Using co-localization priors and microenvironment statistics to reconstruct tissue organization from single-cell data

Yitzchak Vaknin
School of Computer Science and Engineering,
The Hebrew University of Jerusalem, Israel
yitzchak.vaknin@mail.huji.ac.il

Noa Moriel
School of Computer Science and Engineering,
The Hebrew University of Jerusalem, Israel
noa.moriel@mail.huji.ac.il

Mor Nitzan
School of Computer Science and Engineering,
Racah Institute of Physics, and Faculty of Medicine,
The Hebrew University of Jerusalem, Israel
mor.nitzan@mail.huji.ac.il

Abstract

Unveiling spatial expression patterns across tissues has been key for studying developmental processes, division of labor mechanisms, as well as variations in health and disease. Along the rapid development of improved experimental assays, computational methods have been shown to successfully recover spatial information from non-spatial single-cell data using reference atlases and/or assumptions about tissue organization such as relative smoothness of expression. However, spatial reconstruction can still be challenging for complex tissues, especially given a limited reference atlas. Here we show how information about tissue microenvironments statistics, such as cell type neighborhoods, or co-localization priors, can enhance tissue reconstruction in such cases. Specifically, we incorporate co-localization priors as a generalization of novoSpaRc, an optimal transport-based framework for tissue reconstruction given single-cell data, which relies at its core on an interpolation between a structural correspondence assumption between expression and physical space and a potential reference atlas. We demonstrate that incorporating cell type co-localization priors can enhance the reconstruction of the mammalian organ of Corti and testicular spatial structure.

Introduction

The collective behavior and division of labor of cells in tissues, in health and disease, relies on their interactions and global organization [1, 2, 3]. Computationally inferring such organization and spatial expression patterns from single-cell data, such as single-cell RNA-sequencing (scRNA-seq) data, is a rapidly-developing field, which includes diverse methodologies, relying on multiple types of prior knowledge [4, 5, 6]. Still, methods struggle with reconstruction of complex, disordered tissues, especially where a reference atlas is limited. Importantly, different tissues, even complex and globally disordered, were shown to exhibit specific spatial features and cell type co-localization rules [7, 8, 9], which thus can potentially contribute to spatial reconstruction.

Here we integrate such spatial features and generalize novoSpaRc, an optimal-transport-based method for tissue reconstruction [4, 10]. novoSpaRc relies on an interpolation between a potentially available reference atlas and a structural correspondence assumption which expresses a hypothesis

Learning Meaningful Representations of Life (NeurIPS 2022).

about an underlying tissue organization principle reflecting relative, averaged smoothness of gene expression across tissues. This assumption, however, generally holds within a single cell type and not for complex, disordered tissues[4]. We leverage information about local cellular microenvironments statistics, and as a proof of concept, focus on cell-type co-localization statistics, as an additional weighted term in the novoSpaRc framework. Specifically, we first construct a co-localization matrix which captures cell-type neighborhood statistics which can either be learned from existing data or known a-priori. Then, the co-localization matrix is integrated into a quadratic term which is added to the objective of the generalized optimal transport formulation of novoSpaRc. Using two examples of the mouse organ of Corti and testis, we show that co-localization prior can be used for both layered and islet-patterned tissues, and can aid in recovering tissue structure and spatial gene expression patterns.

Methods

Given scRNA-seq gene expression data, $X \in \mathbb{R}^{N\times K}$ (N cells and K genes), and a target space of M locations corresponding to the physical tissue structure we wish to reconstruct, we aim to infer an embedding $T \in [0,1]^{N\times M}$ of N cells to M locations which takes into consideration and interpolates between three factors: (1) structural correspondence assumption, (2) a reference atlas, and (3) a co-localization prior, in a way which generalizes the novoSpaRc framework [4, 10] to include local microenvironment statistics. Together, we optimize T via the following generalized optimal-transport formulation:

$$T^* = \underset{T\in C_{P_{cell},P_{loc}}}{\operatorname{argmin}} \; (1-\alpha-\beta)\, C^{smooth}(T) + \alpha C^{atlas}(T) + \beta C^{coLoc} - \varepsilon H\,(T) \tag{1}$$

where $C_{P_{cell},P_{loc}} = \{T | T \in [0,1]^{N\times M},\ T\mathbf{1} = P_{cell},\ \mathbf{1}T = P_{loc}\}$, $\mathbf{1}P_{cell} = 1,\ P_{loc}\mathbf{1} = 1$. The first three terms are interpolated using the non-negative coefficients α, β such that $\alpha + \beta \in [0,1]$.

The first term corresponds to the structural correspondence assumption,

$$C^{smooth}(T) = \sum_{ijkl} L\left(D^{exp}_{ik}, D^{phys}_{jl}\right) T_{ij}T_{kl} \tag{2}$$

where $D^{exp} \in \mathbb{R}^{N\times N}$ is cell-cell similarity matrix, $D^{phys} \in \mathbb{R}^{M\times M}$ is location-location similarity matrix, and L is a loss function which we take to be the quadratic loss (as described in [4, 10]).

The second term corresponds to the discrepancy between an embedding and knowledge regarding a partial reference atlas,

$$C^{atlas}(T) = \sum_{ij} D^{exp,phys}_{ij} \cdot T_{ij} \tag{3}$$

where $D^{exp,phys} \in \mathbb{R}^{N\times M}$ is a cell-location discrepancy matrix between cells and locations according to an available reference atlas (as described in [4, 10]).

The third term corresponds to co-localization priors,

$$C^{coLoc} = \sum_{ijkl} \overline{L}\left(D^{coLoc}_{ik}, D^{neigh}_{jl}\right) T_{ij}T_{kl} \tag{4}$$

where $\overline{L} = a \cdot b$, $D^{neigh} \in \{0,1\}^{M\times M}$ is the kNN adjacency matrix of locations in the target space. D^{coLoc} is a matrix corresponding to the agreement between the cellular embedding and the cell types co-localization prior,

$$D^{coLoc} = -X_{label} \cdot A \cdot {X_{label}}^T \in \mathbb{R}^{N\times N} \tag{5}$$

where $X_{label} \in [0,1]^{N\times t}$ is a labelled single-cell dataset, and $A \in \mathbb{R}^{t\times t}$ is the co-localization matrix, which captures the ‘attraction’ between t different labels, such as cell types. Specifically, $A_{t_1 t_2}$ describes the likelihood of cell type t_1 to be located in the neighborhood (defined below) of cell type t_2.

A can be based either on quantitative estimates of spatially-informed measurements of labels (such as those inferred from Slide-seq [11, 12] captured expression), or based on qualitative description of organization. In the former case, based on label indicators/compositional annotations across the spatial sample, $X'_{label} \in [0,1]^{M' \times t}$ (M' locations in spatial sample, t labels), and their radius-based nearest neighbors adjacency indicator matrix $B \in \{0,1\}^{M' \times M'}$, we compute $A = {X'_{label}}^{T} \cdot B \cdot X'_{label}$. In special cases, we can construct a proxy for the co-localization matrix A without requiring a spatial sample of the data, as we show for the layered structure of the organ of Corti example below.

The fourth term corresponds to entropic regularization, where $H(T) = -\sum_{ij} T_{ij} \log(T_{ij})$, and ε is a non-negative regularization constant.

Expression over the target space, $S \in \mathbb{R}^{K \times M}$, is consequently inferred by $S = X^T T^*$. Likewise, labels over the target space, $X^*_{label} \in [0,1]^{M \times t}$, are inferred by $X^*_{label} = {T^*}^T X_{label}$. We solve this minimization problem using projected gradient descent, a projection based on the Kullback-Leibler metric rewritten as an instance of entropically-regularized optimal transport ([13]), which is then computed using Sinkhorn's fixed point iteration [14], similarly to [4].

Results

Reconstruction of the organ of Corti

The organ of Corti, a hearing receptor located in the mammalian cochlea, consists of ten layers of different cell types that spiral from its base to its apex [15](Figure 1A). Based on data of the transcriptional profiles of cells in the mammalian organ of Corti collected in [15], the variation in expression across the base-apex axis was recently utilized to computationally order the cells along this axis, for each cell type separately, both in the original study [15] and using novoSpaRc [10]. However, using a similar approach to map all cells together onto a 2-dimensional target space representing the unrolled tissue fails in recovering the layered-cell-type structure (Figure 1C, Figure 1E). We used the layered structure exhibited by different cell types in the organ of Corti to construct a co-localization matrix of cell types, reflecting the distances between layers along the medial-lateral axis (Figure 1I). We then mapped all cells simultaneously onto the 2-dimensional target space, using both the cell type co-localization prior and the structural correspondence prior (Figure 1B, Figure 1D, Figure 1F). We find that the two axes of the target space following reconstruction correspond to medial-lateral cell type ordering (X-axis, Figure 1B, Figure 1D) and basal-apical ordering (Y-axis,Figure 1F) of the cells. The cell type co-localization prior is crucial for faithful reconstruction in this case, which is optimized at an intermediate level of interpolation between the two priors ($\beta = 0.3$; Figure 1H).

Reconstruction mouse testicular microenvironment

The testis are a well-structured organ consisting of seminiferous tubules with spermatogonial stem cells at their basement membrane, and separated from each other with interstitial space that contains supporting cells like macrophages and endothelial, and cells involved in the regulation of spermatogenesis like Sertoli, and Leydig cells. Here we aimed to reconstruct a Slide-seq puck of the mouse testis based on its synthetic dissociation [16] (Figure 2A). First, we tested our method on a patch from the Slide-seq puck, where we both constructed the cell-type co-localization matrix and used it to reconstruct the same patch (Figure 2E). Similarly to the organ of Corti reconstruction, the co-localization prior enhances the quality of reconstruction of the testicular spatial structure (Figure 2B, Figure 2C). When interpolating between the co-localization prior and a reference atlas composed of spatial expression of several marker genes (chosen as highly variable genes), we find that while the quality of reconstruction increases with the size of the reference atlas (the number of marker genes used for reconstruction), optimal reconstruction is achieved at intermediate interpolation value ($\beta = 0.4$), leveraging co-localization information (Figure 2D). To further test whether co-localization statistics can be learned from one sample and transferred to recover the organization of another, we divided the Slide-seq puck into 16 patches and tested the contribution of the co-localization prior across patches (Figure 3A, Figure 3B, Figure 3C). We find that even in this more challenging case given an effectively noisy version of a co-localization prior, it contributes to the spatial reconstruction of the tissue with a similar intermediate interpolation value ($\beta = 0.3$; Figure 3D).

Discussion

Here we showed how cell type co-localization statistics can be used to enhance the spatial reconstruction of complex tissues, including the mammalian organ of Corti and testicular spatial structure. We built on the novoSpaRc [4, 10] framework and demonstrated how interpolating between terms capturing information about either the structural correspondence prior or a reference atlas and co-localization statistics can improve the quality of spatial tissue reconstruction, compared to using only the former terms, respectively.

In this work we aimed to show the potential for using spatial features, such as cell type co-localization statistics, to enhance spatial reconstruction of single-cell data. We envision that the analysis and incorporation of additional, more complex and more diverse, spatial features could improve tissue reconstruction further, although such complexity will pose more challenging optimization problems. More complex spatial features could potentially be analyzed and incorporated using flexible frameworks such as Graph Neural Networks. Together, improving computational methodologies for tissue reconstruction will aid in exposing the structural design principles that shape functional, ordered and disordered tissues.

Acknowledgments and Disclosure of Funding

We thank Zoe Piran and members of our lab for meaningful discussions, and Shahar Nahum for the illustration of the organ of Corti. This work was funded by the Center for Interdisciplinary Data Science Research at the Hebrew University of Jerusalem (NM and MN), the Council of Higher Education (NM), Azrieli Foundation Early Career Faculty Fellowship, Google Research Scholar program, and the European Union (ERC, DecodeSC, 101040660) (MN). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. The authors declare no competing interests.

References

[1] Marietta Zinner, Ilya Lukonin, and Prisca Liberali. “Design principles of tissue organisation: How single cells coordinate across scales”. In: *Current Opinion in Cell Biology* 67 (2020), pp. 37–45.

[2] Aline Xavier da Silveira dos Santos and Prisca Liberali. “From single cells to tissue self-organization”. In: *The Febs Journal* 286.8 (2019), pp. 1495–1513.

[3] Evan Heller and Elaine Fuchs. “Tissue patterning and cellular mechanics”. In: *The Journal of Cell Biology* 211.2 (2015), pp. 219–231.

[4] Mor Nitzan et al. “Gene expression cartography”. In: *Nature* 576.7785 (2019), pp. 132–137.

[5] Xianwen Ren et al. “Reconstruction of cell spatial organization from single-cell RNA sequencing data based on ligand-receptor mediated self-assembly”. In: *Cell Research* 30.9 (2020), pp. 763–778.

[6] Zixuan Cang and Qing Nie. “Inferring spatial and signaling relationships between cells from single cell transcriptomic data”. In: *Nature Communications* 11.1 (2020), p. 2084.

[7] Andrea Behanova, Anna Klemm, and Carolina Wählby. “Spatial Statistics for Understanding Tissue Organization”. In: *Frontiers in Physiology* 13 (2022).

[8] G. Palla et al. “Spatial components of molecular tissue biology”. In: *Nature Biotechnology* (2022).

[9] Inbal Avraham-Davidi et al. “Integrative single cell and spatial transcriptomics of colorectal cancer reveals multicellular functional units that support tumor progression”. In: *bioRxiv* (2022).

[10] Noa Moriel et al. “NovoSpaRc: flexible spatial reconstruction of single-cell gene expression with optimal transport”. In: *Nature Protocols* 16.9 (2021), pp. 4177–4200.

[11] Samuel G. Rodriques et al. “Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution”. In: *Science (New York, N.Y.)* 363.6434 (2019), pp. 1463–1467.

[12] Robert R. Stickels et al. “Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2”. en. In: *Nature Biotechnology* 39.3 (2021), pp. 313–319.

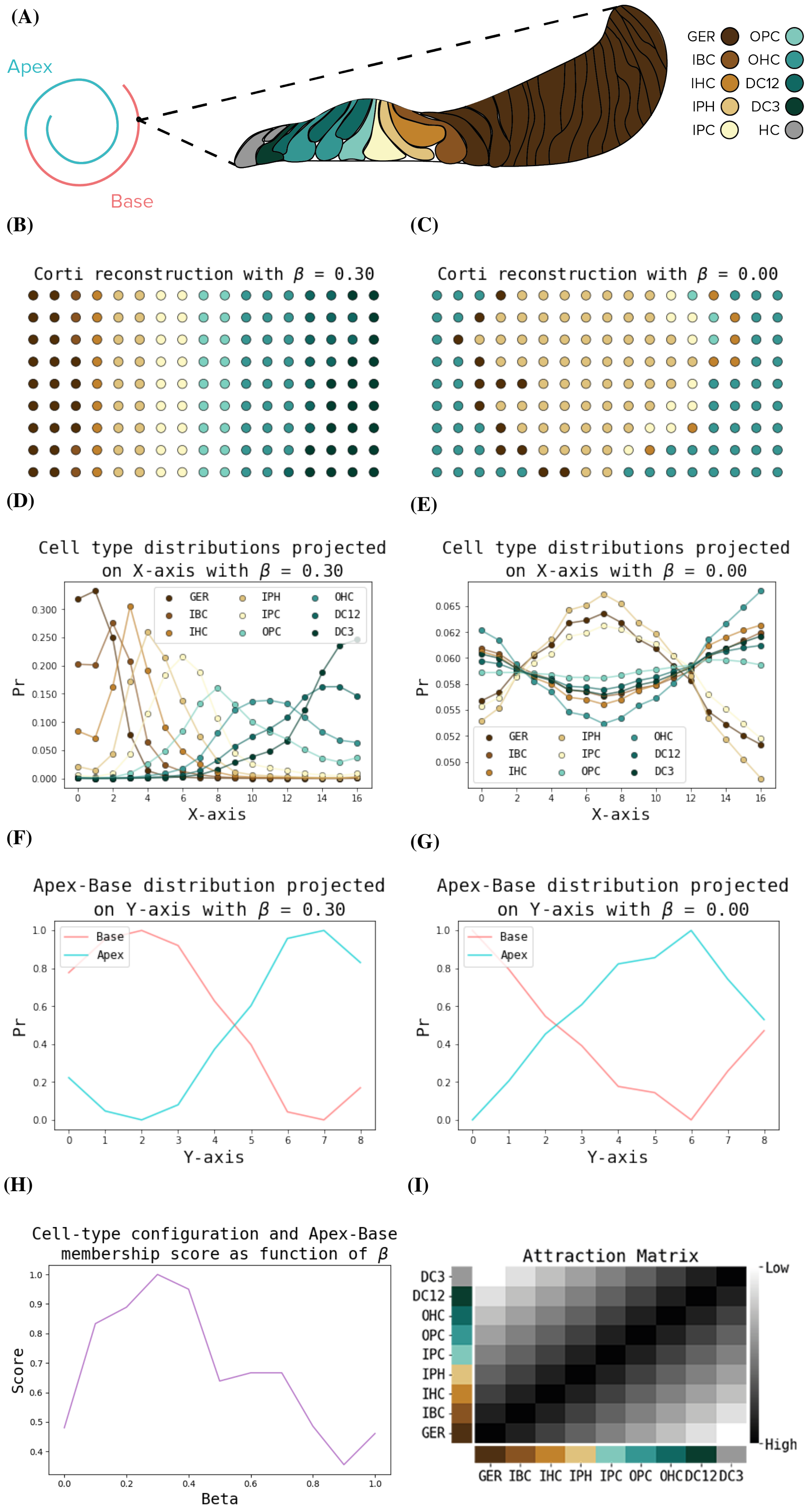
[13] Jean-David Benamou et al. “Iterative Bregman Projections for Regularized Transportation Problems”. In: *SIAM Journal on Scientific Computing* 37.2 (2015).

[14] Richard Sinkhorn. “Diagonal Equivalence to Matrices with Prescribed Row and Column Sums”. In: *The American Mathematical Monthly* 74.4 (1967), pp. 402–405.

[15] Jörg Waldhaus, Robert Durruthy-Durruthy, and Stefan Heller. “Quantitative High-Resolution Cellular Map of the Organ of Corti”. In: *Cell Reports* 11.9 (2015), pp. 1385–1399.

[16] Haiqi Chen et al. “Dissecting mammalian spermatogenesis using spatial transcriptomics”. In: *Cell Reports* 37.5 (2021), p. 109915.

Figure 1: Organ of Corti spatial reconstruction using co-localization priors. **(A)** The organ of Corti illustration. Base-apex and medial-lateral axes (GER to HC). 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), outer hair cell (OHC), Deiters’ cell row 1–2 (DC12) and Deiters’ cell row 3 (DC3), Hensen’s cell (HC). **(B)-(C)** Single cells were spatially reconstructed using a 2-dimensional grid. Grid locations are colored according to the cell type which received the highest probability in the reconstruction for that location, with (left) and without (right) co-localization priors **(D)-(E)** Reconstructed cell type distributions projected on the X-axis. **(F)-(G)** Reconstructed base-apex average membership distributions projected on the Y-axis. **(H)** Combined averaged cell-type configuration score (number of fixed points) and apex-base score (AUC score) as function of β value. **(I)** Attraction matrix heatmap. Parameters: $(\alpha, \beta, 1-\alpha-\beta) = (0, 0.3, 0.7)$ [left column, using co-localization prior, sub-figures **B, D , F**] and (0, 0, 1) [right column, only smoothness, sub-figures **C, E , G**]; $\varepsilon = 5e^{-3}$


(A)
Apex
Base
GER
IBC
IHC
IPH
IPC
OPC
OHC
DC12
DC3
HC
(B)
Corti reconstruction with β = 0.30
(C)
Corti reconstruction with β = 0.00
(D)
Cell type distributions projected on X-axis with β = 0.30
GER
IBC
IHC
IPH
IPC
OPC
OHC
DC12
DC3
Pr
0.000
0.050
0.100
0.150
0.200
0.250
0.300
X-axis
0
2
4
6
8
10
12
14
16
(E)
Cell type distributions projected on X-axis with β = 0.00
GER
IBC
IHC
IPH
IPC
OPC
OHC
DC12
DC3
Pr
0.050
0.052
0.055
0.058
0.060
0.062
0.065
X-axis
0
2
4
6
8
10
12
14
16
(F)
Apex-Base distribution projected on Y-axis with β = 0.30
Base
Apex
Pr
0.0
0.2
0.4
0.6
0.8
1.0
Y-axis
0
1
2
3
4
5
6
7
8
(G)
Apex-Base distribution projected on Y-axis with β = 0.00
Base
Apex
Pr
0.0
0.2
0.4
0.6
0.8
1.0
Y-axis
0
1
2
3
4
5
6
7
8
(H)
Cell-type configuration and Apex-Base membership score as function of β
Score
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Beta
0.0
0.2
0.4
0.6
0.8
1.0
(I)
Attraction Matrix
DC3
DC12
OHC
OPC
IPC
IPH
IHC
IBC
GER
GER
IBC
IHC
IPH
IPC
OPC
OHC
DC12
DC3
Low
High

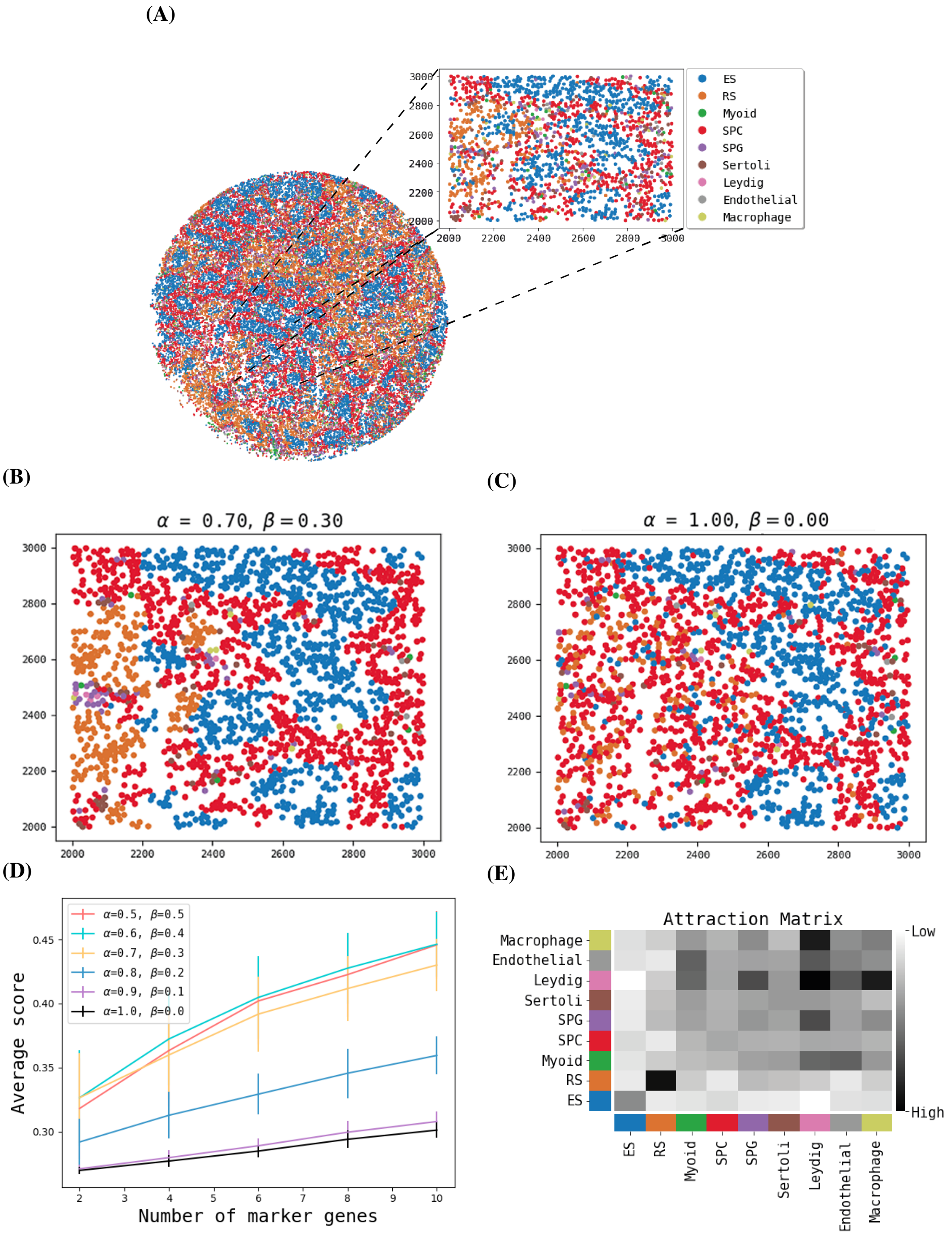


Figure 2: Mouse testis spatial reconstruction using co-localization priors on a single patch of a Slide-seq puck. (A) Focusing in on a patch from a Slide-seq puck of the mouse testis. Cell-type abbreviations: elongating/elongated spermatid (ES), round spermatid (RS), spermatocyte (SPC), spermatogonium (SPG). **(B)** Reconstructed cell type argmax across the patch using two testis marker genes (TNP1 and SYCP1), with a co-localization prior, $(\alpha, \beta, 1-\alpha-\beta) = (0.7, 0.3, 0)$ and **(C)** without a co-localization prior, $(1, 0, 0)$. **(D)** Average reconstruction score as function of number of marker genes (over 40 iterations). Marker genes were randomly sampled among top 100 highly variable genes. **(E)** Mouse testis attraction matrix heatmap. Parameters: $D_1(T)$, $D_2(T)$ were calculated as in [10] with $k = 5$ and $\varepsilon = 5e^{-3}$.

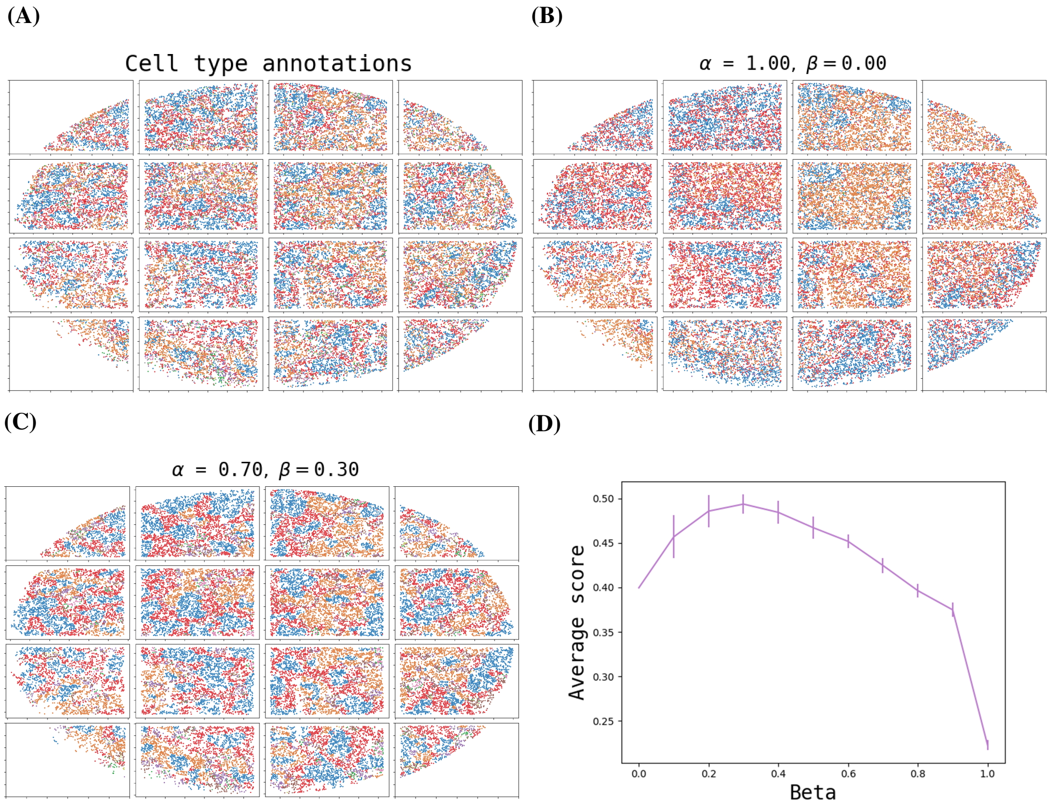


Figure 3: Mouse testis spatial reconstruction using co-localization priors across patches of a Slide-seq puck. **(A)** The full Slide-seq puck split *in silico* to 16 patches. Colors represent cell-type annotations as specified in Figure 2A. **(B)** Reconstructed cell type argmax across the full puck using only a reference atlas prior (based on marker genes TNP1 and and SYCP1). **(C)** Reconstructed cell type argmax across the full puck using both the reference atlas and additionally, co-localization prior based on a single patch. **(D)** Average cell-type reconstruction score as a function of β value. Error bars represent standard deviation of the crossover co-localization-learning and testing across all 16 patches. For $\beta = 1$, no reference atlas information means that the reconstruction is not anchored spatially and therefore does not directly correlate to the original tissue structure.