NovoSpaRc: flexible spatial reconstruction of single-cell gene expression with optimal transport

Moriel N., Senel E., Friedman N., Rajewsky N., Karaiskos N., Nitzan M.

Nature Protocols
2021 SEP ; 16(9): 4177-4200
2021 AUG 04 (first published online: final version)
doi: 10.1038/s41596-021-00573-7

URL: https://www.nature.com/articles/s41596-021-00573-7

Publisher: Nature Publishing Group (Springer Nature)

NovoSpaRc: flexible spatial reconstruction of single-cell gene expression with optimal transport

Noa Moriel[1,*], Enes Senel[2,*], Nir Friedman[1,3], Nikolaus Rajewsky[2], Nikos Karaiskos[2,#] & Mor Nitzan[1,4,5,#]

[1]School of Computer Science and Engineering, The Hebrew University of Jerusalem
[2]Systems Biology of Gene Regulatory Elements, Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany
[3]Institute of Life Sciences, The Hebrew University of Jerusalem
[4]Racah Institute of Physics, The Hebrew University of Jerusalem, Jerusalem
[5]Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem, 91904, Israel
[*]These authors contributed equally
[#]Corresponding: nikolaos.karaiskos@mdc-berlin.de, mor.nitzan@mail.huji.ac.il

Keywords: Single-cell RNA-seq, Spatial transcriptomics, Optimal transport, Spatial reconstruction

EDITORIAL SUMMARY This protocol describes novoSpaRc, a computational pipeline for *de novo* reconstruction of spatial gene expression from single-cell RNA sequencing with the potential to incorporate spatial atlas data to improve the reconstruction.

Abstract

Single-cell RNA-sequencing technologies have revolutionized modern biomedical sciences. A fundamental challenge is to incorporate spatial information to study tissue organization and spatial gene expression patterns. Here, we describe a detailed protocol for using novoSpaRc, a computational framework that probabilistically assigns cells to tissue locations. At the core of this framework lies a structural correspondence hypothesis, that cells in physical proximity share similar gene expression profiles. Given scRNA-seq data, novoSpaRc spatially reconstructs tissues based on this hypothesis, and optionally, by including a reference atlas of marker genes to improve reconstruction. We describe the novoSpaRc algorithm, and its implementation in an open-source Python package (github.com/rajewsky-lab/novosparc). NovoSpaRc maps a scRNA-seq dataset of 10,000 cells onto 1,000 locations in under 5 minutes. We describe results obtained using novoSpaRc to reconstruct the mouse organ of Corti *de novo* based on the structural correspondence assumption, the human osteosarcoma cultured cells based on marker gene information, and provide a step-by-step guide to *Drosophila* embryo reconstruction in the Procedure to demonstrate how these two strategies can be combined.

Introduction

The emergence of single-cell RNA sequencing (scRNA-seq) technologies during the past decade has transformed the biomedical sciences [1,2]. High-throughput methods have enabled the simultaneous profiling of tens of thousands of cellular transcriptomes stemming from the same tissue [3,4], and have been successfully employed throughout multiple discoveries, such as to dissect tissue heterogeneity [5,6], to identify rare cell populations [5,7,8], and to investigate cell states [5,9] and cell differentiation processes [10,11] among others.

Most scRNA-seq methods, however, require dissociation of the tissue, which results in the loss of spatial information. The physical context of the cells is vital for the understanding of biological functions at the global collective scale, such as spatial gene expression patterns [12–15], the organization of cell types in space [8,16,17], as well as heterogeneous responses to perturbations or drug responses throughout diseased tissues [18]. At the local level, spatial information is critical to thoroughly study cell-cell interactions and individual cellular states [19].

A growing number of experimental techniques that preserve spatial information have been developed over the past few years to bridge this gap[20]. While these techniques are generally still at least partially limited in throughput[14,16,17,21,22] spatial resolution[23] and commercially available solutions are often costly and do not offer single-cell resolution [24–26], experimental techniques are constantly diversifying, advancing and improving[27]. However, there is an urgent need to decipher spatial information from the vast single-cell data that already exists. Furthermore, there is a need to leverage the expanding set of high-quality spatial transcriptomic experiments as complementary information for scRNA-seq data and learn how to efficiently integrate these two sources of information.

The challenge of reconstructing spatial gene expression from single-cell data is tackled by multiple computational techniques that require the existence of a spatial atlas of marker genes to be used as a reference guide. Such reference atlas is generally only feasible for stereotypical tissues with robust, relatively simple spatial expression patterns (which can repeat across multiple subunits within the tissue), such as liver lobules, the intestinal epithelium, and some embryos at early developmental stages. In addition, such reference atlas may not be straightforward to construct [28–32]. Recently, we presented novoSpaRc[33], a new computational approach that can spatially reconstruct gene expression without the need of a reference atlas, while being able to incorporate it and enhance performance if such an atlas exists.

NovoSpaRc is based on the hypothesis that physically neighboring cells share similar transcriptional profiles, so that gene expression, on average, does not change abruptly but in a continuous manner for a substantial subset of genes. We formulated this hypothesis within the framework of optimal transport[34,35] (OT), which allows us to probabilistically assign single cells to tissue locations by interpolating between the continuity assumption and other types of prior experimental data, such as the spatial expression of a subset of marker genes (or a reference atlas), the local density of cells in the tissue, and the technical quality of read measurements extracted from single cells. In this manuscript, we provide detailed guidelines for using novoSpaRc to recover the spatial organization of cells and genes in their tissue-of-origin based on single-cell data.

Overview of the algorithm and workflow

The main objective of novoSpaRc is to probabilistically map single cells onto the tissue's physical structure, and infer gene expression patterns across the tissue. To do that, novoSpaRc requires a **gene expression matrix** and a **target space** (coordinates of the physical space). **Atlas expression**, that is, spatial expression of a subset of genes over the tissue, is an additional optional input. Using these inputs, novoSpaRc computes three cost matrices, which together allow us to interpolate between minimizing the deviation of a certain mapping from a structural correspondence assumption between distances of cells in gene expression and physical space, and from a potentially available reference atlas. NovoSpaRc outputs a **transport matrix**, a probabilistic mapping of cells onto the target space locations, using the OT framework, and computes the inferred **spatial gene expression** over the target space.

The workflow is schematically represented in Fig. 1 along a detailed description below of each of these steps, the inputs and outputs of novoSpaRc, and optional validations and follow-up analyses.

Input cells and locations descriptions to construct Tissue object (Steps 1-6)

Cell expression

The main input to the novoSpaRc algorithm is a gene expression matrix that captures single-cell gene expression levels within a population of cells. Cell-by-gene matrices where each entry is the count of RNA molecules retrieved from scRNA-seq are a typical input. However, outputs of other experimental procedures that quantify gene expression levels can be integrated as well, such as using RNA quantization through amplification rounds[36](Fig. 2), fluorescent imaging[14,16,17](Fig. 3), or other sequencing techniques [23,37,38](demonstrated for Slide-seq data in[33]).

Preprocessing of the gene expression matrix can be minimal, such as the standard library-normalization scheme of cell-count normalization and log transformation for scRNA-seq data [39]. Since scRNA-seq protocols suffer from low capture probabilities and expression representation is redundant and extensive in dimensions (e.g. ~20K genes), using a meaningful low-dimensional representation of expression such as a highly variable set of genes or a latent representation of expression (e.g. PCA) can drastically enhance the quality and runtime of reconstruction.

Target space

A target space is a set of coordinates corresponding to the physical locations across the tissue onto which novoSpaRc maps the single cells. The set of locations can span any 1D, 2D, or 3D structure corresponding either to the explicit tissue structure or a representation that captures the structure's inherent spatial symmetries. For optimal reconstruction results, the shape of the target space should resemble the shape (or underlying symmetries) of the tissue-of-origin, as the inherent coordinate relationships will be used for the spatial reconstruction. Note that while a faithful representation of the tissue-of-origin shape or symmetries is ideal, simpler target spaces or ones that only capture local structures and symmetries of the tissue are many times sufficient.

There are two ways to create the target space if no prior reference is available. The most straightforward way is to use novoSpaRc's internal functions and create a basic shape target space, e.g. rectangle, circle, sphere, prism, etc. If we are interested in reconstructing spatial variability along a single axis, for example, such as that corresponding to a one-dimensional gradient of oxygen or morphogenes in the biological system, we should use a linear target space. For example, in the organ of Corti, a 2D spiral organ essential for hearing, gene expression within cell subpopulations mainly varies along a 1D apex-to-base axis. By constructing a corresponding 1D target space, we illustrate novoSpaRc expression reconstruction along this axis (Fig. 2). Additionally, a target space can be created from experimental measurements. Representative images can be processed to determine cell locations (along with their corresponding information for gene expression). This is illustrated in the expression reconstruction of human osteosarcoma cultured cells where cell locations are deduced from microscope imaging obtained using MERFISH (Fig. 3). A 3-dimensional analogue of this case is illustrated in the reconstruction of the *Drosophila* embryo described in the Procedure Section.

(Optional) atlas expression

A reference atlas is an optional input, carrying information about the expression levels of a subset of genes across the target space. Such a reference atlas can be incorporated into novoSpaRc and increase the reconstruction quality by essentially restricting the space of possible reconstruction solutions to those that are consistent with the atlas, or by spatially regulating the mapping process. The reference atlas can guide the selection of the target space. For example, if marker gene expression is measured using *in situ* imaging at single-cell resolution, then we can set the target locations at the cells' centroid locations. To account for atlas information at lower resolution, such as retrieved experimentally from bulk sequencing of sectioned tissue [40–42], or from computational local aggregation of nearby cells due to low signal[23], spatial expression of genes is binned and integrated (e.g. averaged) in order to provide expression over the set of target locations.

In general, there are no special requirements or restrictions regarding the data format, number of genes, and experimental method used to construct the reference atlas. However, reconstruction is likely to benefit from a reference atlas quantifying the expression of spatially informative genes.

Given the target space locations, cellular gene expression, and optionally the reference atlas of spatial expression, we construct a Tissue object, the main object of the novoSpaRc package.

Compute cost matrices (Steps 7-8)

Having the normalized gene expression matrix and the target space at our disposal, and potentially a reference atlas, we continue with computing the cost matrices that are needed for performing the spatial reconstruction.

Computing the cell-cell and location-location cost matrices

The cell-cell cost matrix summarizes the distances between cells in gene expression space, and the location-location cost matrix summarizes the physical distances between locations in the target space. The assumption at the heart of novoSpaRc states that there is a correspondence between the structure of locations in physical space and the structure of cells in gene expression space potentially along a low-dimensional nonlinear manifold. More concretely, it implies that there is a correspondence between pairwise distances of locations in physical space and cells in gene expression space. To capture distances along low-dimensional structures, we construct k-Nearest Neighbors (kNN) graphs (based on Euclidean distances) in physical space and in gene expression space. The corresponding cost matrices are comprised of pairwise distances between cells and locations, computed as the shortest paths along the corresponding kNN graphs. These cost matrices would be used to capture the essence of the structural correspondence assumption, that is, the averaged transcriptional similarity among physically proximal cells.

Computing the reference atlas cost matrix

If a reference atlas is available for a subset of genes, the corresponding cost matrix captures the discrepancy between the expression of these genes in each cell and in each location of the target space. Specifically, we compute the Euclidean distance across the subset of genes composing the reference atlas between the cells and locations.

Compute optimal transport of cells to locations and predict expression over target space (Step 9)

Setting marginal distributions

Here we set the marginal distributions for both the cells and locations. By default, novoSpaRc initializes the marginal distributions to be uniform. This means that the total spatial mapping probability associated with each cell is the same, and the total mapping probability associated with each location is the same. In cases where non-uniform mapping is desired, where prior biological knowledge exists for the physical density of cells, or varying technical quality of cells, this can be readily incorporated at this step.

Setting the alpha parameter

The alpha parameter is used to interpolate between two modes of reconstruction: (1) a *de novo* spatial reconstruction ($\alpha = 0$), based only on the underlying structural correspondence assumption, and (2) a reconstruction based only on the information provided by the reference atlas for the spatial expression of a set of marker genes ($\alpha = 1$). Intermediate values take into account both objectives during the reconstruction. Values closer to $\alpha = 1$ reflect higher confidence in the reference atlas. For example, a reference atlas corresponding to a single-cell expression sample of high quality, and composed of a large number of marker genes, would be expected to generate highly-informative cell-to-location distances, as shown for the MERFISH data (Fig. 3). Values closer to $\alpha = 0$ reflect higher confidence in the structural correspondence assumption, and when spatial smoothness in gene expression is expected, for example due to gradients of oxygen or nutrients, or due to physical progression of the cells, such as in the case of the crypt-to-villus axis in the intestinal epithelium[12,33].

Computing the transport matrix

NovoSpaRc then computes a locally-optimal transport matrix that probabilistically assigns single cells to locations through the extended framework of optimal transport as described below.

Predicting expression over target space

Given single-cell gene expression (as input) and the transport matrix (assigning each of these cells a probability distribution over tissue locations), we can compute the resulting full gene expression expected at each of the locations in the tissue-of-origin. Both the optimal mapping of cells to locations and the predicted expression can be fetched after reconstruction.

Validation of results and follow-up analyses (Steps 10-14)

Expression cross-validation

If available, high spatial correlation between the expression of genes from the reference atlas and their predicted expression indicates successful reconstruction, and therefore can be used to cross-validate the reconstruction (see Fig. 3b). Based on such quantitative evaluation, the algorithm's parameters (such as α) can be tuned and selected. In addition, displaying genes that are expected to be spatially informative (see below) can assist in qualitatively evaluating the reconstruction (as displayed in Step 10 in the Procedure Section).

Localized probabilistic mapping

With novoSpaRc, we recover a probabilistic mapping of cells over locations in the tissue-of-origin (as opposed to a discrete one-to-one mapping), which has several advantages: (1) probabilistic mapping tends to be more robust in cases where the data is noisy and sparse, as in scRNA-seq data, and naturally expresses uncertainty when there is not enough information to pinpoint the exact location of a cell; (2) When locations correspond to an experimentally-acquired reference atlas, it is possible for information of cellular gene expression to be distributed over multiple locations, due to the experimental setup (e.g. a cell's expression may be splitted between adjacent beads in Slide-seq[23,27]); (3) probabilistic mapping is computationally advantageous in the context of entropically regularized optimal transport (see Mathematical Formulation of novoSparc section). Yet, in general we expect a biologically-meaningful mapping to reflect a relatively localized mapping of each of the cells in physical space (see Step 11 in the Procedure Section).

Self-consistency validation

While novoSpaRc probabilistically assigns single cells to tissue locations, the algorithm itself is deterministic in the sense that given a certain input and an initialization of the transport matrix, the algorithm will always converge on the same solution. Varying mildly the transport matrix initialization, the input by subsampling the cells or changing the gene selection, and the algorithm parameters, can aid in assessing the robustness of the reconstruction and optimizing parameter selection (as shown in Step 12 in the Procedure Section).

Verify and identify spatially informative genes

Identifying spatially informative genes, or genes whose expression varies in a meaningful, non-random, pattern across the tissue, is a complex task since meaningful spatial expression patterns can be of diverse forms. Here we use a measure for global spatial auto-correlation, Moran's I, to rank genes as spatially informative. The Moran's I score for a gene with spatial expression x, within a user-defined neighborhood W specifying the relation between location pairs i, j in W_{ij} (we set k-nearest neighboring locations to be the neighborhood of each location with equal contribution) is expressed as follows:

$$I = \frac{m}{s}\frac{\sum_{ij} z_i W_{ij} z_j}{\sum_i {z_i}^2},$$

where $z_i = (x_i - \bar{x})$, $\bar{x}$ is the mean expression, m is the number of locations, and $s = \sum_{ij} W_{ij}$. We compute the Moran's I score and its corresponding one-tailed p-value under normality assumption. A gene whose expression is significantly correlated among neighboring tissue locations (e.g. constant expression over a large region), is considered to be a spatially informative gene. Step 13 in the Procedure Section demonstrates differences in Moran's I values for genes considered to be spatially informative (genes assayed for a reference atlas) and spatially disordered genes.

Find dominating spatial archetypes

To discover the major gene expression patterns over a tissue, we hierarchically cluster the predicted expression of spatially informative genes and extract the averaged spatial expression of main branches within the reconstructed tree, yielding a set of differential and dominant pattern schemes (as shown in Step 14 in the Procedure Section).

Mathematical formulation of novoSpaRc

NovoSpaRc attempts to find a transport matrix $T \in C_{p_{cell},p_{loc}}$, or probabilistic mapping between n cells and m locations. To enforce assignment of all cells over all locations, we are looking for a transport matrix constrained by the marginal cell and location distributions, $p_{cell} \in [0,1]^{n\times 1}, p_{loc} \in [0,1]^{1\times m}$,

$$T^* = argmin_{T \in C_{p_{cell},p_{loc}}} (1-\alpha)\, C^{smooth}(T) + \alpha\, C^{atlas}(T) - \epsilon H(T),$$

where $C_{p_{cell},p_{loc}} = \{T \mid T \in [0,1]^{n\times m}, T1 = p_{cell}, 1T = p_{loc}\}$, $1p_{cell} = 1, p_{loc}1 = 1$. Here, without any additional prior knowledge, we assume uniform marginal distributions for the cells and locations, that is $(p_{cell})_i = 1/n, (p_{loc})_j = 1/m$, for cell i and location j.

We seek to find such T that minimizes the objective composed of the following three terms described below.

The first term is related to the structural correspondence assumption, aiming to minimize the Gromov-Wasserstein discrepancy[43,44] between pairwise distances of cells in gene expression space, $D^{exp} \in R_{+}^{n\times n}$, and pairwise distances of locations in physical space, $D^{phys} \in R_{+}^{m\times m}$, weighted by the transport matrix:

$$C^{smooth}(T) = \sum_{cells\ i,j\ locations\ k,l} L\left(D_{ij}^{exp}, D_{kl}^{phys}\right) T_{ik} T_{jl},$$

where L is a loss function. Here, we use the quadratic loss $L(a,b) = \frac{1}{2}(|a-b|)^2$. The pairwise distances between cells, D^{exp}, and locations, $D^{phys}$are computed as the shortest path distances over the respective kNN graphs constructed over the cells in gene expression space and locations in physical space based on Euclidean distance.

The second term is related to a potentially available reference atlas, aiming to minimize the discrepancy between the reconstructed spatial gene expression and the spatial expression registered by the atlas for the subset of genes it contains,

$$C^{atlas}(T) = \sum_{cell\ i\ location\ k} D_{ik}^{exp,phys} T_{ik},$$

where we use the Euclidean norm to quantify the discrepancy between the expression levels of genes in each of the locations (according to the atlas) and their expression in the set of single cells, given by $D^{exp,phys} \in R_{+}^{n\times m}$.

The algorithm is also compatible for adjusted measures of cell-to-cell expression distances, D^{exp}, physical location distances, D^{phys}, and atlas-based, cell-to-location discrepancy, $D^{exp,phys}$.

The last term is an entropic regularization term that promotes a disperse mapping,

$$H(T) = -\sum_{cell\ i\ location\ k} T_{ik}\ log\ (T_{ik}),$$

The coefficient α controls the interpolation between the structural correspondence objective (C^{smooth}) and the reference atlas discrepancy objective (C^{atlas}). ϵ sets the weight of the entropy regularization. Higher ϵ values drive the solution towards a higher-entropy T (e.g. for uniform marginal distributions, when this term dominates the objective, we would expect to converge onto a nearly-uniform T), while lower ϵ values result in more localized T.

Using this setup, a (locally) optimal mapping, T, can be derived with alternating iterations of projected exponential gradient descent, by minimizing the overall objective and using Kullback-Leibler (KL) projection to constrain the solution to the subspace of transport matrices, $C_{p_{cell}, p_{loc}}$. The integration of the entropy term reduces this process to efficient iterations of Sinkhorn's fixed point algorithm[33].

After obtaining an optimal mapping T^* (`tissue.gw`), the spatial gene expression matrix $S \in R^{g \times m}$, or the expression of each gene for each location in the target space (`tissue.sdge`), can be recovered by multiplying $S = X^T T^*$, for the original gene expression matrix $X \in R^{n \times g}$.

For further mathematical details and full description of the algorithmic approach for computing the novoSpaRc transport matrix, please refer to [33].

Comparison of novoSpaRc to existing baselines

Two seminal papers[29,45] proposed the first computational approaches, scoring the correspondence of cells to locations according to their agreement with a coupled reference atlas and using these scores to infer spatial gene expression. This methodology has been deployed and extended in various biological contexts (e.g.[12,30–33,36,46–48]).

In contrast to these methods, novoSpaRc can utilize the geometric structure of cells in gene expression space compared to the structure of locations of the target space, which relaxes the strong dependency on a reference atlas, yet enables its integration when it is available. Additionally, the novoSpaRc framework is rich in interpretable parameters, such as the effective neighborhood size and cell densities, and therefore, they can often be estimated and leveraged as priors through analysis of the data.

After novoSpaRc was published, additional spatial reconstruction methods were proposed which integrate priors including cell density[32], and ligand-receptor communication[49,50].

A detailed comparison of several spatial reconstruction methods is presented in Table 1.

Limitations of novoSpaRc

Recovering gene expression over a tissue from a cell by gene matrix, with or without using prior information, is challenging. In fact, there are many parameters that limit the success of any spatial reconstruction method, such as the intrinsic stochasticity of expression, the reproducibility of spatial gene expression across multiple tissues of the same type, experimental noise, and the correspondence between scRNA-seq data and a reference atlas. Beyond the quality of expression information, when there is no “anchor” to break symmetries of the target space (like a reference atlas), global transformations have to be considered for any *de novo* reconstruction. While novoSpaRc’s flexibility and probabilistic nature mediate such difficulties, they still pose a challenge that is shared across most spatial reconstruction methods.

Specifically concerning novoSpaRc’s assumptions, the structural correspondence assumption can reconstruct a tissue’s spatial expression solely from individual cells’ expression besides complementing an existing reference atlas. Transcription driven by spatial signals, such as metabolites gradients, and long-range cell-to-cell communication, as well as the absence of cell migration, can boost this assumption as these increase local expression similarities. However, components of expression that result from processes detached from the physical space, spatial combination of multiple cell types, and disorder such as that associated with cancerous conditions, challenge this assumption. Additional steps including conducting separate reconstructions for cells of distinguished cell types (e.g. handling subpopulations of the organ of Corti separately, Fig. 2b-c), accounting for specific spatially informative genes (e.g. using apex-base differentially expressed genes to recover organization in the organ of Corti, Fig. 2d), averaging expression (e.g. cerebellum Slide-seq beads aggregation[33]), and collapsing the physical space to capture inherent expression symmetries (e.g. liver lobule[33]), can also increase the correspondence between physical and expression distances.

When a reference atlas is used for reconstruction, the quality of reconstruction depends on the quality of the reference atlas, and the level of correspondence between it and the single-cell data.

Applications of novoSpaRc

We previously applied novoSpaRc to a variety of tissues, including the mammalian liver, intestinal epithelium, whole-kidney, and sections of brain cerebellum, as well as *Drosophila* and zebrafish embryos [33]. Here, we describe how novoSpaRc successfully reconstructs two additional tissues: (i) the base-to-apex organization of the organ of Corti[47] *de novo*, using the structural correspondence assumption, and (ii) the spatial locations of osteosarcoma cultured cells[14] by using only marker gene information, as the signal of local expression similarity is expected to be weak. Notebooks for reconstructing both examples are available in the github repository. We additionally describe in a step-by-step fashion how to interpolate between the structural correspondence assumption and marker gene information by reconstructing the *Drosophila* embryo[30,51] and use whole-kidney scRNA-seq dataset[52] to benchmark the runtime of novoSpaRc (Fig. 4).

De novo spatial reconstruction of the organ of corti

The organ of Corti, the receptor organ for hearing, contains multiple layers of cell types, extending spirally from base to apex (Fig. 2a). In[47] the organ of Corti is first dissected into apical and basal halves providing a ground truth of the spatial membership of the cells, followed by dissociation to individual cells and quantitative reverse transcription polymerase chain reaction (qRT-PCR) for measuring RNA expression levels of 192 genes. Using the resulting gene expression matrix, we map these cells to a finer linear grid without a reference atlas (*de novo*). We find that cells that originated from either the apex or the base were mapped towards opposite ends of the linear target space (shown in Fig. 2b for outer hair cells (OHC)). In addition, the mapping recapitulates the monotonic expression gradient of genes known to be zonated towards either the base or the apex (Fig. 2c). These results suggest a correspondence between the reconstructed linear grid and the original base-apex axis. Moreover, for each of the cell types in the data, we manage to recover gradual mapping towards opposite ends for apex- and base-originating cells with *de novo* reconstruction using their differentially expressed genes (Fig. 2d).

Reconstruction of spatially disordered expression in cultured osteosarcoma cells

To show how novoSpaRc can be used to reconstruct spatially disorganized expression, we examine a dataset of osteosarcoma cells that were cultured and assayed using a multiplexed imaging method, MERFISH [14] (Fig. 3a). For direct comparison to a ground-truth spatial gene expression, we synthetically dissociate the original spatially-informed MERFISH data to individual cells and use novoSpaRc to infer their original locations. Unlike a tissue, physically proximal cells are not generally expected to exhibit transcriptional similarity. Nevertheless, employing novoSpaRc and using a few randomly selected marker genes as a reference atlas, we manage to reconstruct spatial gene expression for all genes (Fig. 3b). We visualize the increase in the quality of reconstruction as more marker genes are employed, with the recovery of microenvironments, marked with fibroblast growth factor, FGF18 (Fig. 3c).

Step-by-step spatial reconstruction of the *Drosophila* embryo

In the procedure we describe the steps used for reconstructing the spatial organization of the *Drosophila* embryo given scRNA-seq data[30]. The available reference atlas for the *Drosophila* embryo encompasses the expression of 84 transcription factors measured in 3,039 cells using FISH (http://www.cb.uu.se/~cris/BDTNP_Imaging.html)[51]. Throughout the Procedure section, we analyze our underlying assumptions using the extensive reference atlas, discuss the selection of key parameters, evaluate the reconstruction, and discuss potential extensions of the analysis.

Materials

Equipment

Software

- Python version 3.5 or later and the standard Python installation package, pip (https://pip.pypa.io/en/stable/installing/)
- NovoSpaRc package (https://pypi.org/project/novosparc/)

Hardware

- 32- or 64- bit computer running Linux, Windows or Mac OS X
- >= 4GB of RAM
- Internet connection is required for downloading and installing novoSpaRc from PyPi

Data

- A single-cell gene expression matrix is needed as input for novoSpaRc.
- A target space defining the locations within the tissue and atlas gene expression describing expression of certain genes over the target space are optional.

Example data

In the procedure we go through reconstruction of *Drosophila* embryo scRNA-seq expression [30]. As a reference atlas, we use the previously constructed *Drosophila* embryo where expression of 84 transcription factors was quantitatively registered for individual cells based on FISH imaging (http://www.cb.uu.se/~cris/BDTNP_Imaging.html)[51]. A tutorial demonstrating the use of novoSpaRc to reconstruct single-cell gene expression of the Drosophila scRNA-seq dataset is available at:
https://github.com/rajewsky-lab/novosparc/blob/master/reconstruct_drosophila_embryo_tutorial.ipynb

Equipment Setup

Installation

To install novoSpaRc, run:

```
pip install novosparc
```

Pip retrieves the latest novoSpaRc version from PyPy, as well as the various dependencies that are required. To avoid issues with package dependencies, we recommend the use of an isolated Python environment. This can be accomplished via conda (https://docs.conda.io/en/latest/) or pipenv (https://github.com/pypa/pipenv) combined with virtualenv (https://pypi.org/project/virtualenv/) for an isolated python environment.

Imports

Import novoSpaRc along with other packages, and their abbreviations used in this tutorial:

```
# imports
import novosparc

import os
import numpy as np
import pandas as pd
import scanpy as sc
import matplotlib.pyplot as plt
import altair as alt
from scipy.spatial.distance import cdist, squareform, pdist
from scipy.stats import ks_2samp
from scipy.stats import pearsonr

import random
random.seed(0)
```

Procedure

Input cells and locations descriptions to construct Tissue object Timing: 1 second

1. Read gene expression data.
   Here, we use *Drosophila* embryo scRNA-seq expression data[30] to demonstrate the reconstruction process. Start by reading the data into Scanpy's [53] AnnData format (see https://anndata.readthedocs.io/en/stable/anndata.AnnData.html for details):

```
# reading expression data to scanpy AnnData (cells x genes)
data_dir = 'novosparc/datasets/drosophila_scRNAseq/'
data_path = os.path.join(data_dir, 'dge_normalized.txt')
dataset = sc.read(data_path).T
gene_names = dataset.var.index.tolist()

num_cells, num_genes = dataset.shape # 1297 cells x 8924 genes
```

2. Preprocess data.

In this example, the data is saved after preprocessing. In case of an unprocessed count matrix, however, standard[39], or data-tailored preprocessing is recommended. Since novoSpaRc reads the dataset as a Scanpy AnnData object, we can apply standardized preprocessing steps. For example:

```
# preprocess
sc.pp.normalize_per_cell(dataset)
sc.pp.log1p(dataset)
```

After preprocessing, it is worth observing the data size before proceeding. Potentially, we can subsample the number of cells to shorten runtimes and to assess robustness of reconstruction by using different subsets of cells:

```
# optional: subset cells
num_cells = 1000
sc.pp.subsample(dataset, n_obs=num_cells)
```

3. (Optional) Generate low dimensional representation of the data.
   To reduce the noise, capture meaningful expression distances between cells and reduce runtimes, it is advisable to decrease the dimension of gene expression data. This can be done by subsetting the gene expression matrix for highly variable genes or using PCA representation. An order of hundreds (~500-1,000) of highly variable genes, or tens (~50) of PCs are usually a good choice for our runs.

```
# optional: generating a lower representation of expression
dge_rep = None # a representation of cells gene expression
sc.pp.highly_variable_genes(dataset)
is_var_gene = dataset.var['highly_variable']
var_genes = list(is_var_gene.index[is_var_gene])

# alternative A: variable expressed genes representation
dge_rep = dataset.to_df()[var_genes]

# alternative B: pca representation
n_comps = 50
sc.pp.pca(dataset, n_comps=n_comps)
dge_rep = pd.DataFrame(dataset.obsm['X_pca'])
```

4. Create a target space.
   We provide three alternatives for determining the locations of the target space, depending on the existing knowledge about the shape of the tissue (Fig. 5a-c).

   A. **Using a reference atlas**
   i. If a reference atlas is used, then the target space consists of its set locations (Fig. 5a). Here we use the previously constructed virtual *Drosophila* embryo where expression of 84 transcription factors was quantitatively registered for individual cells from FISH imaging (http://www.cb.uu.se/~cris/BDTNP_Imaging.html). Load available target space as follows:

```
# alternative A: target space available apriori
atlas_dir = 'novosparc/datasets/bdtnp/'
target_space_path = os.path.join(atlas_dir, 'geometry.txt')
locations = pd.read_csv(target_space_path, sep=' ')
num_locations = 3039
locations = locations[:num_locations][['xcoord', 'zcoord']].values
```

   B. **Using a prior shape without exact locations**
   i. In cases where we know the general shape of the tissue but we are missing specific locations of the cells, the target space can be generated from a binary image (Fig. 5b). Coordinates of every black pixel within the input image are set as a target location. To decrease the resolution of spatial expression, we can also subsample the locations of the target space (e.g. here 3039 locations are sampled). Generate locations from a binary image as follows:

```
# alternative B: prior shape without exact locations
tissue_path = 'novosparc/datasets/tissue_example.png'
locations = novosparc.gm.create_target_space_from_image(tissue_path)
locations = locations[np.random.choice(locations.shape[0], num_locations), :]
```

C. **No prior knowledge of the target space**

i. NovoSpaRc can create a target space by setting locations on certain basic shapes. Current supported shapes include filled circle, 2D torus projection, rectangular grid, sphere, and torus. Grids can be populated with equidistant or randomly drawn points. Generate a circle filled with equidistant grid as the target space as follows (Fig. 5c):

```
# alternative C: no prior knowledge of target space
locations = novosparc.gm.construct_circle(num_locations=num_locations)
```

Setting target space dimension and collapsing symmetries: although tissues are generally 3D objects (or 2D for a monolayer of cells), there is often a lower dimension subspace that dominates the expression variation. Therefore, to characterize the expression across a tissue, we sometimes prefer to construct a target space of lower dimension. In addition, we may want to collapse symmetrical regions. Here, for example, the *Drosophila* embryo is represented by a 2D projection, and instead of inferring the spatial expression of both sides of the embryo, we make use of the bilateral symmetry and map cells to a single side.

Setting `num_locations`: the number of locations to map cells to, `num_locations`, sets the location density of the tissue. While the number of locations should be upper bound by the number of cells, to increase the robustness of the reconstruction and to identify low-resolution spatial patterns, we often choose a smaller `num_locations` value (for example, to identify 1-dimensional monotonic gene expression gradients it may be sufficient to set `num_locations` to be in the order of 10, as shown for the reconstruction of the liver lobule and intestinal epithelium[33]).

5. (Optional) Read atlas expression. Available reference marker genes can be used to achieve a better reconstruction (Fig. 6). We deploy here the measurements of 84 transcription factors reported from FISH imaging (http://www.cb.uu.se/~cris/BDTNP_Imaging.html) (Fig. 6a). Read the atlas into a Scanpy AnnData object (`atlas`), where columns correspond to the reference atlas genes and rows to positions of the target space locations. If the loaded atlas locations were subset in Step 4, we select the same locations (or rows) here:

```
# reading reference atlas
atlas_path = os.path.join(atlas_dir, 'dge.txt')
atlas = sc.read(atlas_path)
atlas_genes = atlas.var.index.tolist()
atlas.obsm['spatial'] = locations
```

```
pl_genes = ['sna', 'ken', 'eve']
novosparc.pl.embedding(atlas, pl_genes)
```

Expression can also be potentially inferred over a configured target space from images or from previously binned spatial expression, for instance, using interpolation tools (e.g. using `scipy.interpolate` package).

If an atlas is provided, then one can test to what extent the structural correspondence assumption holds over the atlas genes. A proxy for this test is to examine if the expression distances versus their physical distances indeed exhibit an increasing monotonic relationship (code below, Fig. 7a), and that this monotonicity persists when comparing the expression and location cost values (as computed in Steps 7-8 and shown in[33]).

```
# tip: visualizing loc-loc expression distances vs their physical distances
novosparc.pl.plot_exp_loc_dists(atlas.X, locations)
```

Optionally, examine how "spatially informative" the marker genes are, use Moran's I measure for spatial auto-correlation as follows (Fig. 7b):

```
# tip: testing how spatially informative are the atlas' marker genes
mI, pvals = novosparc.an.get_moran_pvals(atlas.X, locations)
df = pd.DataFrame({'moransI': mI, 'pval': pvals}, index=atlas_genes)

gene_max_mI = df['moransI'].idxmax()
gene_min_mI = df['moransI'].idxmin()


novosparc.pl.embedding(atlas, [gene_max_mI, gene_min_mI])
```

6. Construct a Tissue object.
   So far, we constructed the input for the spatial reconstruction. Next, initialize a Tissue object with the cell expression dataset in the form of a Scanpy AnnData object and the target space locations as a two-dimensional numpy ndarray of shape `num_locations x dimensions` (in case the tissue is one-dimensional, `locations` should still be two-dimensional):

```
# construct Tissue object
tissue = novosparc.cm.Tissue(dataset=dataset, locations=locations)
```

Compute cost matrices Timing: 5 seconds

7. Set parameters for cost matrices.
   The OT framework interpolates between two objectives, optimizing structural correspondence and minimizing atlas-based discrepancy.

   First set the parameters for each objective. For the structural correspondence (smoothness) objective, set the number of neighbors to use for constructing the kNN graphs over cells and locations.

   If using the atlas (linear) objective, then novoSpaRc requires a two-dimensional numpy ndarray, `atlas_matrix`, of shape `num_locations x num_markers`, and a one-dimensional numpy array of the corresponding indices of the marker genes in the expression matrix, `markers_to_use`:

```
# params for smooth cost
num_neighbors_s = num_neighbors_t = 5

# params for linear cost
markers = list(set(atlas_genes).intersection(gene_names))
num_markers = len(markers)
atlas_matrix = atlas.to_df()[markers].values
markers_idx = pd.DataFrame({'markers_idx': np.arange(num_genes)},
                           index=gene_names)
markers_to_use = np.concatenate(markers_idx.loc[markers].values)
```

   Setting `num_neighbors_s, num_neighbors_t:` Assume there is an expression "niche", or a microenvironment, around a cell where expression differences within the niche are subtle. Then, `num_neighbors_s`, the number of neighbors to consider for the cell nearest neighbor graph should correspond to the cell radius of this niche. Therefore, we often choose `num_neighbors_s` to be the number of immediate neighbors in space, depending on the grid's dimensionality (e.g. for 1D, ~2 neighbors, for 2D, ~5-8 neighbors). The optimal `num_neighbors_t` value can be tuned depending on the overall number of cells, level of noise, and the dimensionality of the grid the cells are mapped to. Empirically, we found that `num_neighbors_t` values between 3-15 yield robust spatial embeddings.

8. Compute cost matrices.
   We provide several options for computing the cost matrices:

   A. **Handle both objectives together**
   i. If we wish to use all data (e.g. all genes, cells and locations) and compute the corresponding cost matrices, then run:

```
# alternative A: setup both assumptions
tissue.setup_reconstruction(atlas_matrix=atlas_matrix,
                                 markers_to_use=markers_to_use,
                                 num_neighbors_s=num_neighbors_s,
                                 num_neighbors_t=num_neighbors_t)
```

B. **Handle each objective separately**
i. Each of the corresponding cost matrices should be set separately if: (1) we want to use a lower dimensional representation of expression for the structural correspondence objective, such as the PCA representation of the expression matrix (saved in `dge_rep` as a numpy array of shape `num_cells x n_comps`), or (2) we use only one of the objectives. Set the corresponding cost matrices as follows:

```
# alternative B: handling each assumption separately
tissue.setup_smooth_costs(dge_rep=dge_rep)
tissue.setup_linear_cost(markers_to_use, atlas_matrix)
```

C. **Directly set cost matrices**
i. There are cases where the default procedure for deriving a cost matrix is less suitable. For example, when there are very few cells or locations, we may want to skip the kNN graph construction in order to preserve the complete distance information. In that case, we can compute and feed the cost matrices to the Tissue object directly as numpy arrays of shapes `(num_cells x num_locations)`, `(num_cells x num_cells)`, `(num_locations x num_locations)`, for `markers_cost`, `exp_cost`, `loc_cost`, respectively. Normalize all cost matrices (e.g. divide by the maximum value) and, for locations and expression cost matrices also center the distances (e.g. subtract the mean) for comparable scaling:

```
# alternative C: directly set cost matrices.
tissue.costs[‘markers’] = markers_cost
tissue.costs[‘expression’] = exp_cost
tissue.costs[‘locations’] = loc_cost
```

CRITICAL STEP It is best to examine all cost matrices (e.g. using `plt.imshow(tissue.costs[‘expression’])`) to ensure that the matrices are non-uniform and do not have non-finite values (e.g. nan, or inf), otherwise troubleshooting is required.

?Troubleshooting

Compute optimal transport of cells to locations and predicted expression over target space
Timing: 10 seconds

9. Compute OT of cells to locations with a given alpha parameter:

```
# compute optimal transport of cells to locations
alpha_linear = 0.8
epsilon = 5e-3
tissue.reconstruct(alpha_linear=alpha_linear, epsilon=epsilon)
```

Setting `alpha_linear`: `alpha_linear` parameter (α) controls the contribution of the reference atlas relative to the structural correspondence objective. `alpha_linear=0` means that no prior information is available and *de novo* reconstruction is performed. `alpha_linear=1` amounts to only using the marker gene information for the reconstruction. Here we run with `alpha_linear=0.8` as the atlas contains many spatially informative genes.

Setting `epsilon`: The `epsilon` parameter is associated with the entropic regularization term. A low epsilon results in a more localized mapping and a higher epsilon with a higher-entropy (approaching a uniform) mapping. Setting a relatively low epsilon (e.g. `epsilon = 5e-3`) is often necessary to achieve differential expression across positions. However, choosing an epsilon that is too low can lead to numerical errors (see Troubleshooting). For convenience, running the computation with argument `search_epsilon=True` automatically attempts reconstruction with a greater epsilon if numerical errors are observed.

Setting marginal distributions: By default, marginal distributions are set as uniform over cells and locations. For cases where we would like to assign different probabilities to different cells (e.g. the marginal probability of a cell is proportional to its sequenced reads, expressing our confidence in individual cells’ expression measurements), or to different locations (e.g. reflecting varying cellular density), it is also possible to manually set the marginal probability distributions of cells, `p_expression`, and of locations, `p_locations`. For example, we can adjust the marginal distribution over locations to reflect a denser cell composition near the center of the tissue (Fig. 5d):

```
# adjust location marginals gradual cell density from the tissue’s center
rdist = novosparc.gm.prob_dist_from_center(locations)

atlas.obs['Alternative location marginals'] = rdist
novosparc.pl.embedding(atlas, ['Alternative location marginals'])

tissue.reconstruct(alpha_linear=alpha_linear, epsilon=epsilon, p_locations=rdist)
```

Once computed, the transport matrix is available in Tissue's object field tissue.gw (numpy ndarray of dimensions num_cells x num_locations), and the predicted expression in tissue.sdge (numpy ndarray shaped as num_genes x num_locations).

?Troubleshooting

Validation of results and follow-up analyses Timing: 45 seconds

10. Validate predicted expression over target space. tissue.sdge captures the predicted gene expression over the target space locations. Visualize the inferred expression by constructing a Scanpy object for the predicted spatial expression (Fig. 6b):

```
# reconstructed expression of individual genes
sdge = tissue.sdge
dataset_reconst = sc.AnnData(pd.DataFrame(sdge.T, columns=gene_names))
dataset_reconst.obsm['spatial'] = locations

novosparc.pl.embedding(dataset_reconst, pl_genes)
```

An inherent feature of *de novo* reconstruction, not restricted to the novoSpaRc algorithm, is that the orientation of the reconstructed virtual tissue is arbitrary up to global transformations (reflections, rotations and translations), relative to the respective axes of symmetry of the target space (see more detailed discussion in[33]).

Any information regarding the spatial expression of a subset of genes, whether it is location-specific or general lower-resolution information of where, or in what patterns a gene is expressed, can help validate the results of the reconstruction. Retrospectively, we can visually assess the validity of the reconstruction by plotting the reconstructed spatial expression of a gene and compare it to the information at hand, and, if applicable, quantify the correlation between the original and predicted expression.

Moreover, to cross-validate our reconstruction, we can measure expression correlation while varying individual, or multiple parameters, such as the selected expression representation (e.g. highly variable genes, or PCs), num_locations, num_neighbors_s, num_neighbors_t, epsilon and alpha_linear. For example, we probe the choice of alpha_linear when less marker genes are available, using 40 random markers (Fig. 8a):

```
# cross-validation with atlas
```

```
repeats = 10
num_markerss = [40]
alpha_linears = np.arange(0.5, 1.0001, 0.1)

df_corr_atlas, df_corr_repeats = novosparc.an.correlation_random_markers(tissue,
                                              with_atlas=True,
                                              with_repeats=False,
                                              alpha_linears=alpha_linears,
                                              epsilons=[epsilon],
                                              num_markerss=num_markerss,
                                              repeats=repeats)

tit='Correlation with atlas'
alt.Chart(df_corr_atlas, title=tit).mark_boxplot().encode(x='alpha_linear:Q',
                                         y='Pearson correlation:Q')
```

?Troubleshooting

11. Validate localized mapping of individual cells.
The mapping of cells to locations is available in the matrix `tissue.gw` (Fig. 9a). The dimensions of the numpy ndarray are `num_cells x num_locations`, and each entry represents the relative probability that a cell is mapped to a specific spatial position. Cells can be spread across multiple locations, and the spatial expression pattern is computed as a weighted average of cellular expression. Validate mapping of cells as follows:

```
# probability of individual cells belonging to each location
gw = tissue.gw
ngw = (gw.T / gw.sum(1)).T
cell_idx = [1, 12]
cell_prb_cols = ['cell %d' % i for i in cell_idx]
dataset_reconst.obs = pd.DataFrame(ngw.T[:, cell_idx], columns=cell_prb_cols)

novosparc.pl.embedding(dataset_reconst, cell_prb_cols)
```

To overview how localized the transport matrix is, one can examine the distribution (over all cells) of the entropy of the transportation of individual cells (`ent_T`) and statistically compare it (e.g. using Kolmogorov–Smirnov test (KS)) with randomized transportations (e.g. with random mapping, `ent_T_rproj`, Fig. 9b). If cells are uniformly mapped across locations, see the Troubleshooting section. Evaluate the entropy distribution as follows:

```
# evaluate entropy of transportation
ent_T, ent_T_unif, ent_T_rproj, ent_T_shuf =
novosparc.pl.plot_transport_entropy_dist(tissue_with_markers)

ks_2samp(ent_T, ent_T_rproj) # KstestResult(statistic=1.0, pvalue=0)
```

?Troubleshooting

12. Self-consistency analysis.
As the optimal transport framework converges to a local optimum, it is useful to examine the robustness of the results. Self-consistency analysis without a reference atlas can be done by performing multiple reconstruction runs with random initialization (setting `random_ini=True` in `tissue.reconstruct`), subsampling cells used for reconstruction, or adding external noise or induced sparsity to the expression matrix. If we are using a reference atlas for reconstruction as here, it is also possible to vary the reference atlas genes used for reconstruction. Set `with_repeats=True` in `novosparc.an.correlation_random_markers` (Step 10) to quantify the stability of `alpha_linear` values (Fig. 8b).

13. Verify and identify spatially informative genes.
Compare Moran's I values using `tissue.calculate_spatially_informative_genes` for genes that are expected to be spatially informative (for example, marker genes denoted by `atlas_genes`, are often chosen to be assayed because they are spatially informative), and non-informative (e.g. cell-cycle-related genes, `cyc_genes`) as shown in Fig. 10. Once computed, `tissue.spatially_informative_genes` holds a pandas DataFrame with the following columns: `genes`, `mI`, `pval`, corresponding to each of the genes' name, Moran's I value and its corresponding one-tailed p-value computed under normality assumption using permuted locations (the size of neighborhood used can be set using `n_neighbors`), respectively:

```
# verify spatially informative genes
cyc_genes = [g for g in gene_names if g.startswith('Cyc')]
mI_genes = cyc_genes + atlas_genes

tissue.calculate_spatially_informative_genes(mI_genes)
genes_with_scores = tissue.spatially_informative_genes

genes_with_scores.index = genes_with_scores['genes']

gene_groups = {'Atlas': atlas_genes, 'Cell-cycle': cyc_genes}
novosparc.pl.plot_morans_dists(genes_with_scores, gene_groups)

gene_max_mI = genes_with_scores['genes'].iloc[0]
gene_min_mI = genes_with_scores['genes'].iloc[-1]

novosparc.pl.embedding(dataset_reconst, [gene_max_mI, gene_min_mI)
```

Furthermore, using this code, we can also identify novel spatially informative genes.

14. Extract archetypes.
    Beyond examining the recovered spatial expression of individual genes, we can further cluster the spatial expression profiles to `num_clusters` archetypes to capture prototypical expression patterns and identify spatial gene expression programs of spatially informative genes (e.g. extracting archetypes for the inferred spatial expression of atlas marker genes, Fig. 11). The spatial expression `archetypes` (numpy ndarray of `num_clusters x num_locations`), the cluster assignment for each gene, `clusters` (numpy ndarray of length `num_genes`), and the correlation of the gene and its chosen archetype, `gene_corrs` (also numpy ndarray of length `num_genes`), are computed using `novosparc.rc.find_spatial_archetypes`:

```
# extracting archetypes
num_clusters = 10
atlas_indices = pd.DataFrame(np.arange(num_genes),
index=gene_names)[0].loc[atlas_genes].values

archetypes, clusters, gene_corrs =
novosparc.rc.find_spatial_archetypes(num_clusters, sdge[atlas_indices,:])

arch_cols = ['archetype %d'% i for i in np.arange(num_clusters)]
dataset_reconst.obs = pd.DataFrame(index=dataset_reconst.obs.index)
df = pd.DataFrame(archetypes.T, columns=arch_cols)
dataset_reconst.obs = pd.concat((dataset_reconst.obs, df), 1)

novosparc.pl.embedding(dataset_reconst, arch_cols)
```

Troubleshooting

Troubleshooting advice can be found in Table 2.

Timing

Total runtime for the example given in the Procedure is under 1.5 minutes:

Equipment setup, novoSpaRc installation via *pip*: 10 seconds
Steps 1-6, Input cells and locations descriptions to construct Tissue object: 1 second
Steps 7-8, Compute cost matrices: 5 seconds
Step 9, Compute optimal transport of cells to locations and predicted expression over target space: 10 seconds
Steps 10-14, Validation of results and follow-up analyses: 45 seconds

Since runtimes of the setup and reconstruction steps depend on the numbers of cells, locations and genes that are used for reconstruction, we benchmarked the scalability of novoSpaRc on an extensive (~40K cells) dataset of whole-kidney scRNA-seq[52] (Fig 4). All runtime calculations are done on a workstation with Intel(R) Core(TM) i7-9800X CPU @ 3.80GHz and 64 GB RAM.

Anticipated results

The results in the Procedure Section highlight the prerequisites, validity checks, integration and benefits of the structural correspondence objective and the reference atlas objective in the novoSpaRc framework. Specifically, beyond describing conventional steps that are relevant for any reconstruction (e.g. preparation steps 1-4), we leverage the extensive reference atlas to evaluate the structural correspondence assumption and find that gene expression of nearby cells (e.g. within a radius of ~40 cells) is expected to be more similar than expression of cells that are farther away (Fig. 7a). We additionally evaluated the spatial auto-correlation scores of the genes composing the reference atlas (Fig. 7b).

Based on our analysis, we chose `alpha_linear`, the parameter used to interpolate between the two objectives, to lean towards the reference atlas objective. The atlas here holds information for a large number of spatially informative genes (84 transcription factors). In such a case, a maximal `alpha_linear` value is often favored. However, in many scenarios the reference atlas at hand is less informative, resulting in a tradeoff between using a larger `alpha_linear` value, leading to reconstructed spatial expression which is more in agreement with the expression of the reference atlas (Fig. 8a), and using a smaller `alpha_linear` value, biasing the reconstructed spatial expression towards smoother solutions (consistent with the structural correspondence assumption) (Fig. 8b).

We further evaluate the reconstruction by ensuring that the embedding of single cells is relatively localized within the tissue (Fig. 9), and that the spatial expression of marker genes given by the reference atlas is correlated to their inferred spatial expression by novoSpaRc (Fig. 6). Finally, hierarchical clustering of spatial expression patterns into spatial archetypes can generate more robust, interpretable spatial signatures (Fig. 11).

The basic outputs of spatial reconstruction using novoSpaRc are summarized in Table 3.

Data availability

All data analyzed within this protocol are publicly available. The osteosarcoma dataset[14] (Fig. 3) and the organ of Corti data[47] (Fig. 2) can be downloaded from the accompanying supplementary files of their corresponding manuscripts. The *Drosophila* embryo scRNA-seq data [30] used in the Procedure Section was acquired from the GEO database with accession number GSE95025, and the reference BDTNP dataset can be downloaded directly from the BDTNP webpage [51,54]. The whole-kidney dataset [52] used for benchmarking runtimes (Fig. 4) is available in the GEO database with accession number GSE107585.

Code availability

NovoSpaRc is available as a Python package at https://pypi.org/project/novosparc/, and its source code is available on GitHub (https://github.com/rajewsky-lab/novosparc) and on Zenodo[55].

Author Contributions

This protocol is based on a manuscript by M.N., N.K., N.F. and N.R. Here, N.M., E.S., N.K. and M.N. implemented the method and performed computational and data analyses. N.M., E.S., N.F., N.R., N.K. and M.N. wrote the manuscript.

Acknowledgements

N.M. is supported by CIDR (Center for Interdisciplinary Data Science Research). N.K. and E.S were supported by the DFG grant KA 5006/1-1. M.N. is supported by an Early Career Faculty Fellowship by the Azrieli Foundation.

Ethics declarations

Competing interests

The authors declare no competing interests.

Related links

Key references using this protocol

Nitzan, M. et al. Nature 576, 132–137 (2019): https://doi.org/10.1038/s41586-019-1773-3
Hinze, C. et al. J. Am. Soc. Nephrol. 32, 291–306 (2021): https://doi.org/10.1681/ASN.2020070930
Holmes, A. et al. J. Exp. Med. 217, e20200483 (2020): https://doi.org/10.1084/jem.20200483

Key data used in this protocol

Waldhaus, J. et al. Cell Rep. 11, 1385–1399 (2015): https://doi.org/10.1016/j.celrep.2015.04.062]
Xia, C. et al. Proc. Natl. Acad. Sci. U. S. A. 116, 19490–19499 (2019): https://doi.org/10.1073/pnas.1912459116
Park, J., et al. Science 360(6390), 758-763 (2018): https://doi.org/10.1126/science.aar2131
Karaiskos, N. et al. Science 358, 194–199 (2017): https://doi.org/10.1126/science.aan3235
Larkin, A. et al. Nucleic Acids Res. 49(D1) D899–D907(2021): https://doi.org/10.1093/nar/gkaa1026

References

1. Aldridge, S. & Teichmann, S. A. Single cell transcriptomics comes of age. *Nat. Commun.* **11**, 4307 (2020).
2. Kulkarni, A., Anderson, A. G., Merullo, D. P. & Konopka, G. Beyond bulk: a review of single cell transcriptomics methodologies and applications. *Curr. Opin. Biotechnol.* **58**, 129–136 (2019).
3. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
4. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
5. Birnbaum, K. D. Power in Numbers: Single-Cell RNA-Seq Strategies to Dissect Complex Tissues. *Annu. Rev. Genet.* **52**, 203–221 (2018).
6. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155–160 (2015).
7. Plasschaert, L. W. *et al.* A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).
8. Saunders, A. *et al.* Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell* **174**, 1015–1030.e16 (2018).

9. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).

10. Wagner, D. E. & Klein, A. M. Lineage tracing meets single-cell omics: opportunities and challenges. *Nat. Rev. Genet.* **21**, 410–427 (2020).

11. Cannoodt, R., Saelens, W. & Saeys, Y. Computational methods for trajectory inference from single-cell transcriptomics. *Eur. J. Immunol.* **46**, 2496–2506 (2016).

12. Moor, A. E. *et al.* Spatial Reconstruction of Single Enterocytes Uncovers Broad Zonation along the Intestinal Villus Axis. *Cell* **175**, 1156–1167.e15 (2018).

13. Puram, S. V. *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611–1624.e24 (2017).

14. Xia, C., Fan, J., Emanuel, G., Hao, J. & Zhuang, X. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 19490–19499 (2019).

15. Sun, S., Zhu, J. & Zhou, X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat. Methods* **17**, 193–200 (2020).

16. Eng, C.-H. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235–239 (2019).

17. Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, (2018).

18. Gumbleton, M. *et al.* Spatial expression and functionality of drug transporters in the intact lung: objectives for further research. *Adv. Drug Deliv. Rev.* **63**, 110–118 (2011).

19. Arnol, D., Schapiro, D., Bodenmiller, B., Saez-Rodriguez, J. & Stegle, O. Modeling Cell-Cell Interactions from Spatial Molecular Data with Spatial Variance Component Analysis. *Cell Rep.* **29**, 202–211.e6 (2019).

20. Teves, J. M. & Won, K. J. Mapping Cellular Coordinates through Advances in Spatial Transcriptomics Technology. *Mol. Cells* **43**, 591–599 (2020).

21. Goltsev, Y. *et al.* Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* **174**, 968–981.e15 (2018).

22. Qian, X. *et al.* Probabilistic cell typing enables fine mapping of closely related cell types in situ. *Nat. Methods* **17**, 101–106 (2020).

23. Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).

24. 10x Genomics. https://www.10xgenomics.com/products/spatial-gene-expression/.

25. Home - Spatial Transcriptomics. https://spatialtranscriptomics.com/.

26. GeoMx Digital Spatial Profiling. https://www.nanostring.com/products/geomx-digital-spatial-profiler/geomx-dsp?utm_source=AdWords&utm_medium=GeoMx_SearchAd&gclid=EAIaIQobChMIrJiah43N6wIVTuR3Ch2qSAzUEAAYASAAEgKmcPD_BwE.

27. Stickels, R. R. *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0739-1.

28. 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).

29. Achim, K. *et al.* High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* **33**, 503–509 (2015).

30. Karaiskos, N. *et al.* The Drosophila embryo at single-cell transcriptome resolution. *Science* **358**, 194–199 (2017).

31. Okochi, Y., Sakaguchi, S., Nakae, K., Kondo, T. & Naoki, H. Model-based prediction of spatial gene expression via generative linear mapping. 2020.05.21.107847 (2020) doi:10.1101/2020.05.21.107847.

32. Biancalani, T. *et al.* Deep learning and alignment of spatially-resolved whole transcriptomes of single cells in the mouse brain with Tangram. 2020.08.29.272831 (2020) doi:10.1101/2020.08.29.272831.

33. Nitzan, M., Karaiskos, N., Friedman, N. & Rajewsky, N. Gene expression cartography. *Nature* **576**, 132–137 (2019).

34. Peyré, G. & Cuturi, M. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning* **11**, 355–607 (2019).

35. Villani, C. *Topics in Optimal Transportation*. (American Mathematical Soc., 2003).

36. Durruthy-Durruthy, R. *et al.* Reconstruction of the mouse otocyst and early neuroblast lineage at single-cell resolution. *Cell* **157**, 964–978 (2014).

37. Stickels, R. R. *et al.* Sensitive spatial genome wide expression profiling at cellular resolution. 2020.03.12.989806 (2020) doi:10.1101/2020.03.12.989806.

38. 10x Genomics. https://www.10xgenomics.com/products/spatial-gene-expression/.

39. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).

40. Espina, V. *et al.* Laser-capture microdissection. *Nat. Protoc.* **1**, 586–603 (2006).

41. Combs, P. A. & Eisen, M. B. Sequencing mRNA from cryo-sliced Drosophila embryos to determine genome-wide spatial patterns of gene expression. *PLoS One* **8**, e71820 (2013).

42. Junker, J. P. *et al.* Genome-wide RNA Tomography in the zebrafish embryo. *Cell* **159**, 662–675 (2014).

43. Mémoli, F. On the use of Gromov-Hausdorff distances for shape comparison. (2007).

44. Peyre, G., Cuturi, M. & Solomon, J. Gromov-wasserstein averaging of kernel and distance matrices. (2016).

45. 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).

46. Halpern, K. B. *et al.* Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature* **542**, 352–356 (2017).

47. Waldhaus, J., Durruthy-Durruthy, R. & Heller, S. Quantitative High-Resolution Cellular Map of the Organ of Corti. *Cell Rep.* **11**, 1385–1399 (2015).

48. Habib, N. *et al.* Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science* **353**, 925–928 (2016).

49. Ren, X. *et al.* Reconstruction of cell spatial organization from single-cell RNA sequencing data based on ligand-receptor mediated self-assembly. *Cell Res.* **30**, 763–778 (2020).

50. Cang, Z. & Nie, Q. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nat. Commun.* **11**, 2084 (2020).

51. Larkin, A. *et al.* FlyBase: updates to the Drosophila melanogaster knowledge base. *Nucleic Acids Res.* **49**, D899–D907 (2021).

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

53. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

54. FlyBase. [No title]. http://flybase.org/reports/FBlc0003350.html.

55. Moriel, N. *et al. NovoSpaRc: flexible spatial reconstruction of single-cell gene expression with optimal transport*. (2021). doi:10.5281/zenodo.4661199.

Figure 1: Schematic representation of the novoSpaRc algorithm. a, Preparation of inputs for novoSpaRc's Tissue object – constructing a target space and reading gene expression datasets. If a reference atlas is used, then the target space corresponds to its locations. **b,** Computation of cost matrices including physical distances between locations and expression distances between cells, both computed as the shortest path in k-Nearest Neighbors (kNN) graphs. If a reference atlas is used, then an additional cost matrix of atlas correspondence captures the expression discrepancy between locations and cells according to the reference atlas. **c,** Computing the optimal transportation of cells to locations (`tissue.gw`) given a parameter α interpolating between the structural correspondence and atlas correspondence objectives. The predicted expression of genes over locations (`tissue.sdge`) is then computed by matrix multiplication of cellular gene expression and their probabilistic mapping to locations (`tissue.gw`). The output probabilistic embedding and the predicted spatial gene expression can be fetched from the Tissue object.

Figure 2: NovoSpaRc successfully reconstructs *de novo* the organ of Corti. a, Illustration of the organ of Corti where regions of distinct cell types, including outer hair cell (OHC), spirally span the base-to-apex axis of the cochlea. **b,** *De novo* mapping OHCs (based on single-cell data in [47]) to a 1D grid (x-axis), correctly places cells of differing original base/apex membership (blue/orange respectively), with high probability (y-axis) at opposite sides of the grid. **c,** Genes known to be zonated towards either the base or the apex in OHC cells [47] are successfully reconstructed as such, without a reference atlas. The expression level of each gene is normalized to its maximum value. **d,** Base/apex membership values are recapitulated at opposite ends with *de novo* mapping for each of the organ of Corti cell types using their differentially expressed genes. Cell-type abbreviations: Greater Epithelial Ridge (GER), Inner Border Cell (IBC), Inner Hair Cell (IHC), Inner Phalangeal Cell (IPH), Inner Pillar Cell (IPC), Outer Pillar Cell (OPC), Deiters' Cell row 1-2 (DC12), and Deiters' Cell row 3 (DC3). For (b,d) membership values are normalized to the maximum layer mean. Error bars represent the standard error.

Figure 3: Spatial reconstruction of osteosarcoma cultured cells using marker gene information. a, Cellular microenvironments [14] in cultured osteosarcoma cells visualized using FGF18 gene expression level as captured with MERFISH technology[14]. **b,** The quality of reconstruction, computed by the median Pearson correlation between the original gene expression and spatial expression reconstructed by novoSpaRc, improves with the number of marker genes and reaches saturation using 4 marker genes. The results are averaged over 20 different combinations of marker genes. Centre line: median; box limits: q1 and q3 quantiles; whiskers: extend to [q1 - 1.5 * IQR, q3 + 1.5 * IQR] with IQR (interquartile range) = q3-q1; dots: outside the defined range. **c,** Spatial reconstruction of FGF18 gene expression using (1,2,4) marker genes. 2 random markers are sufficient to recover the FGF18 microenvironment signatures.

Figure 4: Computing times of the setup and reconstruction steps. NovoSpaRc's runtimes on a whole-kidney dataset[52] for *de novo* reconstruction with varying numbers of cells and locations. 840 highly variable genes are used for the reconstruction. **a,** Setup consists of computing expression and location cost matrices (Steps 7-8). **b,** Reconstruction computes the optimal transport of cells to locations (Step 9).

Figure 5: Configure target space. When constructing a target space (Step 4) we can set the locations to be, **a,** read from a file when available, **b,** sampled from a binary image given a certain shape, or, **c,** set on a regular grid within one of a few preset shapes such as an oval (locations can then be visualized with `plt.scatter(locations[:,0], locations[:,1], s=1)`). **d,** When running `tissue.reconstruction` (Step 9), we can incorporate more information regarding locations in the target space, such as non-uniform marginal probability distribution over locations (here, corresponding to denser cell concentration near the center of the tissue).

Figure 6: Spatial expression based on reference atlas and novoSpaRc reconstruction. a, Visualizing spatial gene expression patterns of three genes (*sna, ken, eve*) based on the *Drosophila* embryo BDTNP reference atlas[51], and **b,** their reconstruction by novoSpaRc from scRNA-seq data (available in `tissue.sdge`). Corresponding Pearson correlation values are indicated.

Figure 7: Structural correspondence and Moran's I score for the *Drosophila* reference atlas. a, Assessing the structural correspondence assumption in the *Drosophila* reference atlas[51] by plotting the pairwise Euclidean distances between cells in terms of their gene expression vs. their physical locations. Centre line: median; box limits: q1 and q3 quantiles; whiskers: extend to [q1 - 1.5 * IQR, q3 + 1.5 * IQR] with IQR (interquartile range) = q3-q1; dots: outside the defined range. **b,** Visualizing the spatial expression of reference atlas genes *sna, zen2* with maximal (left) and minimal (right) spatial auto-correlation (Moran's I) score, respectively. Mean Moran's I score for reference atlas genes = 0.88.

Figure 8: Cross-validation and self-consistency for selecting alpha-linear. Assessing `alpha_linear` parameter by measuring **a,** the correlation between novoSpaRc reconstruction of the *Drosophila* embryo and the corresponding reference atlas, and **b,** the average correlation within a group of novoSpaRc's reconstructions. For each reconstruction, 40 marker genes are randomly chosen out of the full 84 marker gene atlas. Centre line: median; box limits: q1 and q3 quantiles; whiskers: extend to [q1 - 1.5 * IQR, q3 + 1.5 * IQR] with IQR (interquartile range) = q3-q1.

Figure 9: Inspecting cell-to-location optimal transport values. a, Visualization of the mapping probability distribution over locations (normalized rows `tissue.gw` matrix) for two representative cells, as inferred by novoSpaRc. **b,** To statistically assess how concentrated

the inferred spatial embedding is, we compare the distributions of the entropy values of (i) the inferred transport matrix by novoSpaRc (blue), (ii) random transport matrix (yellow), (iii) outer product of the marginal cell and location probabilities (this is a uniform distribution when the marginals are uniform, green), and (iv) inferred transport matrix by novoSpaRc when the reference atlas is shuffled (each gene is shuffled independently over locations, red).

Figure 10: Analysis of spatially informative genes. Post-reconstruction, we can analyze the spatial auto-correlation score for different gene groups. **a,** Genes composing the BDTNP reference atlas (such as **b,** left, *ImpE2*) are found to be highly spatially informative (average Moran's I values of gene group = 0.85), whereas genes related to the cell-cycle (genes of 'cyc' prefix) have a lower average Moran's I value (0.59), as their expression is disorganized over the tissue (such as **b,** right, *CycK*).

Figure 11: Extracting spatial archetypes. Clustering the spatial expression of spatially informative genes to discover spatial archetypes can generate an effective analysis and summary of global spatial expression patterns, which can help in identifying tissue functions, cellular division of labor, and underlying mechanisms of regulation of processes. We show here three distinct archetypes discovered by clustering the inferred spatial expression of the *Drosophila* embryo (genes used for clustering were restricted to those included in the BDTNP reference atlas).

Table 1: Comparison of spatial reconstruction methods.

	Seurat[45]	**DistMap**[30]	**novoSpaRc**[33]	**Perler**[31]	**Tangram**[32]	**CSOmap**[49]
Spatial mapping with reference atlas	✓	✓	✓	✓	✓	×
Probabilistic mapping of cells to locations	✓	✓	✓	✓	✓	✓
Spatial mapping *de novo*	×	×	✓	×	×	✓
Does not require predetermined shape	✓	✓	×	✓	✓	✓
Incorporates the structural correspondence assumption	×	×	✓	×	×	×
Reference atlas can have continuous values	×	×	✓	✓	✓	-
Does not require data imputation	×	✓	✓	✓	✓	-
Does not require a threshold for binarizing expression	✓	×	✓	✓	✓	-
Requires ligand-receptor communication reference	×	×	×	×	×	✓
Accounts for cell density prior	×	×	✓	×	✓	✓
Programming language	R	R	Python	Python	Python	Matlab

Table 2: Troubleshooting.

Step	Problem	Possible Reason	Solution
8	Uniform cost matrix or non-finite values (nan, inf, -inf) in cost matrix	Self-set cost matrices or non-finite value in input descriptions	If using novoSpaRc cost matrix computation, check inputs - cell, location and atlas descriptions for uniform description and mal values.
9	Numerical errors	Epsilon is too low	Examine `tissue.gw` to see if it is uniform and run with `verbose=True` to see if this happens in the early iterations. If either of these are true, try running `tissue.reconstruct` with a higher epsilon
10	No differential spatial expression pattern is detected when visualizing different genes	NovoSpaRc resulted in a uniform mapping	Check `tissue.gw` and if it is uniform (or even almost uniform), examine Troubleshooting for Step 11
10	Low correlation of true and predicted expression for many genes after *de novo* reconstruction	Reconstruction has to be rotated or reflected	When using *de novo* reconstruction, orientation is arbitrary up to global transformations (reflections, rotations and translations), relative to the respective axes of symmetry of the target space. Visualize side-by-side the original and predicted expression

Figure 1

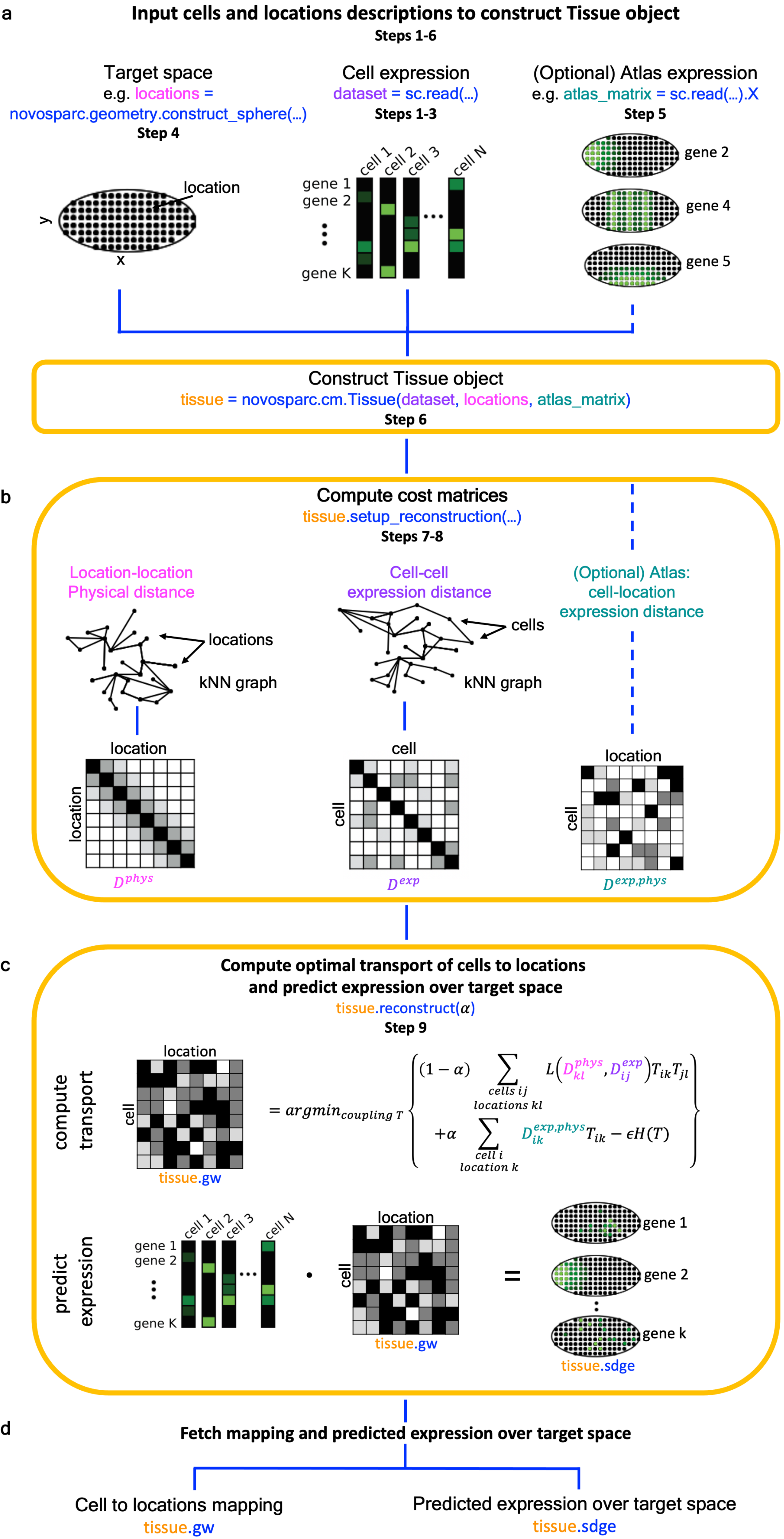

a
Input cells and locations descriptions to construct Tissue object
Steps 1-6
Target space
e.g. locations =
novosparc.geometry.construct_sphere(...)
Step 4
location
y
x
Cell expression
dataset = sc.read(...)
Steps 1-3
cell 1
cell 2
cell 3
cell N
gene 1
gene 2
gene K
(Optional) Atlas expression
e.g. atlas_matrix = sc.read(...).X
Step 5
gene 2
gene 4
gene 5
Construct Tissue object
tissue = novosparc.cm.Tissue(dataset, locations, atlas_matrix)
Step 6
b
Compute cost matrices
tissue.setup_reconstruction(...)
Steps 7-8
Location-location
Physical distance
locations
kNN graph
location
location
D^{phys}
Cell-cell
expression distance
cells
kNN graph
cell
cell
D^{exp}
(Optional) Atlas:
cell-location
expression distance
location
cell
$D^{exp,phys}$
c
Compute optimal transport of cells to locations
and predict expression over target space
tissue.reconstruct(α)
Step 9
compute
transport
location
cell
tissue.gw
$= argmin_{coupling\ T} \left\{ (1-\alpha) \sum_{\substack{cells\ ij \\ locations\ kl}} L\left(D^{phys}_{kl}, D^{exp}_{ij}\right) T_{ik} T_{jl} + \alpha \sum_{\substack{cell\ i \\ location\ k}} D^{exp,phys}_{ik} T_{ik} - \epsilon H(T) \right\}$
predict
expression
cell 1
cell 2
cell 3
cell N
gene 1
gene 2
gene K
location
cell
tissue.gw
=
gene 1
gene 2
gene k
tissue.sdge
d
Fetch mapping and predicted expression over target space
Cell to locations mapping
tissue.gw
Predicted expression over target space
tissue.sdge

Figure 2

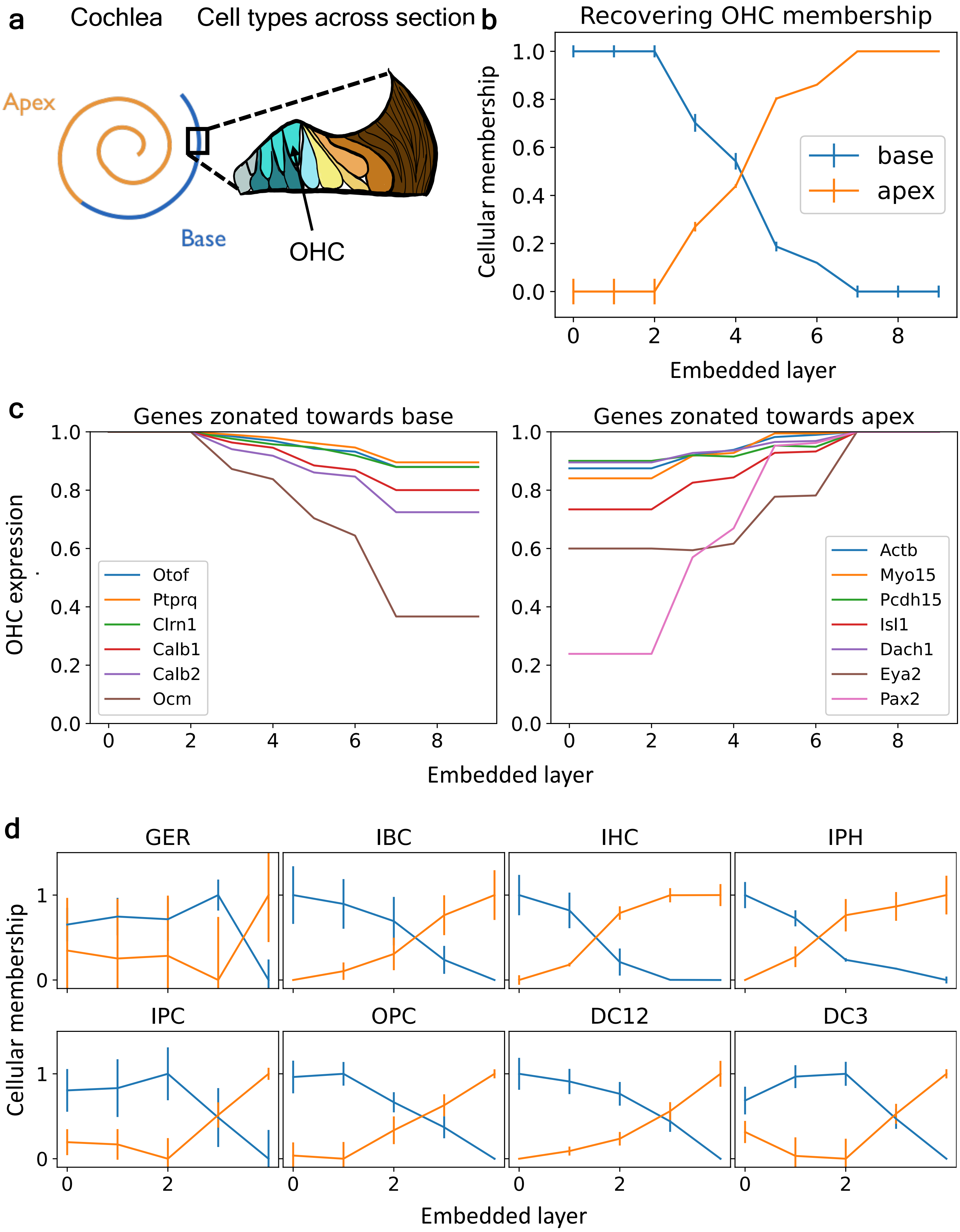
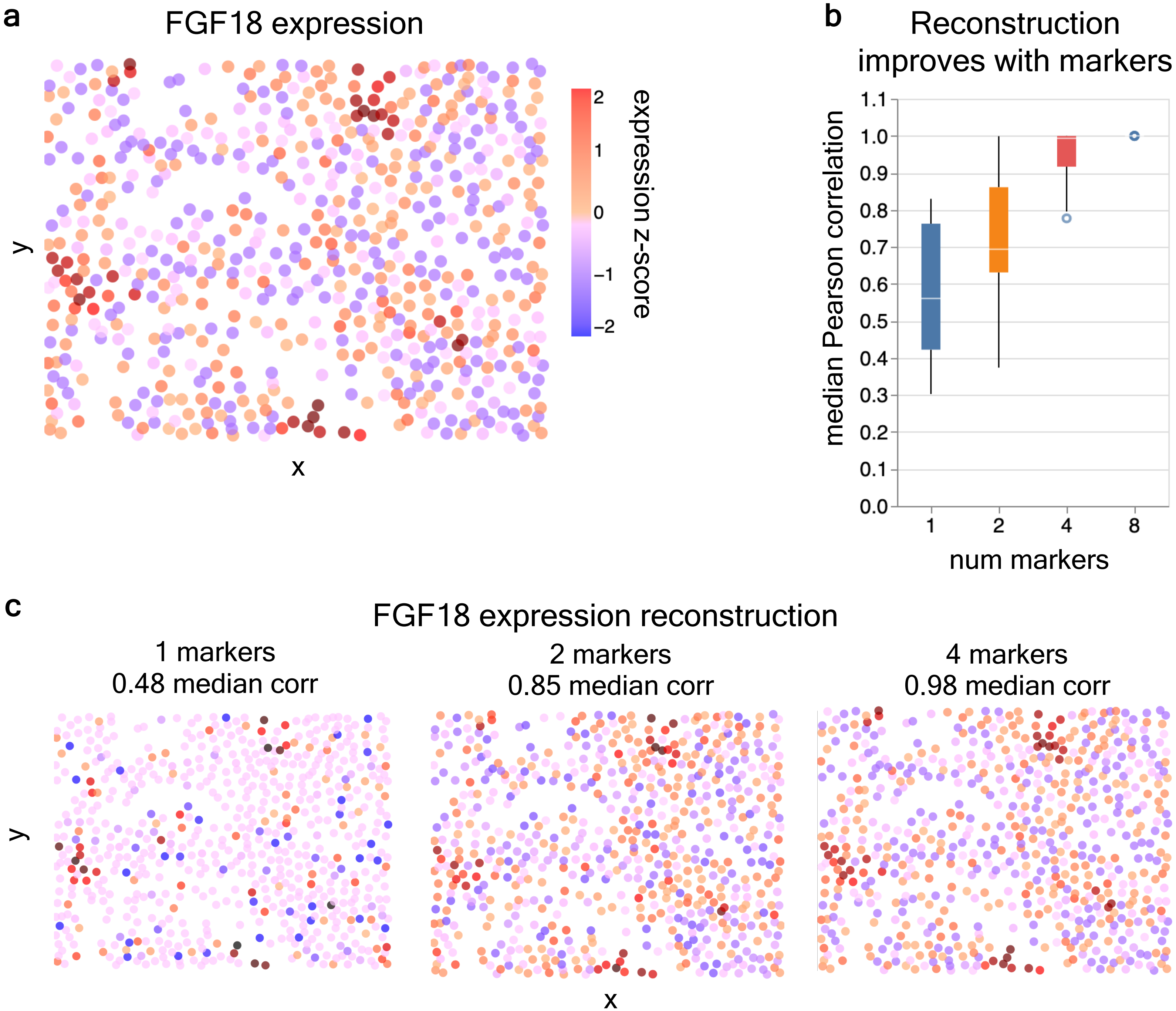

a
Cochlea
Cell types across section
Apex
Base
OHC
b
Recovering OHC membership
Cellular membership
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
Embedded layer
base
apex
c
Genes zonated towards base
Genes zonated towards apex
OHC expression
Otof
Ptprq
Clrn1
Calb1
Calb2
Ocm
Actb
Myo15
Pcdh15
Isl1
Dach1
Eya2
Pax2
Embedded layer
d
GER
IBC
IHC
IPH
IPC
OPC
DC12
DC3
Cellular membership
1
0
0
2
Embedded layer

Figure 3



FGF18 expression

Reconstruction improves with markers

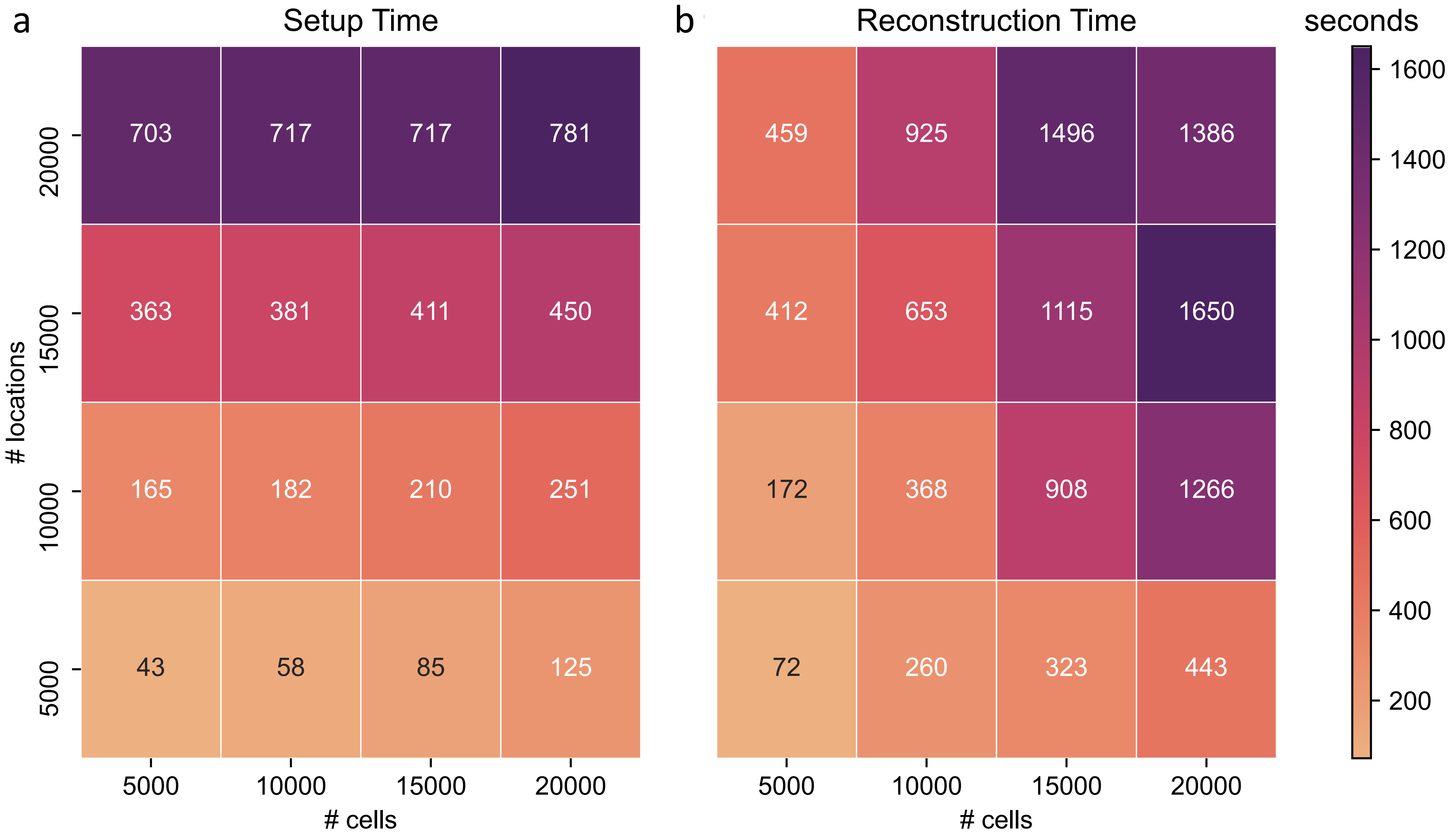
FGF18 expression reconstruction

1 markers
0.48 median corr

2 markers
0.85 median corr

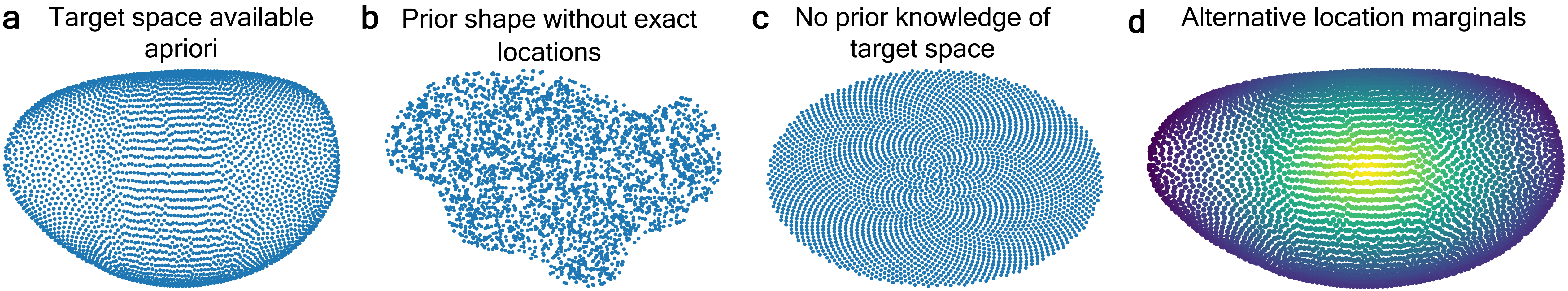
4 markers
0.98 median corr

Figure 4



a
Setup Time
b
Reconstruction Time
seconds
# locations
20000
15000
10000
5000
703
717
717
781
363
381
411
450
165
182
210
251
43
58
85
125
459
925
1496
1386
412
653
1115
1650
172
368
908
1266
72
260
323
443
5000
10000
15000
20000
# cells
5000
10000
15000
20000
# cells
1600
1400
1200
1000
800
600
400
200

Figure 5



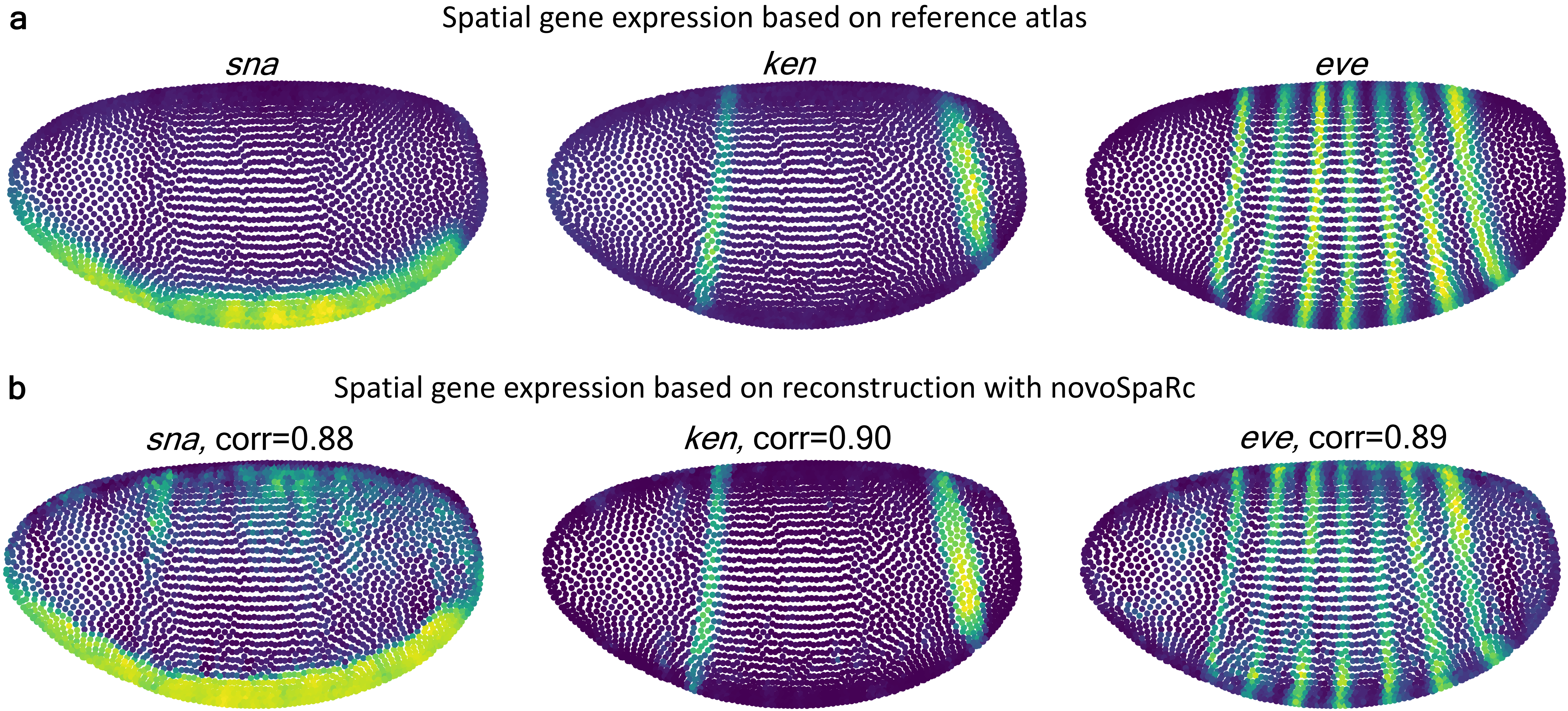
a Target space available apriori

b Prior shape without exact locations

c No prior knowledge of target space

d Alternative location marginals

Figure 6



Spatial gene expression based on reference atlas

sna

ken

eve

Spatial gene expression based on reconstruction with novoSpaRc

sna, corr=0.88

ken, corr=0.90

eve, corr=0.89

Figure 7

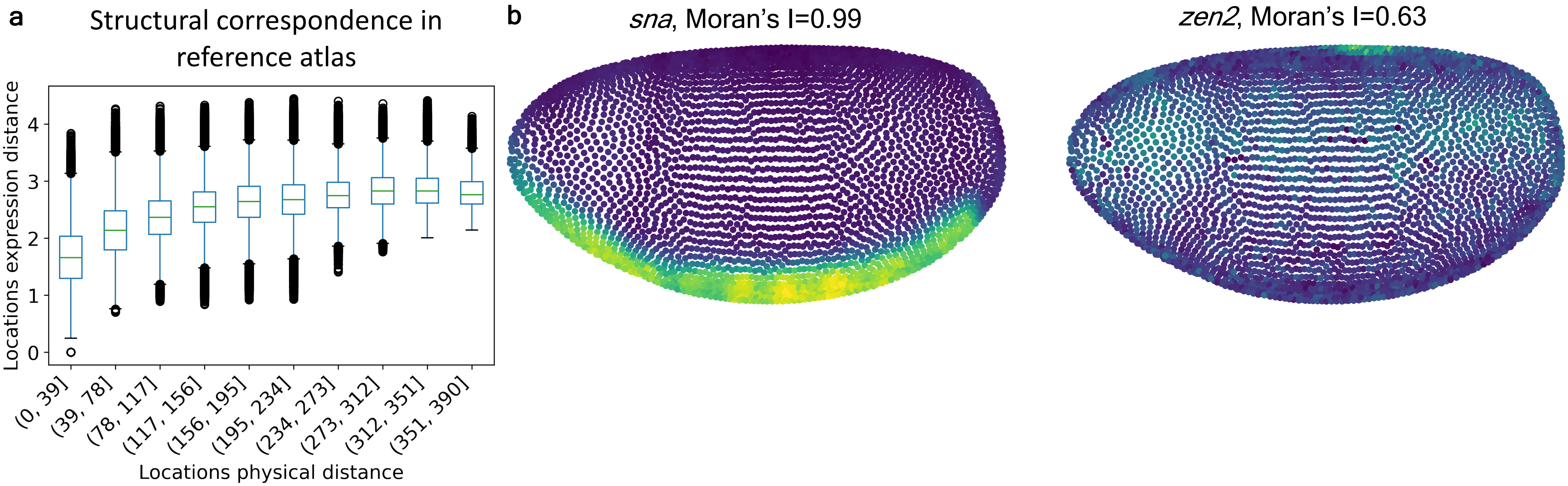
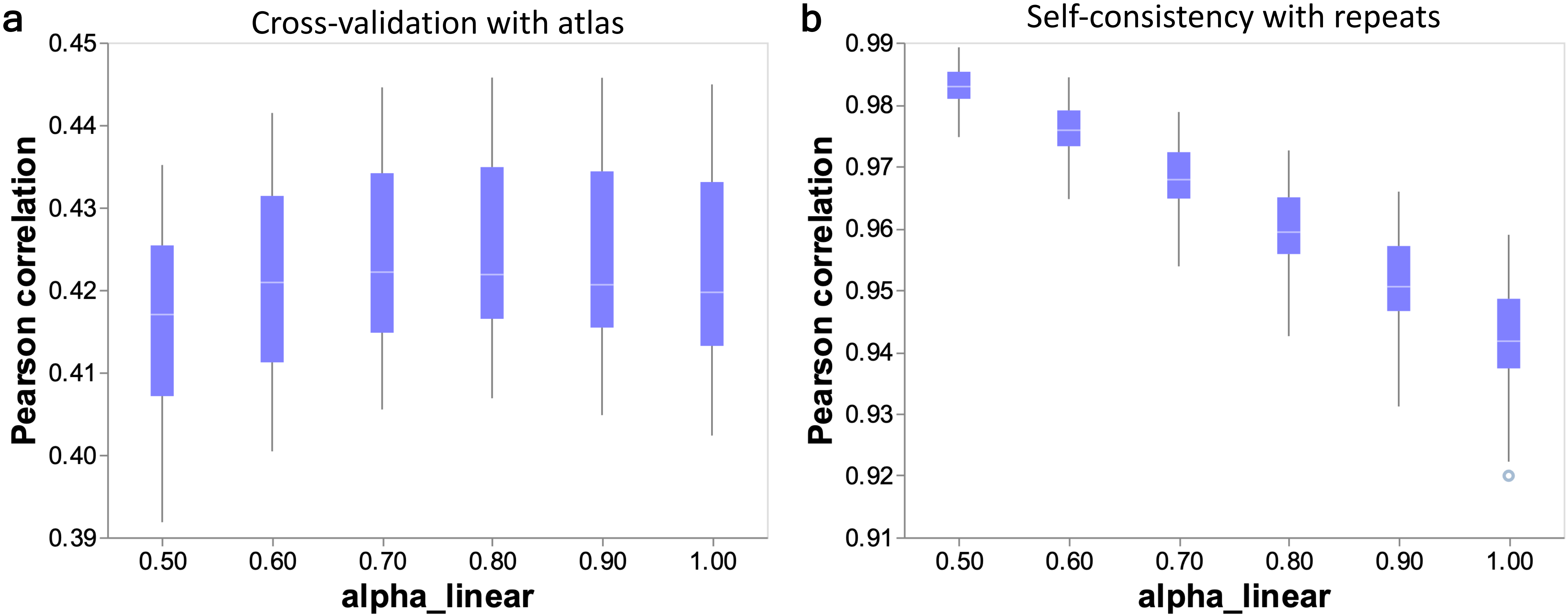
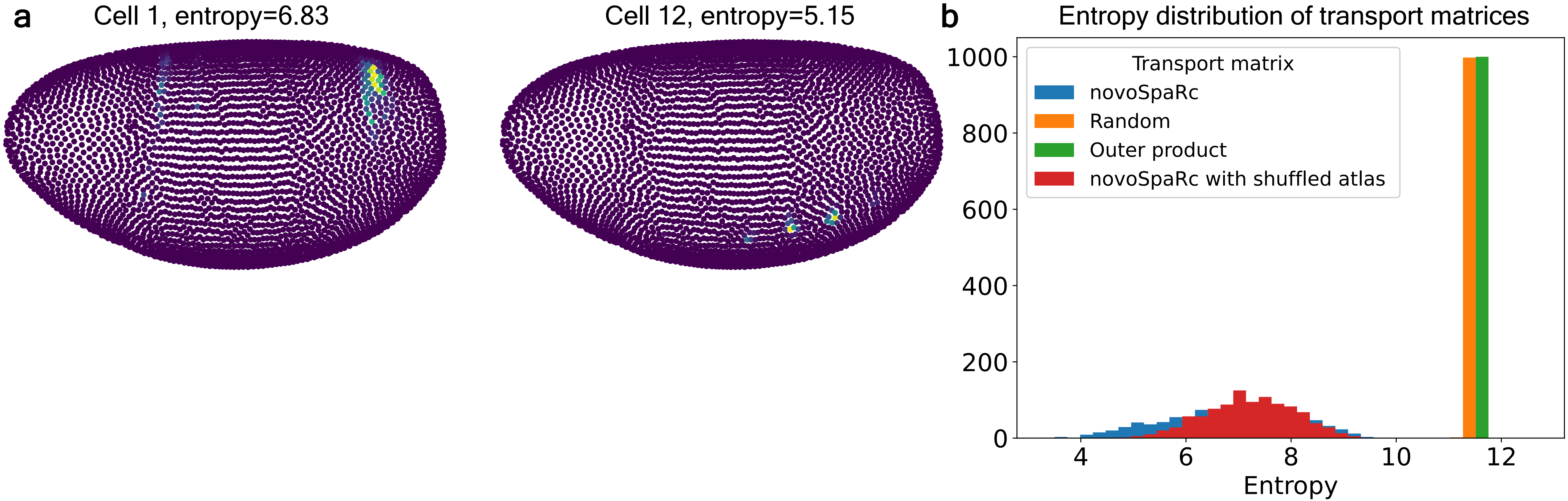

a
Structural correspondence in reference atlas
Locations expression distance
0
1
2
3
4
(0, 39]
(39, 78]
(78, 117]
(117, 156]
(156, 195]
(195, 234]
(234, 273]
(273, 312]
(312, 351]
(351, 390]
Locations physical distance
b
sna, Moran's I=0.99
zen2, Moran's I=0.63

Figure 8



a
Cross-validation with atlas
Pearson correlation
0.45
0.44
0.43
0.42
0.41
0.40
0.39
0.50
0.60
0.70
0.80
0.90
1.00
alpha_linear
b
Self-consistency with repeats
Pearson correlation
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0.50
0.60
0.70
0.80
0.90
1.00
alpha_linear

Figure 9

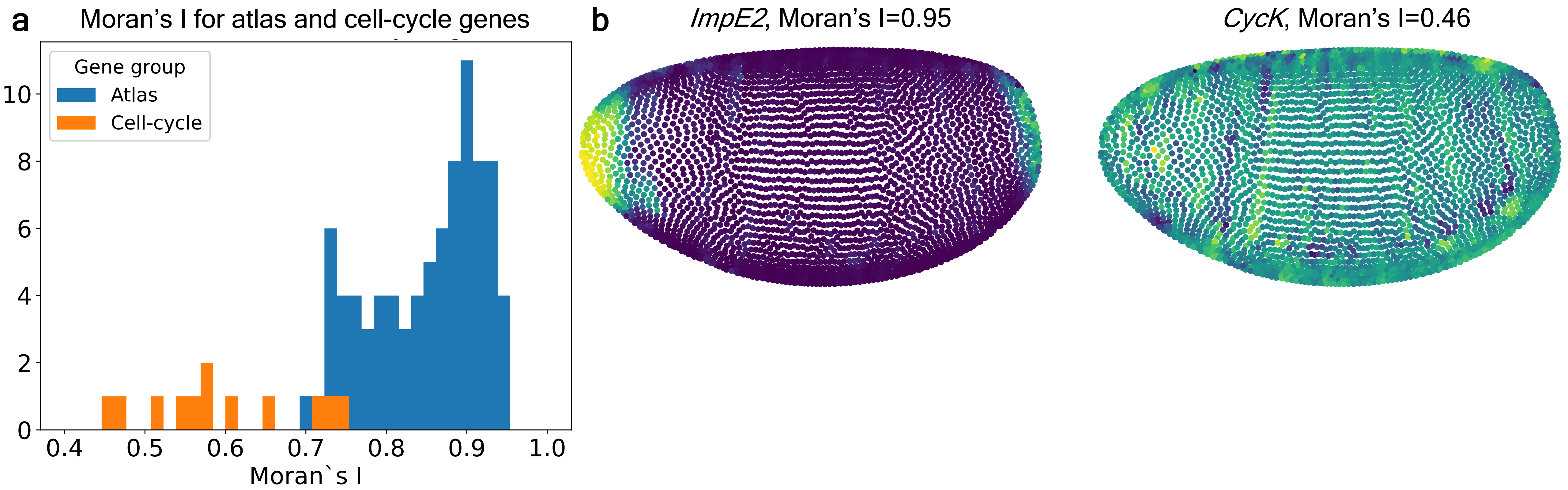


Cell 1, entropy=6.83

Cell 12, entropy=5.15

Entropy distribution of transport matrices

Figure 10

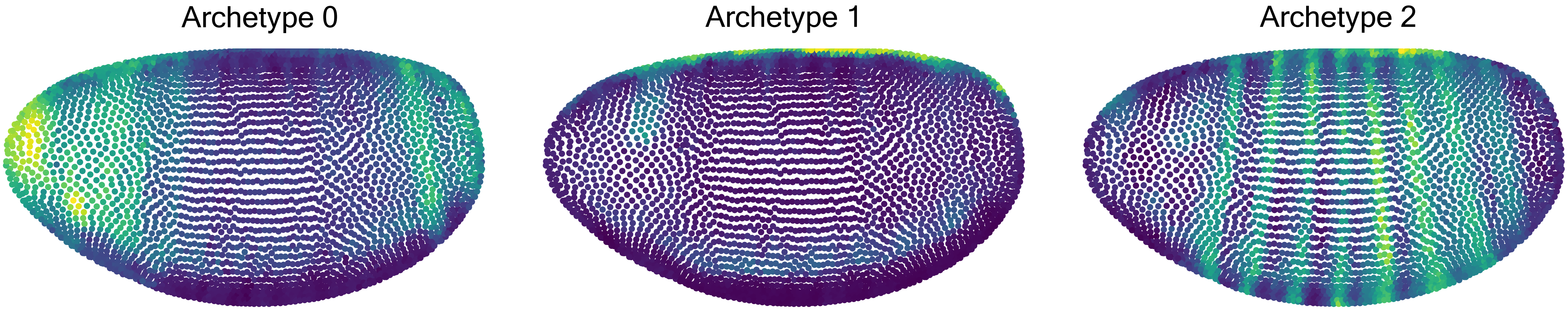


a Moran’s I for atlas and cell-cycle genes

b *ImpE2*, Moran’s I=0.95

CycK, Moran’s I=0.46

Figure 11



Archetype 0

Archetype 1

Archetype 2