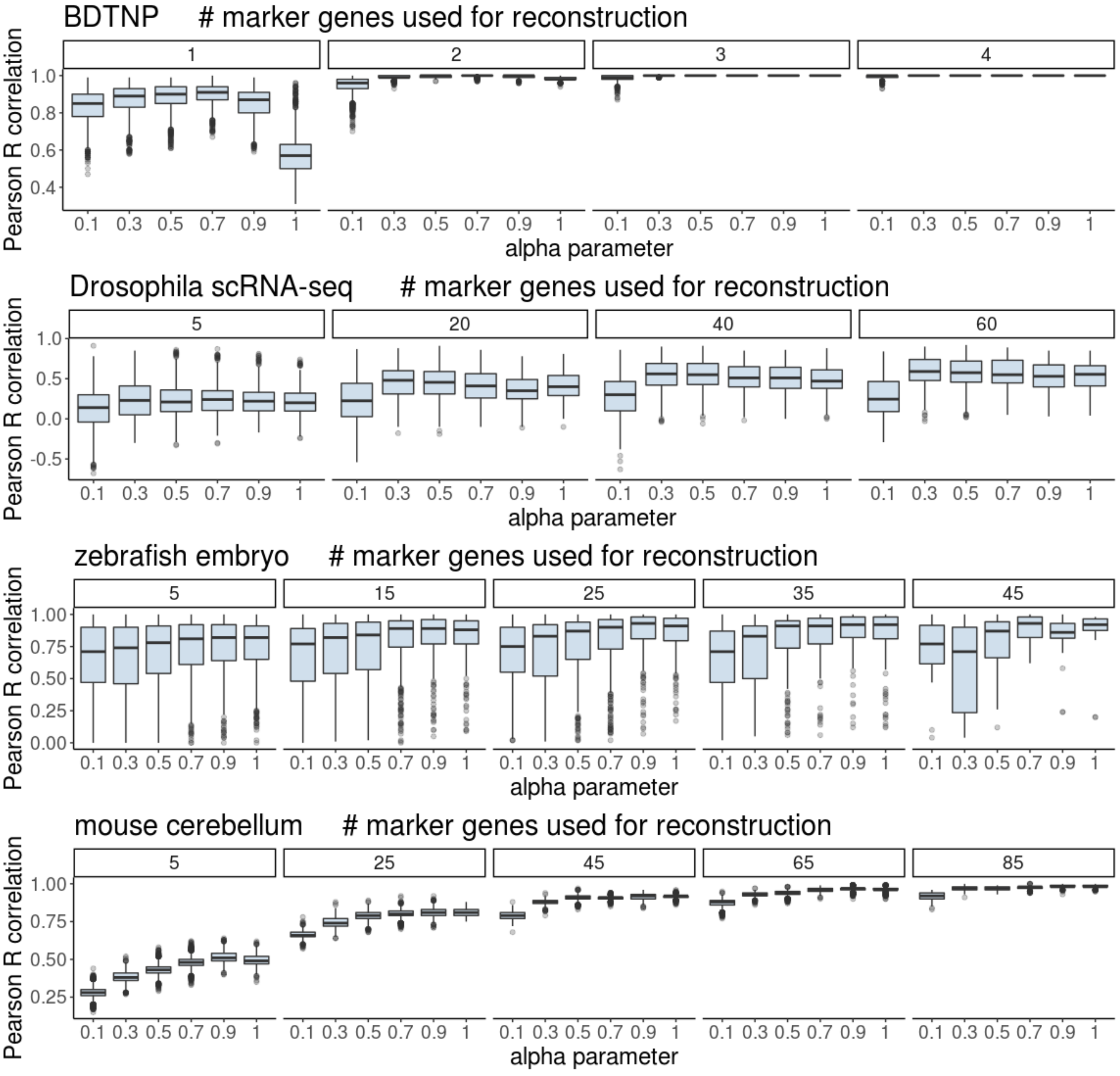
Extended Data Figure 14 | novoSpaRc reconstruction of the zebrafish embryo improves with the number of marker genes. a, Histograms assessing the increase in the accuracy of novoSpaRc reconstruction, measured by the Pearson correlation with FISH data[9], with increasing number of marker genes. **b**, Top row: FISH data[9]; following rows: novoSpaRc predictions by using 15, 30 and 47 marker genes. Genes shown were not used for any of the reconstructions.



Extended Data Figure 15 | novoSpaRc successfully reconstructs a whole-kidney dataset *de novo*. a, Sketch of the major cell types 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 and bottom rows show the novoSpaRc predicted gene expression 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).



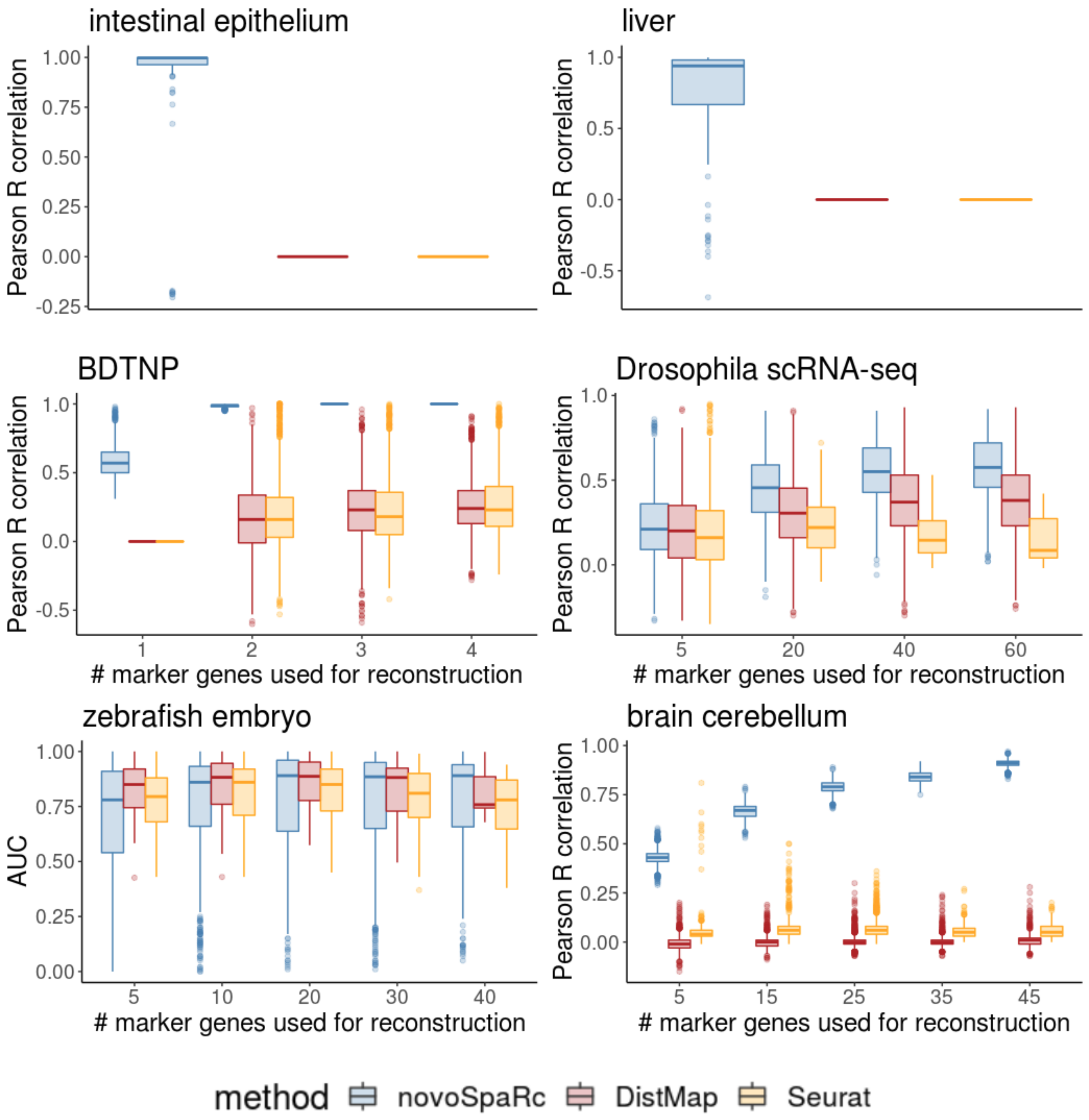
Extended Data Figure 16 | NovoSpaRc successfully reconstructs single *Drosophila* embryos. The averaged original expression of four gap genes (**a**) and four pair-rule genes (**e**) is shown for 101 and 177 individual *Drosophila* embryos, respectively. Solid line: mean; dark shadow: std; light shadow: 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. Center point, mean; error bars, SD. (**c,g**) Pearson correlation increases with the number of marker genes used by novoSpaRc for the reconstruction of the remaining genes (α=0.5) for both gap (**c**) and pair-rule genes (**g**). Using a reference atlas corresponding to the individual embryo being reconstructed ('individual atlas') achieves consistently higher reconstruction quality than using an averaged reference atlas over all embryos ('averaged atlas'). Example of the reconstruction of the expression patterns across a single random embryo, where the reconstruction of each of the four genes is performed using the three complement genes as a reference (α=0.5), for both gap (**d**) and pair-rule genes (**h**). Notice 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,c,e,g) is normalized to the maximum value over the mean expression of all embryos.



Extended Data Figure 17| Reconstruction quality varies with alpha parameter. Reconstructions of the BDTNP dataset, the *Drosophila* and zebrafish embryos and the brain cerebellum with varying number of marker genes used for the reconstruction and different values of the alpha parameter. The reconstruction quality is quantified by calculating Pearson correlations between the predicted and the original gene expression patterns for all genes that were not used as markers for the reconstruction. Reconstruction quality decreases for α=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 alpha is an interpolation parameter (defined in the section ‘Mathematical formulation of novoSpaRc’ in Methods), between using only a reference atlas (α = 1) and using both structural information (driven by the structural correspondence assumption) and a reference atlas. Center line: median; whiskers: +/-2.698SD.

	Seurat	**DistMap**	**novoSpaRc**
Spatial mapping with reference atlas	✔	✔	✔
Reference atlas can have continuous values	X	X	✔
Spatial mapping *de novo*	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 Table 1 | Comparison of spatial reconstruction with novoSpaRc with available methods that fully rely on a reference atlas. The intrinsic characteristics of novoSpaRc are compared against Seurat [9] and DistMap [18].



Extended Data Figure 18 | Comparison of spatial reconstruction with novoSpaRc with available methods that fully rely on a reference atlas. The Pearson correlation of the predicted against 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, the *Drosophila* and zebrafish embryos, and the brain cerebellum. 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 since 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 DistMap[18], and for the zebrafish embryo novoSpaRc performs comparably to or better than Seurat[9], although those methods were developed and tailored for the *Drosophila* and zebrafish embryos, respectively, and the best-performing threshold was chosen for DistMap. Center line: median; whiskers: +/-2.698SD.

Intestine: predicted by NovoSpaRc to be zonated towards the crypt

	Hist1h2ap	2810417H13Ri	Top2a	Kcne3	Cenpa	Ccdc34	Hmgb2	Impdh2	Ptma	Cdca3
Reconstructed as zonated towards the crypt [16]	X	X	X	X	X	X	X	X	X	X
Reported to be expressed in the crypt				[37]						
Reported to be overexpressed in the crypt vs the villus (in human)			[38, 39]							
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 [40].
[b] Associated with cell division
[c] Reported to regulate cell proliferation, apoptosis and migration in bladder [41].
[d] Inferred to be involved in regenerative process, proliferation, or stem cell identity [42].
[e] Gene ontology process: cell cycle and cell division [43].

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 [16]	X	X	X	X	X	X	X	X	X	X
Protein abundance was associated with V6 [16]	X			X		X	X	X	X	X
Reported to be overexpressed in the villus vs the crypt (in human)										[38, 39]

Extended Table 2 |. 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 [13]	X	X	X	X	X	X	X	X	X	X
Reported as zonated towards CV	[44-47]		[48, 49] [47]					[50, 51] [47, 49]		
Differentially methylated towards CV [47]	X		X	X				X		
Higher expression in Axin2$^+$ pericentral hepatocytes [52]		X	X	X		X				X
Additional support							[a]			

[a] A gene found to increase in liver of mice exposed to chronic hypoxia [53].

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 [13]	X	X	X	X	X	X	X	X	X	X
Reported as zonated towards PV		[47]								
Differentially methylated [47]	X	X		X		X			X	X
Lower expression in Axin2$^+$ pericentral hepatocytes [52]	X				X					
Additional support		[c]			[a]		[b]			[c]

[a] A gene found to increase in liver of mice exposed to chronic hypoxia [53].
[b] secretory antioxidase that protects against oxidative damage, whose overexpression reduced local and systemic oxidative stress generated by BDL [54].
[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 [47].

Extended Table 3 |. 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. Shapiro, E., T. Biezuner, and S. Linnarsson, *Single-cell sequencing-based technologies will revolutionize whole-organism science.* Nat Rev Genet, 2013. **14**(9): p. 618-30.
2. Wagner, A., A. Regev, and N. Yosef, *Revealing the vectors of cellular identity with single-cell genomics.* Nat Biotechnol, 2016. **34**(11): p. 1145-1160.
3. Altschuler, S.J. and L.F. Wu, *Cellular heterogeneity: do differences make a difference?* Cell, 2010. **141**(4): p. 559-63.
4. Kolodziejczyk, A.A., et al., *The technology and biology of single-cell RNA sequencing.* Mol Cell, 2015. **58**(4): p. 610-20.
5. Crosetto, N., M. Bienko, and A. van Oudenaarden, *Spatially resolved transcriptomics and beyond.* Nat Rev Genet, 2015. **16**(1): p. 57-66.
6. Lein, E., L.E. Borm, and S. Linnarsson, *The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing.* Science, 2017. **358**(6359): p. 64-69.
7. Moor, A.E. and S. Itzkovitz, *Spatial transcriptomics: paving the way for tissue-level systems biology.* Curr Opin Biotechnol, 2017. **46**: p. 126-133.
8. Chen, X., S.A. Teichmann, and K.B. Meyer, *From Tissues to Cell Types and Back: Single-Cell Gene Expression Analysis of Tissue Architecture.* Annual Review of Biomedical Data Science, 2018. **1**: p. 29-51.
9. Satija, R., et al., *Spatial reconstruction of single-cell gene expression data.* Nat Biotechnol, 2015. **33**(5): p. 495-502.
10. Achim, K., et al., *High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin.* Nat Biotechnol, 2015. **33**(5): p. 503-9.
11. Regev, A., et al., *The Human Cell Atlas.* Elife, 2017. **6**.
12. Rozenblatt-Rosen, O., et al., *The Human Cell Atlas: from vision to reality.* Nature, 2017. **550**(7677): p. 451-453.
13. 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.
14. Durruthy-Durruthy, R., et al., *Reconstruction of the mouse otocyst and early neuroblast lineage at single-cell resolution.* Cell, 2014. **157**(4): p. 964-78.
15. Waldhaus, J., R. Durruthy-Durruthy, and S. Heller, *Quantitative High-Resolution Cellular Map of the Organ of Corti.* Cell Rep, 2015. **11**(9): p. 1385-99.
16. 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.
17. Habib, N., et al., *Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons.* Science, 2016. **353**(6302): p. 925-928.
18. Karaiskos, N., et al., *The Drosophila embryo at single-cell transcriptome resolution.* Science, 2017. **358**(6360): p. 194-199.
19. *Berkeley drosophila transcription network project*. Available from: http://bdtnp.lbl.gov/Fly-Net/bioimaging.jsp.
20. Monge, G., *Memoire sur la theorie des deblais et des remblais.* Mem. de l'Ac. R. des Sc., 1781: p. 666–704.
21. Villani, C., *Topics in optimal transportation*. 2003: American Mathematical Soc.
22. Villani, C., *Optimal transport: old and new*. Vol. 338. 2008: Springer Science & Business Media.
23. Schiebinger, G., et al., *Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming.* BioRxiv, 2017: p. 191056.
24. Forrow, A., et al., *Statistical Optimal Transport via Geodesic Hubs.* arXiv preprint arXiv:1806.07348, 2018.
25. Mémoli, F., *On the use of Gromov-Hausdorff distances for shape comparison.* 2007.

26. Peyré, G., M. Cuturi, and J. Solomon. *Gromov-Wasserstein averaging of kernel and distance matrices*. in *International Conference on Machine Learning*. 2016.

27. Cuturi, M. *Sinkhorn distances: Lightspeed computation of optimal transport*. in *Advances in neural information processing systems*. 2013.

28. Petkova, M.D., et al., *Optimal decoding of information from a genetic network.* arXiv preprint arXiv:1612.08084, 2016.

29. 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.

30. Park, J., et al., *Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease.* Science, 2018. **360**(6390): p. 758-763.

31. Petkova, M.D., et al., *Optimal Decoding of Cellular Identities in a Genetic Network.* Cell, 2019. **176**(4): p. 844-855 e15.

32. 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.

33. Nusslein-Volhard, C. and E. Wieschaus, *Mutations affecting segment number and polarity in Drosophila.* Nature, 1980. **287**(5785): p. 795-801.

34. Rey, S.J. and L. Anselin, *PySAL: A Python Library of Spatial Analytical Methods.* The Review of Regional Studies, 2007. **37**(1): p. 5-27.

35. Stuart, T., et al., *Comprehensive integration of single cell data.* BioRxiv, 2018: p. 460147.

36. Tomancak, P., et al., *Global analysis of patterns of gene expression during Drosophila embryogenesis.* Genome Biol, 2007. **8**(7): p. R145.

37. 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.

38. Gassler, N., et al., *Molecular characterisation of non-absorptive and absorptive enterocytes in human small intestine.* Gut, 2006. **55**(8): p. 1084-9.

39. Olsen, L., et al., *CVD: the intestinal crypt/villus in situ hybridization database.* Bioinformatics, 2004. **20**(8): p. 1327-8.

40. Grootjans, J., et al., *Epithelial endoplasmic reticulum stress orchestrates a protective IgA response.* Science, 2019. **363**(6430): p. 993-998.

41. 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.

42. 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.

43. 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.

44. 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.

45. 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.

46. 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.

47. Brosch, M., et al., *Epigenomic map of human liver reveals principles of zonated morphogenic and metabolic control.* Nat Commun, 2018. **9**(1): p. 4150.

48. 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.

49. Braeuning, A., et al., *Differential gene expression in periportal and perivenous mouse hepatocytes.* FEBS J, 2006. **273**(22): p. 5051-61.

50. Hailfinger, S., et al., *Zonal gene expression in murine liver: lessons from tumors.* Hepatology, 2006. **43**(3): p. 407-14.

51. Gebhardt, R., *Metabolic zonation of the liver: regulation and implications for liver function.* Pharmacol Ther, 1992. **53**(3): p. 275-354.

52. Wang, B., et al., *Self-renewing diploid Axin2(+) cells fuel homeostatic renewal of the liver.* Nature, 2015. **524**(7564): p. 180-5.

53. 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.

54. Zhang, J., et al., *Protective Effects of Peroxiredoxin 4 (PRDX4) on Cholestatic Liver Injury.* Int J Mol Sci, 2018. **19**(9).