Cellular Location from Single-Cell and Spatial Transcriptomics Using Machine Learning Method

Bang Tran

Department of Computer Science, College Of Engineering & Computer Science, California State University

Contact: s.tran@csus.edu, Website: https://webpages.csus.edu/s.tran/

Background

Spatial transcriptomics (ST) that was first featured in 2020 [1] can both profile the transcriptome of the cells and preserve its spatial information within tissue section. As the technology underwent rapid development in recent years, spatial transcriptomics technologies have become primary tools for biologists to understand cells, their microenvironments [2], tumor development [3], and treatment response [4]. However, the technologies are still in early stage where the assays can only measure small regions with mixtures of cells and are unable to provide single-cell information.

Objectives

We present Single-cell and Spatial transcriptomics Alignment (SSA), a novel technique that employs an optimal transport algorithm to assign individual cells from a scRNA-seq atlas to their spatial locations in actual tissue based on their expression profiles.

Results

Data: Downloaded dataset contains 100,064 cells with known. We transform the high-resolution ST data into 01 low-resolution ST dataset and 10 scRNA-seq datasets.

Metric: Euclidean distance, Manhattan distance**,** and KL-divergence [5]

Methods: four state-of-the-art methods, SpaOTsc [6], Tangram [7], Seurat [8], and DistMap [9]

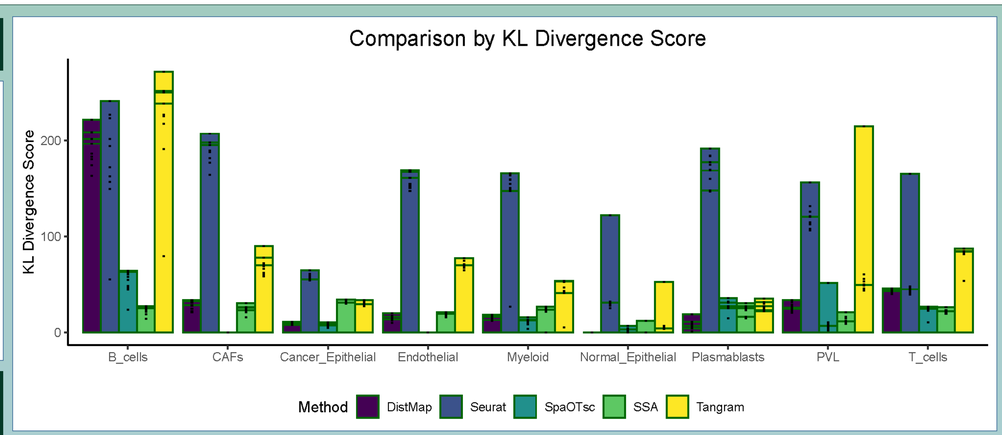
Results: SSA can recover the cells' spatial location with minimal difference and lowest KL-divergence score for each cell type.

Table 1: Comparisons using average Euclidean distance.

Datasets	SSA	SpaoTsc	Tangram	Seurat	DistMap
Dataset-1	**1.786**	3.007	5.364	11.627	4.624
Dataset-2	**1.802**	2.977	5.365	12.057	4.718
Dataset-3	**1.778**	2.952	5.351	11.998	4.534
Dataset-4	**1.818**	2.955	5.291	11.821	5.345
Dataset-5	**1.757**	2.970	5.357	12.072	4.519
Dataset-6	**1.738**	2.944	5.376	11.806	4.804
Dataset-7	**1.784**	2.955	5.388	12.274	4.989
Dataset-8	**1.799**	2.991	5.296	11.751	4.484
Dataset-9	**1.806**	3.076	5.351	11.961	1398.808
Dataset-10	**1.789**	2.939	5.407	12.200	4.502

Table 2: Comparisons using average Manhattan distance.

Datasets	SSA	SpaoTsc	Tangram	Seurat	DistMap
Dataset-1	**2.259**	3.819	6.878	14.769	5.839
Dataset-2	**2.277**	3.781	6.874	15.309	5.967
Dataset-3	**2.245**	3.755	6.862	15.179	5.727
Dataset-4	**2.298**	3.758	6.771	14.98	6.701
Dataset-5	**2.227**	3.778	6.858	15.333	5.705
Dataset-6	**2.193**	3.747	6.895	15.096	6.040
Dataset-7	**2.252**	3.750	6.925	15.603	6.279
Dataset-8	**2.272**	3.801	6.797	14.912	5.663
Dataset-9	**2.280**	3.918	6.858	15.229	1697.566
Dataset-10	**2.263**	3.736	6.930	15.11	5.677

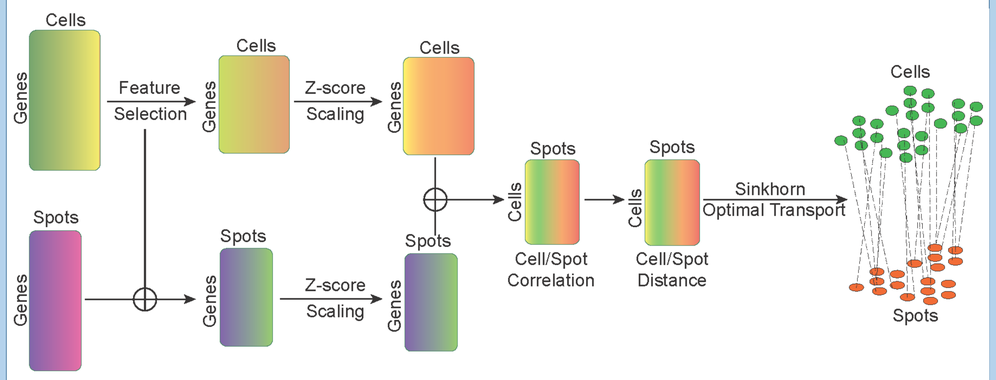


Methodology

Feature Selection and Data Transformation: Select 5,000 genes with the highest variance and use Z-score transformation to scale and center the data.

Cell to Spot Alignment using Sinkhorn Algorithm:

- Given two X and Y as the scaled scRNA-seq and ST matrices, we calculate the pairwise Pearson's correlation. Then, we calculate the pair wise distance between cells and spots.
- Given the distance matrix, we will use Sinkhorn algorithm to compute the optimal transport plan from cells-to-spots. This step involves solving an optimization problem that seeks to find the "cheapest" way to transport mass from the cells to the spots, where the "cost" of transporting mass is given by the distance matrix.
- The output of the Sinkhorn algorithm is a matrix Tm×n where each value represents the mass of a cell transported to a spot. We then transform it into a probability matrix with the same dimension and assign cells to spots based on the maximum probability.



Conclusion

- Outperforms existing state-of-the-art approaches.
- scCAN is the fast method for big data.
- scCAN is robust to dropouts.
- scCAN is the best method to predict true number of cell types.

Future work

Expanding scan to work with other data types such as multi-omics data [10].

Acknowledgement

NSF (grant no. 2343019 and 2203236), NASA (grant no. 80NSSC22M0255, subaward 23-42), NIGMS (grant no. 1R44GM152152-01), NCI (grant no. 1U01CA274573-01A1),and California State University, Sacramento Probationary Faculty Development Grant.

References

1. A Xiaowei. et al. Method of the Year 2020: Spatially resolved transcriptomics. Nature Methods, 18(1), 2021.
2. Leeat K. et al. A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. Cell, 174(6):1373–1387, 2018.
3. Christian M. . et al. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. Cell, 182(5):1341–1359, 2020.
4. Rodrigo N .et al. Tissue-resident FOLR2+ macrophages associate with CD8+ T cell infiltration in human breast cancer. Cell, 185(7):1189–1207, 2022.
5. Bogdan A. et al. Atlas of clinically distinct cell states and ecosystems across human solid tumors. Cell, 184(21):5482–5496, 2021.
6. Zixuan Cang. et al. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. Nature Communications, 11:2084, 2020.
7. Tommaso B. et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. Nature Methods, 18(11):1352–1362, 2021.
8. Stuart T. et al. Comprehensive integration of single-cell data. Cell, 177(7):1888–1902, 2019.
9. Karaiskos N. et al. The Drosophila embryo at single-cell transcriptome resolution. Science, 358(6360):194–199, 2017.
10. Nguyen et al. (2017). A novel approach for data integration and disease subtyping. Genome research, 27(12), 2025-2039.