**SpaOTsc: spatial optimal transport of single cell transcriptomics data** (10 words or 90 characters)

Zixuan Cang1,3 and Qing Nie1,2,3\*

1 Department of Mathematics, 2 Department of Developmental and Cell Biology, and 3 The NSF-Simons Center for Multiscale Cell Fate Research, University of California, Irvine, CA 92697, USA

**Single-cell RNA sequencing (scRNA-seq) provides unprecedented expression details for individual cells; however, crucial spatial information is lost** **. We present SpaOTsc, a toolbox relying on structured optimal transport to recover spatial properties of scRNA-seq data by utilizing other spatial measurements of a small number of genes. Space-constrained cell-cell communications and intercellular gene-gene information flows are inferred using partial information decomposition and ensemble of trees.** (3 sentences and no more than 70 words)

The diverse structures and functions of multicellular organisms arise from regulated heterogeneity among cells that form tissues. Single-cell transcriptomic methods enable analysis of the heterogeneity of gene expressions in individual cells1. Dissociating tissues into single cells is a necessary step for high-throughput genomics, but this causes loss of crucial spatial information. Measuring gene expression in intact tissues provide spatial resolutions; however, the small number of genes examined need to be selected in advance. Is it possible to provide spatial information for the single cells with transcriptomics measured?

Recent computational methods have attempted to pair scRNA-seq data with spatial information using other spatial imaging data (e.g. *in situ* hybridization). Spatial information was inferred at the cell cluster levels by identifying spatial domains with coherent gene expressions in spatial imaging data combined with scRNA-seq data2. At an individual cell level, a similarity measurement based on correlation coefficients3 or correspondence scores4 between commonly examined genes in both spatial imaging data and scRNA-seq data was used to reconstruct spatial gene expression or map cells in scRNA-seq data to their potential spatial origins. Posterior probability estimates on spatial data described by a mixture model5 or simplified to one-dimensional bins6 were used to assign spatial origins to individual cells . Canonical correlation was used to connect cells in scRNA-seq data to their locations in spatial data7 to identify anchors across the both datasets through data integration8. Finally, non-negative matrix factorization method could also be used to construct common low-dimensional spaces of multiple datasets9. The goal of those existing methods is to assign individual cells in scRNA-seq data with estimated spatial origins in a tissue.

Here, we present a different method to integrate the spatial data using optimal transport10 with scRNA-seq data (SpaOTsc). Recently, optimal transport has been applied to various problems such as inferring developmental trajectories from scRNA-seq data11 and handling batch effects12. In this study we utilize a modified optimal transport method to not only infer spatial distances between individual cells in scRNA-seq data but also map those cells to their spatial origins.

Unlike existing approaches that emphasizes a “score” connecting individual cells and the spatial locations in pairs, SpaOTsc treats the two datasets as two distributions, and uses dissimilarity measurements within each dataset to refine a mapping between these two distributions through the structured optimal transport13, leading to probability distributions of individual cells over space (**Fig. 1a**). Next, the spatial distance between any two cells in the scRNA-seq data, referred to as the *cell-cell distance*, is inferred by computing the Wasserstein distance between these distributions (**Fig. 1b**), which is then used for *spatial cell clustering*. Similarly, by interpreting genes as distributions over cells annotated with cell-cell distance, one can quantify the differences between spatial gene expression patterns to generate a *spatial gene atlas*.

To identify possible communications among cells mediated by ligand-receptor interactions, we formulate an optimal transport problem that transports a distribution of ligand expression on cells to a target distribution described by receptors and downstream target genes. The cell-cell distance is used as the “transport cost” to constrain the signaling, and the corresponding optimal transport plan is interpreted as likelihoods for *cell-cell communications*. To further infer under what spatial range one gene in a cell may regulate another gene in a different cell due to ligand-receptor communications, we analyze a collection of trained random forest models with the downstream genes as outputs and the receptors as sample weights. The genes that highly correlate to the downstream genes and the ligands from cells located within a spatial range are the input features, and the ligand feature importance indicates how likely the communications happen under this spatial range (**Fig. 1c**). We compute the *spatial map of intercellular gene-gene regulatory information flow* (“spatial map")by estimating the unique information about a gene in a cell provided by another gene expressed in its neighboring cells, taking into account the information given by a collection of other genes in this cell.

We applied SpaOTsc to different zebrafish embryo datasets5, and carried out leave-one-out cross-validation on the spatial pattern of landmark genes, with scRNA-seq data and the training spatial data used for prediction. Both representative predictions and the average AUC show good accuracy of SpaOTsc (**Fig. 1d,e**). Similar analysis was carried out for a drosophila embryo dataset3, which showed consistent accuracy and performance (**Fig. 1f,g**). We found that the incorporation of both unbalanced and structured optimal transport delivers more accurate predictions than conventional optimal transport (**Supplementary Fig. 1-3**). We also analyzed a mouse visual cortex dataset15, 16 which was previously annotated with spatial information. SpaOTsc had a micro F1 score of 0.48 as characterized by multiclass classification (**Fig. 1h,i**).

Next, we used the inferred cell-cell distance for scRNA-seq data to further separate cell clusters and to identify their potential spatial origins (**Fig. 2a-c** and **Supplementary Fig. 4-8**). Classical dimension reduction methods, such as tSNE, cannot distinguish such spatially separated subclusters presented in the scRNA-seq data (**Fig. 2a**). We quantitatively compared the spatial expression patterns of a set of highly variable genes, and categorized them into distinct modes of spatial gene expression patterns (**Supplementary Fig. 9,10**), such as clusters of genes which are highly expressed in the L2/3, L5, and L6 regions in the mouse visual cortex dataset (**Fig. 2d**).

Finally, we used SpaOTsc to analyze cell-cell communications for multiple drosophila embryo datasets3, with a focus on Wingless (Wg) and Decapentaplegic (Dpp) signalings (**Fig. 2e-h**). Wg signaling was previously shown to act in a range of 50 to 100 µm17. The effective distance inferred using SpaOTsc was about 100 µm (**Fig. 2i** top). After estimating the probability of signaling between each pair of cells constrained by the spatial distance, the cell-cell communications could be summarized to the communications among subclusters (**Fig. 2f**). Interestingly, a thin strip of cells located near the lateral-ventral part of the embryo were found to be both sources and targets of Wg signaling (**Fig. 2f**), and Wg signaling was abundant at the lateral side of the embryo with a bias toward the posterior. This finding explains a previous observation that Wg signaling is crucial to growth of the posterior18, and predicts that cells at the posterior-ventral domain receive Wg signaling from their neighbors.

The Dpp signaling was found to have a longer effective distance of 120 µm (**Fig. 2i** bottom), and the strongest Dpp communications occurred at the lateral where Sog was predicted to be abundant (**Fig. 2h**), supporting a prior result that Dpp signaling undergoes long-range transport facilitated by Sog during dorsal-ventral patterning19. Interestingly, the strong Wg source located near the ventral side was also identified as a strong target of Dpp, receiving signals from the dorsal side (**Fig. 2h**). Both Wg and Dpp cell-cell communications inferred by SpaOTsc were compared with another inference method20 without using spatial information (**Supplementary Fig. 11-13**), showing biological connections more consistent with the literature. Finally, we inferred a spatial map for a set of highly variable genes under different spatial ranges to predict which gene in one cell may affect another gene in a different cell located within an estimated maximal distance (**Fig. 2j**).

,We have shown the capabilities of SpaOTsc to (1) map between scRNA-seq data and spatial data, (2) infer spatial distances between single cells, (3) quantitatively compare spatial gene expression patterns, and (4) reconstruct cell-cell spatial communications. The capabilities of SpaOTsc have been validated by gene expression reconstruction of zebrafish embryo and drosophila embryo datasets, along with assigning spatial origins to the scRNA-seq data of mouse visual cortex, have demonstrated good mapping accuracy of SpaOTsc. By attaching spatial distance to scRNA-seq data using SpaOTsc, one can carry out spatial analysis of all genes at single-cell resolution, which retains cellular heterogeneity. Integrating such spatial analysis with network inference techniques enables construction of spatially constrained intercellular signaling to identify pairs of signal sending/receiving cells or cell clusters. SpaOTsc is generally applicable to datasets where reasonable similarity measurement between single cells and spatial locations are obtainable.

**ACKONWLEDGEMENTS**

This work was supported by a NIH grant U01AR073159, a NSF grant DMS1763272, and a grant from the Simons Foundation (594598, QN).

**REFERENCES**

1. Svensson, V., Vento-Tormo, R. & Teichmann, S.A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat Protoc* **13**, 599-604 (2018).

2. Zhu, Q., Shah, S., Dries, R., Cai, L. & Yuan, G.C. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. *Nat Biotechnol* (2018).

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

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

5. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **33**, 495-U206 (2015).

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

7. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).

8. Stuart, T. et al. Comprehensive integration of single cell data. *BioRxiv*, 460147 (2018).

9. Welch, J. et al. Integrative inference of brain cell similarities and differences from single-cell genomics. *bioRxiv*, 459891 (2018).

10. Chizat, L., Peyre, G., Schmitzer, B. & Vialard, F.X. Scaling Algorithms for Unbalanced Optimal Transport Problems. *Math Comput* **87**, 2563-2609 (2018).

11. Schiebinger, G. et al. Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell* **176**, 1517 (2019).

12. Forrow, A. et al. in The 22nd International Conference on Artificial Intelligence and Statistics 2454-2465 (2019).

13. Vayer, T., Chapel, L., Flamary, R., Tavenard, R. & Courty, N. Optimal Transport for structured data. *arXiv preprint arXiv:1805.09114* (2018).

14. Chan, T.E., Stumpf, M.P. & Babtie, A.C. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell systems* **5**, 251-267. e253 (2017).

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

16. Tasic, B. et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72-78 (2018).

17. Waghmare, I. & Page-McCaw, A. Wnt Signaling in Stem Cell Maintenance and Differentiation in the Drosophila Germarium. *Genes (Basel)* **9** (2018).

18. Martin, B.L. & Kimelman, D. Wnt signaling and the evolution of embryonic posterior development. *Curr Biol* **19**, R215-219 (2009).

19. Wang, Y.C. & Ferguson, E.L. Spatial bistability of Dpp-receptor interactions during Drosophila dorsal-ventral patterning. *Nature* **434**, 229-234 (2005).

20. Wang, S., Karikomi, M., MacLean, A.L. & Nie, Q. Cell lineage and communication network inference via optimization for single-cell transcriptomics. *Nucleic Acids Research* (2019).

**CAPTIONS**

**Figure 1** (**a**) In SpaOTsc the unbalanced transport relaxes the mass conservation constraint (e.g. lines between circles), and the structured transport utilizes additional information (e.g. dotted links) to refine the mapping (e.g. blue hexagon). (**b**) SpaOTsc infers cell-cell distance by computing Wasserstein distance of the distributions (rows of  in **a**). (**c**) Calculated cell-cell distance, along with partial information decomposition and random forest models, was used to infer spatial distance of signaling and then construct space-constrained cell-cell communications. (**d**) Predicted spatial expressions for the zebrafish embryo (spatial data: 64 bins and 47 genes; scRNA-seq data: 851 cells and 10946 genes)5. (**e**) Leave-one-out cross-validation of the spatial expression prediction for the zebrafish embryo data. (**f**) Predicted spatial expressions for the drosophila embryo (spatial data: 3039 locations and 84 genes; scRNA-seq data: 1297 cells and 8925 genes)3. (**g**) Leave-one-out cross-validation for the drosophila embryo spatial data. (**h**) Assignment of spatial origins to the scRNA-seq data for the mouse visual cortex (spatial data: 1549 locations and 1020 genes15; scRNA-seq data: 15413 cells and 45768 genes16). Each column depicts all cells from the spatial data in visual cortex. In column one, the color of cells represents the average probability of the spatial origin of the 890 cells in scRNA-seq data labeled with spatial origin L1. Columns two to five represent similar probabilities for the single cells known to possess four other spatial regions. (**i**) Violin plots along L1-L6 axis of the mapped spatial origins for single cells from each subregion.

**Figure 2** (**a**) Clustering of zebrafish embryo scRNA-seq data using the tSNE method. (**b**) Further SpaOTsc spatial subclustering of (a) based on spatial cell-cell distance constraints, showing three distinct subclusters 1, 2, 3. (**c**) The three subclusters are spatially localized in zebrafish embryo. (**d**) Top: Clustering of genes from mouse visual cortex spatial dataset using the knn-graph+Louvain method where the knn-graph is constructed based on SpaOTsc gene-gene distance. Bottom: Average spatial patterns of gene for three different clusters. (**e**) Cell-cell communications of Wg signaling at the single-cell level using a visualization constrained by cell-cell distance. The color of the link is marked by the color of the sending cells, based on the non-spatial clustering similar to (a). (**f**) Cluster-cluster communication of Wg signaling based on SpaOTsc spatial subclustering. (**g**,**h**) Dpp signaling in space plotted similar to (e, f). (**i**) Spatial ranges of Wg and Dpp signaling inferred using consensuses of random forest models. (**j**) A spatial map of intercellular gene-gene regulatory information flow for the top 20 variable genes in drosophila embryo scRNA-seq data. For example, gene Twist in the 20 μm shell is connected with gene Snail (red curve), suggesting Snail is directly or indirectly affected by Twist in neighbor cells within a spatial distance of 20 μm. These two genes are known to be important during mesoderm formation.

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

5. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **33**, 495-U206 (2015).

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

16. Tasic, B. et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72-78 (2018).

**METHODS**

Full details of the theoretical background and implementation of SpaOTsc can be found in **Supplementary Note 1**.

**SpaOTsc model.**

SpaOTsc constructs a mapping between the  cells in scRNA-seq data and the  locations in spatial data by solving an optimal transport problem21 given three dissimilarity/distance matrices,  for the gene expression dissimilarity between cells and locations,  for the gene expression dissimilarity among cells, and  for the distances among spatial locations. The optimal transport plan  is obtained by solving



where is a weight vector and measures the difference between scaled dissimilarities/distances. The first term quantifies the major transport cost, the second penalty term promotes weight conservation (unbalanced transport)10, and the last term preserves the distance within datasets through the mapping (structured transport)13. The spatial cell-cell distance  is then computed based on  using the Wasserstein metric:



One can carry out three major tasks immediately after obtaining  and : (1) prediction of spatial gene expression by  where is a scaled version of  with each row summing to  and  is the expression vector for a gene in scRNA-seq data; (2) identification of spatially localized cell subclusters by distance-based clustering using  within each previously identified cluster; and (3) visualization of scRNA-seq data constrained by cell-cell distances using the distance matrix .

The intercellular gene-gene regulatory information flow is inferred by using partial information decomposition14, 22, 23. We estimate how much unique information about a gene (target gene) can be provided by another gene (source gene) in its spatial neighborhood through the calculation of the accumulated unique information:



where is the variable for target gene expression in the cells,  is the variable for source gene expression in  neighborhoods of cells whose observation is estimated using , and is a collection of genes with high intracellular correlation with the target gene. The unique information  measures how much unique information provides about in addition to . For the case of intercellular signaling with known ligands, receptors, and their downstream genes, we use random forest models24, 25 to infer the spatial distance of signaling. The ligand expressions of cells in a neighborhood of distance of  ,denoted as , together with other genes highly correlated to a downstream target gene of the ligand-receptor interaction are used as features to fit a random forest model outputting the target gene. The receptor expressions are used as sample weights. The  under which  has the highest feature importance is considered to be the spatial distance of this signaling.

Knowing the ligands, receptors and downstream genes involved in intercellular signaling and , we then infer cell-cell communication by solving another optimal transport problem. First, the source distribution over the cells is constructed to be proportional to expression of ligand gene. Next a destination distribution  is constructed based on the expression of receptors and downstream genes to represent the probability of a cell to receive the signal. And a cell highly expressing receptors with downstream genes consistent with the up-/down-regulation relationships is assigned with a high probability. With this information we solve the following optimal transport problem



The optimal transport plan  is interpreted as likelihoods of cell-cell communications, e.g. its element describes how likely cell  receives signal from cell . When spatial distances for signaling are available, we can simply adjust the cost matrix  by setting entries greater than this distance to a large number to enforce a spatial constraint on communications identification.

**Data sets and processing.**

*Zebrafish embryo.* We downloaded the accompanying data files (<https://www.dropbox.com/s/ev78jelev0jgu5s/seurat_files_zfin.zip?dl=1>) for the seurat tutorial (<https://satijalab.org/seurat/seurat_spatial_tutorial_part1.html>). The scRNA-seq data is stored in the file “zdata.matrix.txt” and the spatial data (*in situ* hybridization) is stored in “Spatial\_ReferenceMap.xlsx”5. The scRNA-seq data is also available through the accession code GEO: [GSE66688](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE66688). We binarized the scRNA-seq data and selected a set of highly variable genes following the same tutorial. For the scRNA-seq data matrix , a log transformation was performed elementwise for the analyses.

*Drosophila embryo.* The scRNA-seq data and the spatial data (*in situ* hybridization) were downloaded from the Dream Single cell Transcriptomics Challenge through Synapse ID (syn15665609)3. The files “bdtnp.txt” and “binarized\_bdtnp.csv” were used for numerical and binary spatial data, respectively. The files “dge\_normalized.txt” and “dge\_binarized\_distMap.csv” were used for the numerical and binary scRNA-seq data. The coordinate of each cell in the spatial data is assigned according to the file “geometry.txt”. We used Scanpy26 to select highly variable genes for downstream analysis (the script used is included in SpaOTsc tutorial files).

*Mouse visual cortex.* The spatial data (**S**patially-resolved **T**ranscript **A**mplicon **R**eadout **map**ping) was downloaded from STARmap Resources (<https://www.dropbox.com/sh/f7ebheru1lbz91s/AABYSSjSTppBmVmWl2H4s_K-a?dl=0>)15. We used the data named “20180505\_BY3\_1kgenes” from folder “visual\_1020”. The scRNA-seq data was downloaded from Allen Brain Atlas16, 27 (<http://celltypes.brain-map.org/api/v2/well_known_file_download/694413985>), and specifically the file “mouse\_VISp\_2018-06-14\_exon-matrix.csv” was used. The spatial data contains 1020 genes and quantifying similarity by directly computing correlation coefficients might include too much noise and inconsistency across datasets. Therefore, we used the “cca” utility in Seurat7 which determines a low-dimensional common space for the two datasets and the script for processing is included in SpaOTsc tutorial files.

**Code availability.** An open source Python implementation of SpaOTsc is available at GitHub (<https://github.uci.edu/zcang/SpaOTsc>). Tutorials are included to reproduce the presented analyses.

**Data availability.** All data used for analysis are also available at the GitHub repository (<https://github.uci.edu/zcang/SpaOTsc>).

**ADDITIONAL REFERENCES**

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

5. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **33**, 495-U206 (2015).

7. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).

10. Chizat, L., Peyre, G., Schmitzer, B. & Vialard, F.X. Scaling Algorithms for Unbalanced Optimal Transport Problems. *Math Comput* **87**, 2563-2609 (2018).

13. Vayer, T., Chapel, L., Flamary, R., Tavenard, R. & Courty, N. Optimal Transport for structured data. *arXiv preprint arXiv:1805.09114* (2018).

14. Chan, T.E., Stumpf, M.P. & Babtie, A.C. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell systems* **5**, 251-267. e253 (2017).

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

16. Tasic, B. et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72-78 (2018).

21. Flamary, R. & Courty, N. (2017).

22. Williams, P.L. & Beer, R.D. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515* (2010).

23. James, R.G., Ellison, C.J. & Crutchfield, J.P. dit: a Python package for discrete information theory. *Journal of Open Source Software* **3**, 738 (2018).

24. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R news* **2**, 18-22 (2002).

25. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *Journal of machine learning research* **12**, 2825-2830 (2011).

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

27. Tasic, B. et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat Neurosci* **19**, 335-346 (2016).