EPiC Series in Computing

Volume 101, 2024, Pages 25–38

Proceedings of the 16th International Conference on Bioinformatics and Computational Biology (BICOB-2024)

SSA: a novel method for Single-cell and Spatial transcriptomics Alignment

Bang Tran[1], Dao Tran[2], and Tin Nguyen*[2]

[1] Department of Engineering & Computer Science, California State University
6000 J St, Sacramento, CA 95819
s.tran@csu.edu

[2] Department of Computer Science and Software Engineering
345 W Magnolia Ave, Auburn, AL 36849
dqt0001@auburn.edu

[3] Department of Computer Science and Software Engineering
345 W Magnolia Ave, Auburn, AL 36849
tinn@auburn.edu

Abstract

Single-cell RNA sequencing (scRNA-seq) provides expression profiles of individual cells but fails to preserve crucial spatial information. On the other hand, Spatial Transcriptomics technologies are able to analyze specific regions within tissue sections, but lack of the capability to examine in single-cell resolution. To overcome these issues, 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. SSA has demonstrated superior performance compared to existing methods SpaOTsc, Tangram, Seurat and DistMap using 10 semi-simulated datasets generated from a high-resolution spatial transcriptomics human breast cancer dataset with 100,064 cells. This advancement provides a refined tool for researchers to delve deeper in understanding of the relationship between cellular spatial organization and gene expression.

1 Introduction

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 [3–6], and treatment response [7]. 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. For example, high-resolution smFISH-based techniques such as seqFISH+ and ExM-MERFISH can only cover a small area of tissue, comprised of around 10,000 cells with a few dozens to a few hundreds of genes [8]. Platforms like 10x Visium [9] are restricted to collective gene expression evaluations, in which each spatially pinpointed expression profile is the average expression of many cells.

H. Al-Mubaid, T. Aldwairi and O. Eulenstein (eds.), BICOB-2024 (EPiC Series in Computing, vol. 101), pp. 25–38

One practical solution to the abovementioned challenge is to perform another single-cell experiment and then map individual cells to spatial locations of the spatial data. Computational methods to perform this task are commonly referred as spatial reconstruction of scRNA-seq data. In this research, we focus only on methods which can output a mapping scheme between individual cells and spatial locations. These methods can be classified into five main categories: i) ad hoc scoring, ii) generative modeling, iii) shared latent space and iv) other techniques.

Methods in the first category generally measure similarity score between each cell and each location based on gene expression and use the obtained score matrix as a mapping scheme. For example, DistMap [10] binarizes both scRNA-seq and ST data before computing the Matthew correlation coefficient (MCC) [11] matrix between cells and spots. Despite some initial success, ad hoc scoring methods often requires binarization step which fails to capture quantitative nuances of gene expression patterns.

Methods in the second category, including Seurat1 [12] and sstGPLVM [13], tend to use probable statistical models to transform the data before measuring relationship between single-cell data and spatial data. Seurat1 uses different techniques to infer expression of landmark genes in scRNA-seq data and ST data, then it compares gene expression between cells and locations to infer cells' spatial locations. sstGPLVM is a generative model method that perform cell-to-spot alignment based on shared latent space. The method uses black box variational inference to estimate the joint Gaussian-based latent space between scRNA-seq and ST data, then imputes missing data in gene expression and covariates and maps individual cells to spatial locations. Generative modeling methods are usually biased because of strong assumptions made about data distribution.

Methods in the third category are based on shared latent space between datasets without using generative modeling to build the mapping matrix, as Seurat3 [14] and Harmony [15]. Seurat3 performs canonical correlation analysis (CCA) [16] to project scRNA-seq and ST datasets into a shared low-dimensional latent space, then selects pairs of mutual nearest neighbors (MNNs) [17] and calculates their corresponding weight score to construct the mapping scheme. Harmony implements an altered K-means algorithm [18] to cluster data projected by Principal Component Analysis (PCA) [19] before using mixture models [20] to align scRNA-seq and ST data in a common latent space. Overal, both generative modeling and shared latent space methods require high quality of data to work on and they might not be robust to outliers.

Methods in the last category which are Tangram [21], SpaOTsc [22], Novosparc [23] follow a variety of strategies. Tangram infers the mapping matrix of probability between cells and locations by minimizing KL divergence between mapped and actual cell density in each location. SpaOTsc constructs an optimal transport plan [24] from single cells to locations using two gene expression dissimilarity matrices among cells, between cells and locations and a spatial distance matrix of locations. Novosparc constructs neighborhood graphs for single cells and spatial locations using gene expression distances and physical distances respectively. Then, considering the spatial autocorrelation among cells, the method computes an optimal transport to match between the two constructed graphs. Overall, these methods make extra assumptions about expression correlation across cells or locations, which requires excessive computational power to maintain. Therefore, methods in this category are usually slow on big datasets.

Here, we propose a new approach, SSA, that can accurately map scRNA-seq data to Spatial transcriptomics spots. We formulate the cell-to-spot mapping as an optimal transport problem and find the optimal mapping using the Sinkhorn algorithm [25]. We validate the proposed method using semi-simulated datasets generated from a real high-resolution spatial transcriptomics dataset with 100,064 cells. We compare the performance of our method with 04 state-of-the-art techniques. Our results show that the proposed approach outperforms all

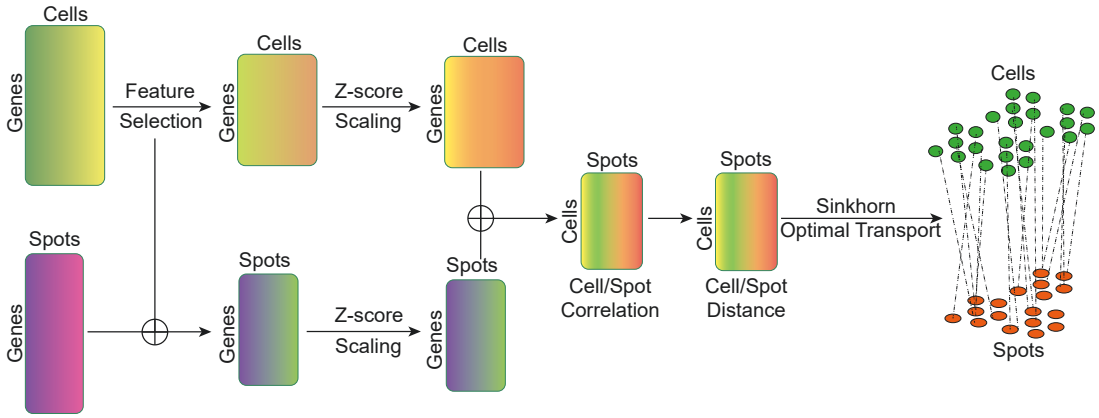

Figure 1: The overall pipeline of SSA. The input are scRNA-seq and ST gene expression matrices. In scRNA-seq matrix, rows are genes and columns are cells/samples. For spot expression matrix, rows are genes and columns are spots. The initial stage is gene selection on the scRNA-seq data to get 5,000 genes with the highest variance, which are subsequently used to subset two expression matrices. Following this, the Z-score scaling technique is used to normalize and transform both filtered matrices. The next step involves calculating the Pearson's correlation between gene expression of cells and spots and transforming it into an expression distance matrix. Finally, SSA leverages Sinkhorn algorithm to accurately map the cells onto corresponding spots.

competing methods by having the highest accuracy.

2 Methods

Figure 1 shows the overall analysis pipeline of SSA. The input of SSA includes the expression matrices of scRNA-seq (genes by cells) and ST (genes by spots). SSA first performs gene selection on scRNA-seq data to keep genes with the highest variance and then applies the Z-score scaling technique to transform both filtered scRNA-seq and spot expression matrices to z-scores. The method then calculates Pearson's correlation between gene expression of cells and spots and transforms it into an expression distance matrix. Finally, SSA applies the Sinkhorn algorithm to accurately map single-cells to spatial spots.

2.1 Feature Selection and Data Transformation

To efficiently perform cells-to-spots alignment, we need to remove genes that are uninformative from both scRNA-seq and spatial transcriptomics data. We first compute the variance for each gene in the scRNA-seq data and select 5,000 genes with the highest variance. We presume that these selected genes also play a pivotal role in spot expression data. Then, we subset both of the expression matrices using this gene set for consistency and comparative analysis. As a result, we obtain two new scRNA-seq and spot expression matrices that have same genes.

For each of the obtained matrices, we use Z-score transformation to scale and center the data. For each gene g, we determine the parameters μ and σ of the Gaussian distribution using all expression values. Then, for each expression value x_{ij} in the gene g we calculate the

Z-score using $Z_{ij} = (x_{ij} - \mu_j)/\sigma_j$. This step helps to reduce the impact of outliers, noise and technical variability in the gene expression data, which can improve the robustness of cells-to-spots alignment algorithm. The outputs from this step are two matrices with expression values are in Z-score scale.

2.2 Cell to Spot Alignment using Sinkhorn Algorithm

In this section, we will describe the main steps to perform cells-to-spots alignment using Sinkhorn algorithm. Given two X and Y as the scaled scRNA-seq and ST matrices obtained from the previous section. First, we calculate the pair wise Pearson's correlation between cells and spots, $\rho(X,Y) = \frac{Cov(X,Y)}{\sigma(X)*\sigma(Y)}$ where $Cov(X,Y)$ is the covariance between X and Y. Using obtained cells/spots correlation matrix, we calculate the pair wise distance between cells and spots $D(X,Y) = 1 - |\rho(X,Y)|$.

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 (in this case, expression profiles) from the cells to the spots, where the "cost" of transporting mass is given by the distance matrix. In details, Sinkhorn algorithm works as follows:

1. Initialization:

- A distance matrix $D(X,Y)$ (cost matrix) that represents distances between single cells and spatial spots.
- Two probability vectors p and q that represent the expression distribution of single cells and spatial spots respectively. Here, p are row-wise sum of gene expression values for each cell in scRNA-seq data and q are row-wise sum of gene expression values for each spot in ST data.
- A regularization term $\lambda = 0.05$.
- A kernel matrix $K = e^{\frac{-\lambda * D(X,Y)}{max(D(X,Y)}}$
- Two vectors $a \in R^m$ and $b \in R^n$ with all entries equal to 1

2. Update:

Repeat until convergence:

- Update: $b = \frac{q}{K^T a}$
- Update: $a = \frac{p}{Kb}$

3. Compute transport plan:

$T = diag(a)Kdiag(b)$

The output of the Sinkhorn algorithm is a matrix $T^{m \times n}$ where each value represents the mass (expression value) 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.

3 Results

In this section, we assess the performance of SSA in the following capabilities: (1) correctly mapping cells to spatial locations, (2) correctly assigning cells to spatial locations within specific cell types. We assess SSA performance against four state-of-the-art methods, SpaOTsc [22], Tangram [21], Seurat [14] and DistMap [10] using 10 semi-simulated scRNA-seq datasets.

3.1 Data Preparation

This section presents the process of preparing data for our analyses. we download a high-resolution spatial transcriptomics human breast cancers data from Gene Expression Omnibus (GEO) under accession number GSE176078 [26]. This dataset contains 100,064 cells in which cells expression profile, cells coordinates in real tissue and cells types are available. We transform the high-resolution ST data into 01 low-resolution ST dataset and 10 scRNA-seq datasets.

First, we partition the spatial domain into a grid structure where each grid cell, or "spot", is larger than the spatial extent of individual cells. The location of each spot is determined by averaging the 2-D coordinates of all cells falling within its boundaries, effectively representing the centroid of cellular locations within each spot.

Second, the gene expression profile for each spot is generated by averaging the gene expression levels of all cells residing within the corresponding grid cell. In this manner, the heterogeneity of gene expression at the cellular level is retained while transitioning to a coarser spatial resolution. Each spot now serves as a representative transcriptomic snapshot of the multiple individual cells within its confines. As a result, we was able to obtain an ST dataset of 3,615 spots and the mapping of spatial location between cells and spots. The spatial transcriptomic landscape of spots is shown in Figure 2. Here, we only use ST expression profile to perform cells-to-spots alignment. We later use spatial mapping information *a posteriori* to assess the performance of all methods.

Finally, we sub-sample the high-resolution ST data to get 10 equally smaller datasets and use them as scRNA-seq datasets for analysis. For each dataset, we perform *log* transformation to re-scale the data, i.e., $log(\mathbf{A}+1)$ where $\mathbf{A}$ is the expression matrix. We then use ST expression profile and single-cell expression profile to perform cells-to-spots alignment. We keep spatial mapping information as *a posteriori* to assess the performance of all methods.

3.2 SSA accurately maps cells to spot reference spatial locations

In ten datasets obtained from the previous section, the assignment of cells to respected spots locations are known. We only use this information for validation of SSA and other methods. For each dataset, we have scRNA-seq and ST expression matrices that serve as the inputs of each cells-to-spots alignment method. After running each method, we have a table in which the first column contain cells IDs and the second column has predicted spots IDs mapped to cells in the first columns. In order to assess if the cells are accurately assigned to the true spots, we measure the average Euclidean and Manhattan distances calculated from predicted spots and true spots spatial location. Smaller distances indicate the better cells-to-spots alignment results.

Table 1 shows average Euclidean distance values obtained for each method from ten datasets. For each row, cells highlighted in bold have the lowest values. For each of the ten datasets analyzed, the average Euclidean distance values obtained for SSA are substantially lower than those of SpaoTsc, Tangram, Seurat and DistMap ($p = 6.531 \times 10^{-7}$ using Wilcoxon test) demonstrating the SSA can accurately align cells back to the true spatial location in the tissue.

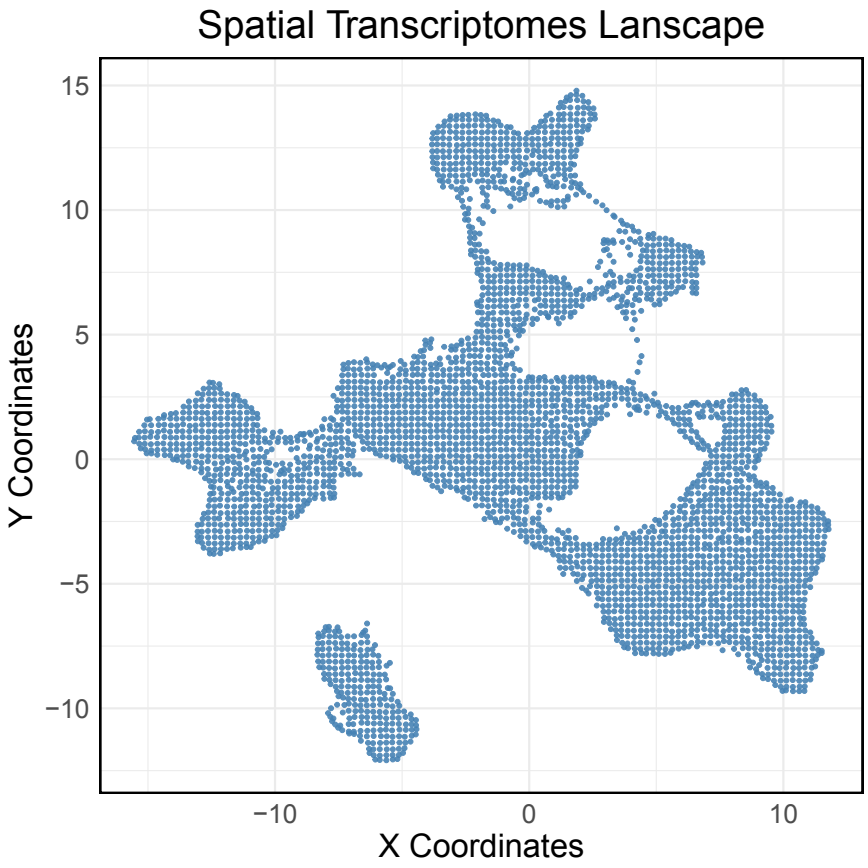

Figure 2: The spatial locations of simulated spots. The scatter plots show the x and y coordinates of each spot. Each spot's location is determined by averaging the 2-D coordinates of all cells within its grid boundaries.

The average Euclidean distance values of SSA is 1.786 which is substantially lower than those of other methods (2.976, 5.355, 11.957 and 144.133 for SpaoTsc, Tangram, Seurat and DistMap, respectively). Especially, DistMap fails to align individual cells to spots in dataset 9 making the average Euclidean distance is very high (1398.808).

Tables 2 shows average Manhattan distance obtained from SSA, SpaoTsc, Tangram, Seurat and DistMap obtained from 10 datasets. Again, this metric confirms that SSA is the best among the competing methods. The Manhattan distance values of SSA are significantly lower than those of other methods ($p = 6.534 \times 10^{-7}$ using Wilcoxon test). The average Manhattan distance values of SSA across all datasets is 2.257 while the average values of SpaoTsc, Tangram, Seurat and DistMap are 3.784, 6.865, 15.152 and 175.116, respectively. Similar to result presented in Table 1, DistMap unable to map cells in dataset 9 to a dedicated spots resulting in excessively high distance (1697.566).

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

3.3 SSA accurately maps cells to spatial location within cell types

In this section, we conduct another quantitatively evaluation of the SSA performance. We compare the cells-to-spots alignment results of SSA and other methods using cell type level Kullback-Leibler (KL) divergence (see Appendix A for more details). KL-divergence score is a measure of how the probability distribution of predicted cells locations is different from the ground truth. Given the true and predicted cells locations generated from SSA and other methods, we calculate the KL-divergence score for all cells in the same cell type. The smaller KL-divergence means the two distributions are very similar.

Figure 3 shows the box plots of KL-divergence scores calculated for each method across nine cell types. Overall, KL-divergence scores calculated from SSA results are substantially lower than those of SpaoTsc, Tangram, Seurat and DistMap across all cell types ($p = 2.838 \times 10^{-7}$ using Wilcoxon test). SSA KL-divergence scores are also significantly lower than the ones produced from other methods in *CAFs, PVL, B cells, T cells, Normal Epithelial* cell type with

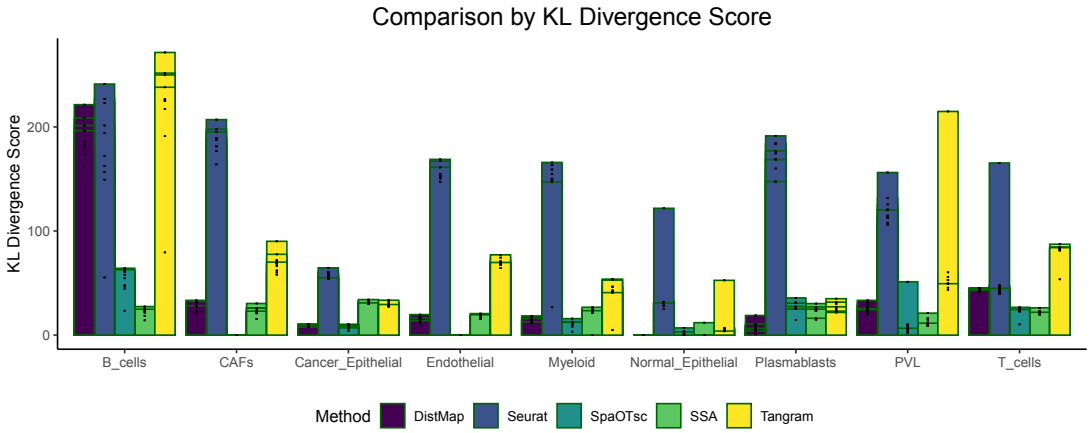

Figure 3: The KL-divergence results generated from SSA and other methods respected to cell types.

p-values of 3.4×10^{-2}, 2.6×10^{-3}, 3.7×10^{-9}, 3.8×10^{-7} and 1.7×10^{-2}, respectively. Those results show that SSA is not only capable of accurately aligning individual cells to spots on entire dataset but also it can generate predicted spatial distribution that is closer to the true distribution for a specific cell type.

4 Conclusion

In this article, we have introduced a novel computational tool, SSA, for alignment of individual cells to physical spatial locations in a tissue based on scRNA-seq and ST data. SSA directly projects individual cells to their spatial coordinates in tissue sections using Sinkhorn algorithm and therefore takes full advantage of the scRNA-seq data. The contribution of SSA approach is two folds. First, SSA accurately maps single-cells to spatial locations for the whole cells population with very minimal differences in distance. Second, SSA efficiently reconstructs a cellular spatial map within a specific cell type. Our extensive analysis shows that SSA substantially outperforms existing state-of-the-art approaches in different benchmark scenarios. SSA can be seamlessly incorporated into other analysis pipelines.

For future work, we plan to integrate this alignment technique with other techniques developed in our research lab for single-cell analysis to group cells with similar patterns together before performing cells-to-spots alignment [27–31]. This will improve the process of finding cells that have highly correlated expression profile with spots. We also plan to extend this work to improve the data for data integration and pathway analysis [32–44], cancer research [45–52], and space biology and drug development [53–58].

5 Acknowledgments

This work was partially supported by 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. Any opinions, findings, and conclusions or recommendations expressed in

this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

A Evaluation metrics

KL-divergence measures how one probability distribution is different from a second, reference probability distribution. Given two sets of true and predicted ST spatial coordinates. Each point should represent the position of a cell. Here, we use two-dimensional kernel density estimation (KDE) to estimate probability density function of cells locations. Given a set of cells with coordinates $\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, the formula for KDE is as follows:

$$f(x, y) = \frac{1}{N} \sum_{j=1} K_h((x - x_j), (y - y_j)) \tag{1}$$

where N is the total number of points, K_h is the kernel function with bandwidth h (in this case, a 2D Gaussian kernel). We apply KDE on both true and predicted ST spatial coordinates of cells to obtain two distributions $P(i)$ and $Q(i)$. Here, the KL divergence measures the distance from the approximate distribution Q to the true distribution P. KL-divergence is defined mathematically as below:

$$D_{KL}(P||Q) = \Sigma_{i \in I} P(x) log \frac{P(i)}{Q(i)} \tag{2}$$

Where the "||" operator indicates "divergence" or Ps divergence from Q.

References

[1] A Xiaowei. Method of the Year 2020: Spatially resolved transcriptomics. *Nature Methods*, 18(1), 2021.

[2] Leeat Keren, Marc Bosse, Diana Marquez, Roshan Angoshtari, Samir Jain, Sushama Varma, Soo-Ryum Yang, Allison Kurian, David Van Valen, Robert West, and Sean C. Bendall. 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. Schürch, Salil S. Bhate, Graham L. Barlow, Darci J. Phillips, Luca Noti, Inti Zlobec, Pauline Chu, Sarah Black, Janos Demeter, David R. McIlwain, Shigemi Kinoshita, Nikolay Samusik, Yury Goltsev, and Garry P. Nolan. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell*, 182(5):1341–1359, 2020.

[4] Hartland W. Jackson, Jana R Fischer, Vito RT. Zanotelli, H Raza Ali, Robert Mechera, Savas D. Soysal, Holger Moch, Simone Muenst, Zsuzsanna Varga, Walter P. Weber, and Bernd Bodenmiller. The single-cell pathology landscape of breast cancer. *Nature*, 578(7796):615–620, 2020.

[5] Bogdan A Luca, Chloé B Steen, Magdalena Matusiak, Armon Azizi, Sushama Varma, Chunfang Zhu, Joanna Przybyl, Almudena Espín-Pérez, Maximilian Diehn, Ash A Alizadeh, Matt van de Rijn, Andrew J. Gentles, and Aaron M. Newman. Atlas of clinically distinct cell states and ecosystems across human solid tumors. *Cell*, 184(21):5482–5496, 2021.

[6] Barbara T Grunwald, Antoine Devisme, Geoffroy Andrieux, Foram Vyas, Kazeera Aliar, Curtis W McCloskey, Andrew Macklin, Gun Ho Jang, Robert Denroche, Joan Miguel Romero, Prashant Bavi, Peter Bronsert, Faiyaz Notta, Grainne O Kane, Julie Wilson, Jennifer Knox, Laura Tamblyn, Molly Udaskin, Nikolina Radulovich, Sandra E. Fischer, Melanie Boerries, Steven Gallinger, Thomas Kislinger, and Rama Khokha. Spatially confined sub-tumor microenvironments in pancreatic cancer. *Cell*, 184(22):5577–5592, 2021.

[7] Rodrigo Nalio Ramos, Yoann Missolo-Koussou, Yohan Gerber-Ferder, Christian P Bromley, Mattia Bugatti, Nicolas Gonzalo Núñez, Jimena Tosello Boari, Wilfrid Richer, Laurie Menger, Jordan Denizeau, Christine Sedlik, Pamela Caudana, Fiorella Kotsias, Leticia L. Niborski, Sophie Viel, Mylène Bohec, Sonia Lameiras, Sylvain Baulande, Laëtitia Lesage, André Nicolas, Didier Meseure, Anne Vincent-Salomon, Fabien Reyal, Charles-Antoine Dutertre, Florent Ginhoux, Lene Vimeux, Emmanuel Donnadieu, Bénédicte Buttard, Jérôme Galon, Santiago Zelenay, William Vermi, Pierre Guermonprez, Eliane Piaggio, and Julie Helft. Tissue-resident FOLR2+ macrophages associate with CD8+ T cell infiltration in human breast cancer. *Cell*, 185(7):1189–1207, 2022.

[8] Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. *Nature methods*, pages 1–13, 2022.

[9] Kaitlyn H. Hajdarovic, Doudou Yu, Lexi-Amber Hassell, Shane A. Evans, Sarah Packer, Nicola Neretti, and Ashley E. Webb. Single-cell analysis of the aging female mouse hypothalamus. *Nature Aging*, 2(7):662–678, 2022.

[10] Nikos Karaiskos, Philipp Wahle, Jonathan Alles, Anastasiya Boltengagen, Salah Ayoub, Claudia Kipar, Christine Kocks, Nikolaus Rajewsky, and Robert P. Zinzen. The Drosophila embryo at single-cell transcriptome resolution. *Science*, 358(6360):194–199, 2017.

[11] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

[12] Rahul Satija, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33:495–502, 2015.

[13] Archit Verma and Barbara Engelhardt. A bayesian nonparametric semi-supervised model for integration of multiple single-cell experiments. *bioRxiv*, pages 2020–01, 2020.

[14] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

[15] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, 16(12):1289–1296, 2019.

[16] Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution*, pages 162–190. Springer, 1992.

[17] Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, 2018.

[18] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 4, pages 281–297. California, USA, 1967.

[19] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933.

[20] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

[21] Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R. Vanderburg, Åsa Segerstolpe, Meng Zhang, Inbal Avraham-Davidi, Sanja Vickovic, Mor Nitzan, Sai Ma, Ayshwarya Subramanian, Michal Lipinski, Jason Buenrostro, Nik Bear Brown, Duccio Fanelli, Xiaowei Zhuang, Evan Z. Macosko, and Aviv Regev. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nature Methods*, 18(11):1352–1362, 2021.

[22] Zixuan Cang and Qing Nie. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature Communications*, 11:2084, 2020.

[23] Noa Moriel, Enes Senel, Nir Friedman, Nikolaus Rajewsky, Nikos Karaiskos, and Mor Nitzan. Novosparc: flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nature Protocols*, 16(9):4177–4200, 2021.

[24] Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE signal processing magazine*, 34(4):43–59, 2017.

[25] Philip A Knight. The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

[26] Sunny Z Wu, Ghamdan Al-Eryani, Daniel Lee Roden, Simon Junankar, Kate Harvey, Alma Andersson, Aatish Thennavan, Chenfei Wang, James R. Torpy, Nenad Bartonicek, Taopeng Wang, Ludvig Larsson, Dominik Kaczorowski, Neil I. Weisenfeld, Cedric R. Uytingco, Jennifer G. Chew, Zachary W. Bent, Chia-Ling Chan, Vikkitharan Gnanasambandapillai, Charles-Antoine Dutertre, Laurence Gluch, Mun N. Hui, Jane Beith, Andrew Parker, Elizabeth Robbins, Davendra Segara, Caroline Cooper, Cindy Mak, Belinda Chan, Sanjay Warrier, Florent Ginhoux, Ewan Millar, Joseph E. Powell, Stephen R. Williams, X. Shirley Liu, Sandra O'Toole, Elgene Lim, Joakim Lundeberg, Charles M. Perou, and Alexander Swarbrick. A single-cell and spatially resolved atlas of human breast cancers. *Nature Genetics*, 53(9):1334–1347, 2021.

[27] Bang Tran, Duc Tran, Hung Nguyen, Nam Sy Vo, and Tin Nguyen. RIA: a novel Regression-based Imputation Approach for single-cell RNA sequencing. In *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–9. IEEE, 2019.

[28] Duc Tran, Hung Nguyen, Bang Tran, Carlo La Vecchia, Hung N. Luu, and Tin Nguyen. Fast and precise single-cell data analysis using hierarchical autoencoder. *Nature Communications*, 12:1029, 2021.

[29] Duc Tran, Bang Tran, Hung Nguyen, and Tin Nguyen. A novel method for single-cell data imputation using subspace regression. *Scientific Reports*, 12:2697, 2022.

[30] Bang Tran, Duc Tran, Hung Nguyen, Seungil Ro, and Tin Nguyen. scCAN: single-cell clustering using autoencoder and network fusion. *Scientific Reports*, 12:10267, 2022.

[31] Yifan Zhang, Duc Tran, Tin Nguyen, Sergiu M Dascalu, and Frederick C. Harris. A robust and accurate single-cell data trajectory inference method using ensemble pseudotime. *BMC Bioinformatics*, 24(1):1–21, 2023.

[32] Tin Nguyen, Rebecca Tagett, Michele Donato, Cristina Mitrea, and Sorin Draghici. A novel bi-level meta-analysis approach-applied to biological pathway analysis. *Bioinformatics*, 32(3):409–416, 2016.

[33] Tin Nguyen, Diana Diaz, Rebecca Tagett, and Sorin Draghici. Overcoming the matched-sample bottleneck: an orthogonal approach to integrate omic data. *Scientific Reports*, 6:29251, 2016.

[34] Tin Nguyen, Cristina Mitrea, Rebecca Tagett, and Sorin Draghici. DANUBE: Data-driven meta-ANalysis using UnBiased Empirical distributions - applied to biological pathway analysis. *Proceedings of the IEEE*, 105(3):496–515, 2017.

[35] Tin Nguyen, Cristina Mitrea, and Sorin Draghici. Network-based approaches for pathway level analysis. *Current Protocols in Bioinformatics*, 61(1):8–25, 2018.

[36] Tuan-Minh Nguyen, Adib Shafi, Tin Nguyen, and Sorin Draghici. Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biology*, 20:203, 2019.

[37] Adib Shafi, Tin Nguyen, Azam Peyvandipour, Hung Nguyen, and Sorin Draghici. A multi-cohort and multi-omics meta-analysis framework to identify network-based gene signatures. *Frontiers in Genetics*, 10:159, 2019.

[38] Adib Shafi, Tin Nguyen, Azam Peyvandipour, and Sorin Draghici. GSMA: an approach to identify robust global and test Gene Signatures using Meta-Analysis. *Bioinformatics*, 36(2):487–495, 2019.

[39] Hung Nguyen, Sangam Shrestha, Duc Tran, Adib Shafi, Sorin Draghici, and Tin Nguyen. A comprehensive survey of tools and software for active subnetwork identification. *Frontiers in Genetics*, 10:155, 2019.

[40] Edward Cruz, Hung Nguyen, Tin Nguyen, and Ian Wallace. Functional analysis tools for post-translational modification: a post-translational modification database for analysis of proteins and metabolic pathways. *The Plant Journal*, 99(5):1003–1013, 2019.

[41] Hung Nguyen, Duc Tran, Bang Tran, Bahadir Pehlivan, and Tin Nguyen. A comprehensive survey of regulatory network inference methods using single-cell RNA sequencing data. *Briefings in Bioinformatics*, 22(3):1–15, 2021.

[42] Tin Nguyen, Adib Shafi, Tuan-Minh Nguyen, A. Grant Schissler, and Sorin Draghici. NBIA: a network-based integrative analysis framework–applied to pathway analysis. *Scientific Reports*, 10:4188, 2020.

[43] Hung Nguyen, Duc Tran, Jonathan M. Galazka, Sylvain V. Costes, Afshin Beheshti, Sorin Draghici, and Tin Nguyen. CPA: A web-based platform for consensus pathway analysis and interactive visualization. *Nucleic Acids Research*, 49(W1):W114–W124, 2021.

[44] Zeynab Maghsoudi, Ha Nguyen, Alireza Tavakkoli, and Tin Nguyen. A comprehensive survey of the approaches for pathway analysis using multi-omics data integration. *Briefings in Bioinformatics*, 23(6):bbac435, 2022.

[45] Tin Nguyen, Rebecca Tagett, Diana Diaz, and Sorin Draghici. A novel approach for data integration and disease subtyping. *Genome Research*, 27:2025–2039, 2017.

[46] Hung Nguyen, Sangam Shrestha, Sorin Draghici, and Tin Nguyen. PINSPlus: A tool for tumor subtype discovery in integrated genomic data. *Bioinformatics*, 35(16):2843–2846, 2019.

[47] Duc Tran, Hung Nguyen, Uyen Le, George Bebis, Hung N. Luu, and Tin Nguyen. A novel method for cancer subtyping and risk prediction using consensus factor analysis. *Frontiers in Oncology*, 10:1052, 2020.

[48] Quang-Huy Nguyen, Hung Nguyen, Tin Nguyen, and Duc-Hau Le. Multi-omics analysis detects novel prognostic subgroups of breast cancer. *Frontiers in Genetics*, 11:1265, 2020.

[49] Hung Nguyen, Duc Tran, Bang Tran, Monikrishna Roy, Adam Cassell, Sergiu Dascalu, Sorin Draghici, and Tin Nguyen. SMRT: Randomized data transformation for cancer subtyping and big data analysis. *Frontiers in Oncology*, 11:725133, 2021.

[50] Thi Hai Yen Nguyen, Tin Nguyen, Quang-Huy Nguyen, and Duc-Hau Le. Re-identification of patient subgroups in uveal melanoma. *Frontiers in Oncology*, 11:731548, 2021.

[51] Quang-Huy Nguyen, Tin Nguyen, and Duc-Hau Le. Identification and validation of a novel three hub long noncoding RNAs with m6A modification signature in low-grade gliomas. *Frontiers in Molecular Biosciences*, 9:801931, 2022.

[52] Quang-Huy Nguyen, Tin Nguyen, and Duc Hau Le. DrGA: cancer driver gene analysis in a simpler manner. *BMC Bioinformatics*, 23(1):86, 2022.

[53] Adib Shafi, Cristina Mitrea, Tin Nguyen, and Sorin Draghici. A survey of the approaches for identifying differential methylation using bisulfite sequencing data. *Briefings in Bioinformatics*, 19(5):737–753, 2018.

[54] Michael P. Menden, Dennis Wang, Mike J. Mason, Bence Szalai, Krishna C. Bulusu, Yuanfang Guan, Thomas Yu, Jaewoo Kang, Minji Jeon, Russ Wolfinger, Tin Nguyen, Mikhail Zaslavskiy, AstraZeneca-Sanger Drug Combination DREAM Consortium, In Sock Jang, Zara Ghazoui, Mehmet E. Ahsen, Robert Vogel, Elias C. Neto, Thea Norman, Eric K. Y. Tang, Mathew J. Garnett, Giovanni Y. Di Veroli, Christian Zwaan, Stephen Fawell, Gustavo Stolovitzky, Justin Guinney, Jonathan R. Dry, and Julio Saez-Rodriguez. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nature Communications*, 10:2674, 2019.

[55] John C Stansfield, Duc Tran, Tin Nguyen, and Mikhail G Dozmorov. R tutorial: Detection of differentially interacting chromatin regions from multiple Hi-C datasets. *Current Protocols in Bioinformatics*, 66(1):e76–e76, 2019.

[56] Benjamin T. Caswell, Caio C. de Carvalho, Hung Nguyen, Monikrishna Roy, Tin Nguyen, and David C. Cantu. Thioesterase enzyme families: Functions, structures, and mechanisms. *Protein Science*, 31(3):652–676, 2022.

[57] Evagelia C. Laiakis, Maisa Pinheiro, Tin Nguyen, Hung Nguyen, Afshin Beheshti, Sucharita M. Dutta, William K. Russell, Mark R. Emmett, and Richard Britten. Quantitative proteomic analytic approaches to identify metabolic changes in the medial prefrontal cortex of rats exposed to space radiation. *Frontiers in Physiology*, DOI: 10.3389/fphys.2022.971282, 2022.

[58] Egle Cekanaviciute, Duc Tran, Hung Nguyen, Alejandra Lopez Macha, Eloise Pariset, Sasha Langley, Giulia Babbi, Sherina Malkani, Sébastien Penninckx, Jonathan C. Schisler, Tin Nguyen, Gary H. Karpen, and Sylvain V. Costes. Mouse genomic associations with in vitro sensitivity to simulated space radiation. *Life Sciences in Space Research*, DOI: 10.1016/j.lssr.2022.07.006, 2022.