Genome Biology

METHOD

Open Access

scDOT: optimal transport for mapping senescent cells in spatial transcriptomics

Nam D. Nguyen[1], Lorena Rosas[2], Timur Khaliullin[2], Peiran Jiang[1], Euxhen Hasanaj[3], Jose A. Ovando-Ricardez[2], Marta Bueno[4], Irfan Rahman[5], Gloria S. Pryhuber[6], Dongmei Li[7], Qin Ma[8], Toren Finkel[9], Melanie Königshoff[4], Oliver Eickelberg[4], Mauricio Rojas[2], Ana L. Mora[2], Jose Lugo-Martinez[1*] and Ziv Bar-Joseph[1,3*]

*Correspondence:
jlugomar@andrew.cmu.edu;
zivbj@andrew.cmu.edu

[1] Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
[2] Dorothy M. Davis Heart and Lung Research Institute, Division of Pulmonary, Critical Care and Sleep Medicine, Department of Internal Medicine, the Ohio State University, Columbus, OH, USA
[3] Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
[4] Division of Pulmonary, Allergy and Critical Care Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
[5] Department of Environmental Medicine, University of Rochester Medical Center, Rochester, NY, USA
[6] Department of Pediatrics, University of Rochester Medical Center, Rochester, NY, USA
[7] Department of Clinical and Translational Research, University of Rochester Medical Center, Rochester, NY, USA
[8] Department of Biomedical Informatics, College of Medicine, Ohio State University, Columbus, OH, USA
[9] Aging Institute, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

Abstract

The low resolution of spatial transcriptomics data necessitates additional information for optimal use. We developed scDOT, which combines spatial transcriptomics and single cell RNA sequencing to improve the ability to reconstruct single cell resolved spatial maps and identify senescent cells. scDOT integrates optimal transport and expression deconvolution to learn non-linear couplings between cells and spots and to infer cell placements. Application of scDOT to lung spatial transcriptomics data improves on prior methods and allows the identification of the spatial organization of senescent cells, their neighboring cells and novel genes involved in cell-cell interactions that may be driving senescence.

Background

Recent advancements in genomics technologies have facilitated the profiling of gene expression at the single-cell level, unveiling valuable insights regarding the molecular heterogeneity of complex biological systems. While single-cell RNA sequencing (scRNA-seq) has significantly enhanced our comprehension of cell-type diversity, it lacks spatial information due to the dissociation of cells. Spatial transcriptomics (ST) techniques enable the preservation of spatial information within tissue samples but typically offer lower resolution or coverage compared to scRNA-seq data. Hence, the integration of scRNA-seq and ST data becomes imperative for acquiring a spatially informed single-cell resolution dataset [1]. This integration approach not only ensures a more comprehensive understanding of the molecular heterogeneity within complex biological systems but also retains the spatial context of gene expression.

Existing methods for integrating single-cell and spatial transcriptomics data primarily focus on cell-type deconvolution. These methods decompose gene expression in a spatial spot into linear combinations of fractions attributed to different cell types, utilizing the single-cell data solely as a reference [2–9]. While successful, these methods often struggle when it comes to cell types with only a few cells [10–12]. Moreover, in cases where these smaller cell types are very similar to cell types with larger number of cells, the assignment of deconvolution methods often completely ignore these smaller cell types as shown in Results.

Cellular senescence, a state of permanent growth arrest, is implicated in various age-related diseases. Understanding cellular senescence requires analyzing cell-cell communications at the individual cell level, as the process exhibits heterogeneity, where only a few cells within a given cell type enter a senescent state simultaneously. Additionally, paracrine senescence, in which a senescent cell can induce senescence in neighboring cells, is of significant importance. Effective communication between senescent cells and neighboring cells is crucial for the progression and maintenance of the senescent phenotype [13, 14]. Senescent cells actively engage in intercellular communication, primarily through the secretion of senescence-associated secretory phenotype (SASP) factors, influencing neighboring and distant cells [14, 15]. However, the mechanisms underlying these communications remain poorly understood. To address this gap, and to enable the study of cell-cell interactions for these small number of senescent cells within a cell type using spatial transcriptomics, we propose an innovative computational framework that integrates single-cell and spatial transcriptomics data. This approach allows us to infer cell-cell communications based on the proximity of cells, whether short- or long-range, shedding light on the intricacies of senescence-associated intercellular signaling. This method offers a superior alternative to organoids, where only cell types interact in an artificial environment.

Mapping individual cells to their spatial origins requires fine-grained mapping, which is prone to imprecise results due to the similarity within cell types and the non-linear relationship between gene expression levels in scRNA-seq and spatial transcriptomics [16]. Methods proposed for this task compute a similarity score in a shared latent space. This similarity score is then coupled with a statistical test to determine the significance of the assignment [16, 17]. Other techniques, e.g., canonical correlation analysis or non-negative matrix factorization, for constructing shared latent space have also been used [18–20]. In contrast, here, we utilize optimal transport [21, 22], a mathematical framework that allows for the comparison and matching of probability distributions. Specifically, we use optimal transport to learn the non-linear coupling between cells and spots by aligning the distributions of gene expression profiles across these two datasets. Our approach employs a probabilistic mapping, where the precision of the mapping is modulated by incorporating the coarse-grained mapping of cell types obtained from the deconvolution task. We solve these two complementary optimization tasks using a bilevel optimization approach [23], based on the differentiable deep declarative network [24] (Fig. 1).

Our approach incorporates two types of data, namely scRNA-seq and spatial transcriptomics, as inputs. It employs iterative computations to perform cell type deconvolution and cell-to-spot spatial mapping. As a result, it produces a coupling matrix between cells and spots that serves as an initial integration outcome. This coupling matrix is subsequently used to infer the cell-to-cell spatial neighborhood graph by aligning cells with spots possessing known spatial coordinates (see Fig. 1). Essentially, the spot coordinates play a crucial role in determining the physical closeness between cells.

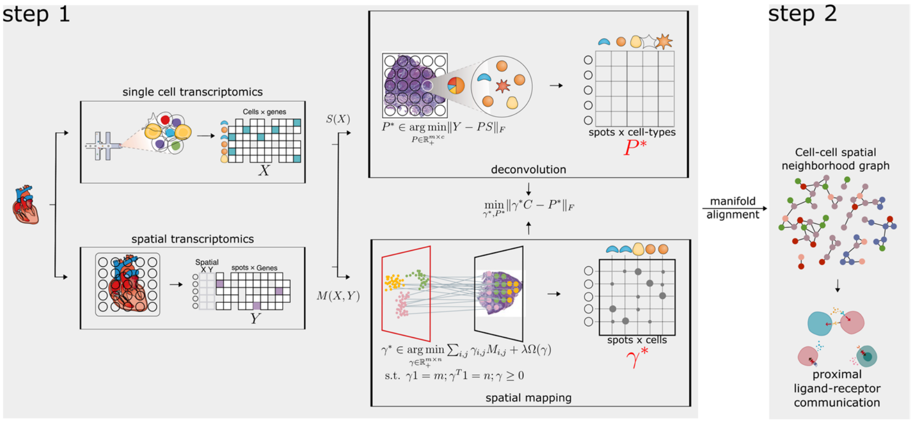
step 1
single cell transcriptomics
Cells × genes
X
spatial transcriptomics
Spatial
X Y
spots × Genes
Y
$S(X)$
$M(X,Y)$
$P^* \in \arg\min_{P\in\mathbb{R}_+^{m\times c}} \|Y - PS\|_F$
spots x cell-types
P^*
deconvolution
$\min_{\gamma^*,P^*} \|\gamma^* C - P^*\|_F$
$\gamma^* \in \arg\min_{\gamma\in\mathbb{R}_+^{m\times n}} \sum_{i,j} \gamma_{i,j} M_{i,j} + \lambda\Omega(\gamma)$
s.t. $\gamma 1 = m; \gamma^T 1 = n; \gamma \geq 0$
spots x cells
γ^*
spatial mapping
manifold alignment
step 2
Cell-cell spatial neighborhood graph
proximal ligand-receptor communication

Fig. 1 Method workflow: Step 1: scDOT takes gene expression profiles from a scRNA-seq dataset and a spatial transcriptomics dataset as inputs. Additionally, cell type information for cells in the scRNA-seq data and spatial coordinates for spots in the spatial transcriptomics data are provided. scDOT simultaneously and in parallel learns the cell type fraction of each spot (deconvolution task) and the mapping between individual cells in the scRNA-seq data and individual spots in the spatial transcriptomics data (spatial reconstruction task). Step 2: The resulting mapping matrix between cells and spots is then utilized to construct the cell-cell spatial neighborhood graph, where cells are connected if they are in close physical proximity

We tested scDOT on both, simulated and new spatial data. As we show, it can accurately assign cells to their spot of origin outperforming prior methods for this task. For the new samples for idiopathic pulmonary fibrosis (IPF), scDOT identifies the spatial distribution and cell-cell interactions between senescence and non-senescence cells and the set of genes involved in these interactions.

Results

We developed an optimal transport (OT) method for mapping scRNA-Seq data to spatial transcriptomics data. The method, illustrated in Fig. 1, performs iterative computations for cell type deconvolution and cell-to-spot spatial mapping, resulting in the generation of the coupling matrix γ as an upstream integration outcome. This coupling matrix is then utilized to infer the cell-to-cell spatial neighborhood graph by aligning cells to spots with known spatial coordinates.

scDOT efficiently reconstructs individual cells to their spatial origins

We first tested scDOT on two simulation datasets where ground truth is known (Methods). The outcome of reconstructing single-cell data, i.e., the coupling matrix γ, when using simulation dataset 1 reveals that it successfully recovers the spatial origins of a high fraction of cells (56% to 76%, depending on a predefined threshold to determine high probability). γ represents probabilistic couplings and so a specific cell can be mapped to several location with different probabilities (which sum up to 1). We found that in most cases the distribution $\gamma_{:,j}$ exhibits is extremely heavy-tailed and places a disproportionately high amount of probability densities at 0. We therefore defined a high probability of associating with a location based on distribution properties (99th, 95th, 90th quantile, or the 75th quantile (the third quantile) plus 1.5 times the interquartile range (IQR) (Turkey's fences)). Obviously, the stricter the threshold, the fewer cells that are correctly matched. However, even for a very high cutoff, we find very large percentage of correct matches (70% of cells at a threshold above the 90th quantile and 56% of cells at a threshold above the 99th quantile when using synthetic data 1). However, the slower decay of reconstruction results due to a more strict threshold is desirable and can be achieved through a heavier tail in the distribution $\gamma_{:,j}$.

In addition, previous studies show that cell type deconvolution methods tend to miss rare cell type [10]. In contrast, when using OT, we are able to map rare cell types to their spatial origins (Fig. 2b). In our simulation data, four types of cells can be classified as rare: 2-Mesothelium and Submucosal Secretory have only 1 cell each, Myofibroblasts has 2, and Fibromyocytes has 7. The boxplots indicate that our approach successfully assigned all these rare cell types to their correct spatial positions.

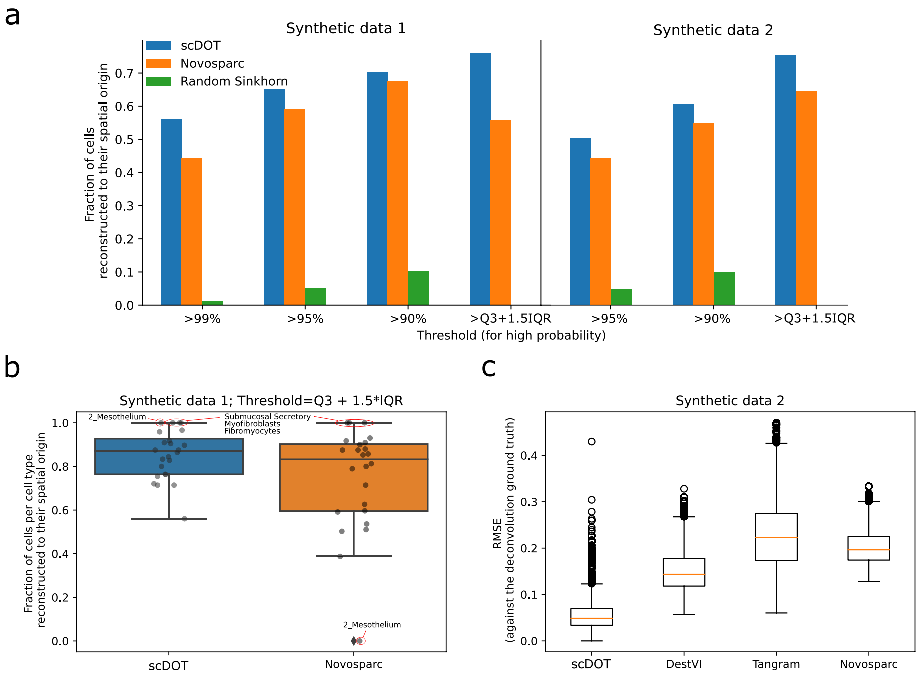

Fig. 2 Performance on synthetic datasets. **a** OT results of simulation datasets 1 and 2 demonstrate that by using different thresholds to define a high probability, we can assign nearly 80% of cells to their spatial origin. scDOT was benchmarked against two other methods: Novosparc, a spatial reconstruction method based on Gromov-Wasserstein distance, and Random Sinkhorn, a naive method that learns the optimal transport coupling with a random cost matrix. The results demonstrate the superior performance of scDOT in all cases. **b** Detailed results of simulation data 1 (with a threshold higher than the 3rd quantile plus 1.5 times the IQR) highlight the effectiveness of scDOT and spatial mapping methods in general for rare cell types. The boxplots illustrate the fraction of correctly reconstructed cells per cell type. Each point represents a single cell type ($c = 24$). Among the considered rare cell types (2-Mesothelium and Submucosal Secretory with 1 cell, Myofibroblasts with 2 cells, and Fibromyocytes with 7 cells), indicated within red circles, scDOT successfully mapped these rare cell types to their exact spatial locations (fraction = 1.0), while Novosparc failed to map 2-Mesothelium to its spatial location (fraction = 0.0). **c** The root-mean-square-error (RMSE) of the deconvolved cell-type proportions compared to the ground truth is evaluated for synthetic data 2, consisting of 9 cell types across 3072 spots. scDOT, along with other methods including DestVI, Tangram, and Novosparc, is compared in terms of RMSE. The boxplots demonstrate that scDOT outperforms the other methods, as indicated by the lower RMSE values. The boxplots display the median (middle line), 25th and 75th percentiles (box), and 5th and 95th percentiles (whiskers)

Comparison to other methods on spatial mapping and cell type deconvolution

Spatial mapping We evaluate the performance of scDOT in spatial mapping and compare it with other existing methods. Figure 2a presents the results for synthetic data 1, where the threshold is set above Q3 + 1.5×IQR. scDOT achieves the highest outcome at this threshold, while the outcome of Novosparc is drastically decreased compared to the outcome at thresholds above the 90th and 95th quantiles. This observation suggests that our probabilistic mapping exhibits a heavier-tailed characteristic, which is a more desirable property for accurate spatial mapping.

Moreover, we find that the reconstruction results are influenced by the dataset used. For synthetic data 2, scDOT achieves a high outcome when the threshold is set above Q3 + 1.5×IQR, with 76% of cells successfully reconstructed. However, stricter thresholds lead to a more rapid decay in the outcomes, with only 50% of cells being reconstructed at the threshold above the 95th quantile. Nevertheless, across all cases, scDOT consistently outperforms both Novosparc and the naive baseline of Random Sinkhorn. We also tested scDOT using another simulated dataset based on a new Xenium single-cell data from a breast cancer tumor microenvironment [25]. Again, we observed that scDOT performs well and is superior to Novosparc, achieving very high accuracy (Additional file 1: B.2, Fig. S1).

In terms of accurately mapping rare cell types to their spatial positions, scDOT successfully assigns all four rare cell types with a fraction of 1.0. However, Novosparc failed to accurately map 2-Mesothelium to its spatial location, as indicated by a fraction of 0.0. Also, as indicated in Fig. 2a, scDOT mapped 76% cells correctly while Novosparc mapped 56% cells correctly; these 20% differences is not shown in Fig. 2b since the difference in the number of cells per cell type is not considered. Finally, we conducted an analysis of scDOT's performance on rare cell types by varying the fraction of B cells in the dataset. Results, presented in the Additional file 1: B.3, demonstrate the robustness of scDOT for mapping rare cell types.

Deconvolution To benchmark the results of cell type deconvolution, we applied scDOT to synthetic data 2 and compared it with three other methods: DestVI [7], Tangram [26], and Novosparc [27]. The synthetic dataset comprised nine cell types distributed across 3072 spots. We specifically chose these three deconvolution methods as they represent distinct computational techniques tailored for spatial transcriptomics data. DestVI is a probabilistic-based method, Tangram utilizes deep learning, and Novosparc is an OT-based method. All three methods require spatial transcriptomics data as input and scRNA-seq data as a reference. Comparing the root-mean-square-error (RMSE) of the deconvolved cell type proportions with the ground truth, scDOT outperformed the other three methods (see Fig. 2c). The mean RMSE scores for scDOT, DestVI, Tangram, and Novosparc were 0.06, 0.15, 0.23, and 0.20, respectively. It is worth noting that Novosparc is not designed for direct computation of cell type deconvolution but rather for mapping cells to spots. As a result, the deconvolution results are calculated by multiplying the coupling matrix γ with the cell-by-cell type relation matrix C, i.e., $P = \gamma \times C$.

We also used Pearson correlation coefficient (PCC) analysis for this benchmark study. Result still demonstrate the superiority of scDOT over other methods, albeit with high variance (see Fig. S2).

Identfiying the spatial patterns of the distribution of specific cell types

We used paired IPF scRNA-Seq and spatial dataset to test the ability of our mapping method to infer cell-cell interactions (Fig. 3). Among the 29 cell types (Methods), multiciliated, secretory, and basal cells exhibited prominent and distinct spatial patterns. Notably, multiciliated, secretory, and basal cells were found to be in close proximity to each other, both in the upper lobe and lower lobe of the tissue. This observation aligns with the traditional view of the airway epithelial mucosal layer, which incorporates basal cells in close proximity to secretory and ciliated cells, forming a tight unit. This unit serves as a physical barrier while remaining responsive to the inhaled environment through interactions with submucosal fibroblasts, smooth muscle cells and cells and molecules from the immune system [28].

Secretory and multiciliated cells are known to be located in close proximity to each other within the respiratory tract, including the lungs. Together, they form a self-clearing mechanism that efficiently removes inhaled particles from the upper airways, preventing their transfer to deeper lung zones [29]. The coordinated action of multiciliated cells, with their motile cilia, and secretory cells, responsible for mucus production and secretion, enables the effective clearance of inhaled particles and maintains the integrity of the respiratory system [30].

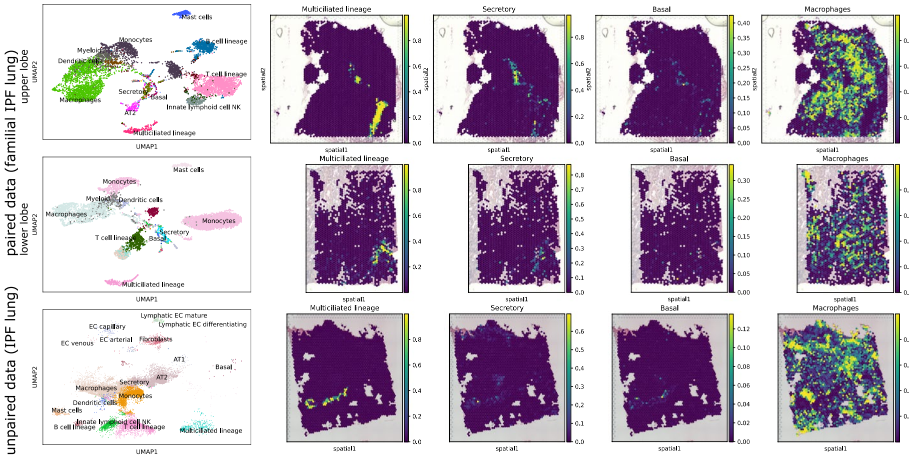

Fig. 3 Spatial distribution patterns of multiciliated, secretory, basal, and macrophage cells across different datasets. *Top*: A UMAP representation of scRNA-seq data, along with the spatial patterns of the selected cell types in the upper lobe slice of the paired familial IPF lung. *Middle*: A UMAP representation of scRNA-seq data and the corresponding cell types in the lower lobe slice of the same sample. *Bottom*: A UMAP representation and spatial distribution of selected cell types in the unpaired IPF lung sample. Notably, multiciliated, secretory, and basal cell types exhibit distinct and prominent spatial patterns. Importantly, these cell types consistently exhibit close proximity to each other across all three datasets, consistent with previous studies on the organization of the respiratory system [28–30]

Basal cells, positioned closer to the basement membrane, further contribute to the organization and functioning of the airway epithelium. They provide structural support and are responsible for the regeneration and repair of the airway epithelial layer [28].

The spatial organization of multiciliated, secretory, and basal cells in close proximity to each other emphasizes their interdependence and coordinated functioning in maintaining the respiratory barrier and facilitating efficient clearance mechanisms. This finding underscores the significance of the spatial arrangement and interactions of diverse cell types within the airway epithelium for the overall homeostasis and defense of the respiratory system.

Conversely, immune cell types such as Macrophages and T cells lineage, which were characterized by a larger number of cells, displayed a more scattered distribution throughout the tissue. Yet, the spatial distribution of these two cell types are complementary to some degree (Figs. 3 and 4, Fig. S3), reflecting the fact that they are both important components of the immune system and play complementary roles in

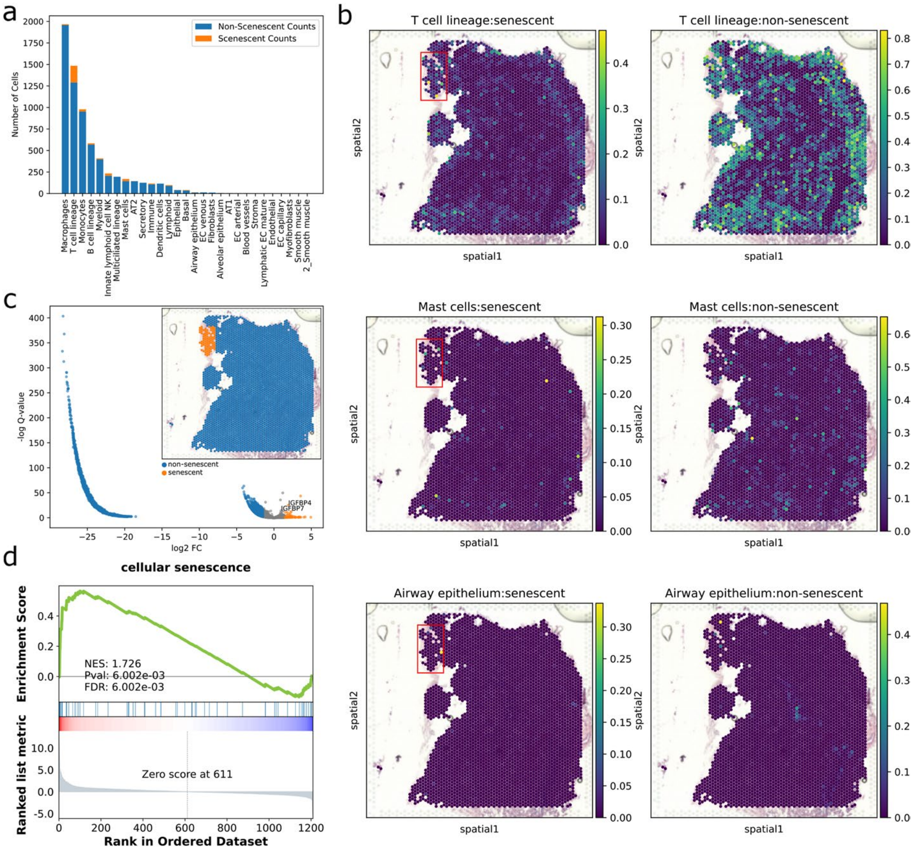

Fig. 4 Analysis of cellular senescence reveals the spatial collocation of senescent cells. **a** The number of senescent cells and non-senescent cells for each cell type is depicted. T cell lineage, mast cells, and airway epithelium exhibit the highest fraction of senescent cells. **b** Spatial distribution of senescent and non-senescent cells for the three aforementioned cell types. Notably, the three different senescent cell types are spatially collocated in the upper left corner of the tissue. **c** Differentially expressed genes for the manually annotated senescent region (colored in orange) in the upper left corner of the tissue (as depicted in **b** of this figure and the upper right corner of this panel). Among the top-ranked DEGs are IGFBP4 and IGFBP7, which are also senescent marker genes. **d** Gene set enrichment analysis (GSEA) plot. The top-ranked DEGs (as shown in panel (**c**) of this figure) are enriched in the gene set consisting of 340 senescent marker genes

defending against infections and maintaining immune homeostasis. On the other hand, cell types with a smaller cell count, such as smooth muscle (consisting of only 2 cells in total), exhibited a spatial arrangement in adjacent spots (Fig. S3).

These patterns were also observed in the unpaired data, particularly with regards to the multiciliated lineage and secretory cell types (Fig. 3), demonstrating the generality of our approach on unpaired datasets. The difference between paired and unpaired data is quantified in detail in Additional file 1 (B.4).

Cell-cell proximity analysis

To quantitatively illustrate the spatial distribution and proximity of multiciliated, secretory, and basal cells described in the previous section and shown in Fig. 3, we employed the neighborhood enrichment score. This score between two cell types represents the z-score derived from a permutation test that tallies the neighboring spots consisting of either cell type. Consistent with the spatial patterns depicted in Sect. 3.3 and Fig. 3, we observed the highest enrichment score between the multiciliated lineage and itself across various datasets (69.46 in the upper lobe of familial IPF paired data, 29.31 in the lower lobe of the same data, and 47.98 in the IPF unpaired data). The score between multiciliated and secretory cell types is also one of the highest (19.40 in the upper lobe of the paired dataset, 12.25 in the lower lobe, and 5.06 in the unpaired dataset). In contrast, the scores between macrophages and T cells are among the lowest across datasets, with scores of − 25, − 5.75, and − 15.83 in the upper lobe, lower lobe, and unpaired dataset, respectively. These scores reflect the fact that they are complementary, as indicated in Sect. 3.3 (see Fig. S3). It is important to note that the neighborhood enrichment scores were estimated at the spot-level and only considered the dominant cell type of each spot, which is defined as the cell type with the highest proportion within that particular spot.

At the cell level, we constructed a cell-cell spatial proximity graph based on OT placement (see the Methods section). The graph was then summarized by cell types, quantifying the physical proximity between each cell type by counting the direct neighboring cells within the same type (see Fig. S3d and e). Once again, the multiciliated lineage exhibited the highest normalized counts with itself across datasets, consistent with the results obtained from the enrichment score and described in Sect. 3.3. In the paired dataset, basal and secretory cells also demonstrated a strong association with the airway epithelium, providing additional evidence for the spatial organization of the respiratory system as discussed in Sect. 3.3. In contrast, immune cells such as T cells and macrophages displayed connections to various cell types, reflecting their dispersed distribution throughout the tissue. Notably, in the IPF lung sample, fibroblast cells exhibited a distinct spatial pattern and were found to be in close proximity to 2-smooth muscle cells and myofibroblast cells, supporting previous research suggesting that α smooth muscle actin-expressing fibroblasts, referred to as myofibroblasts, serve as markers of progressive lung injury and play a central role in detrimental remodeling and disease progression [31, 32] (Fig. S3, Sect. 3.6).

Identification of senescent markers

For cellular senescence analysis, we profiled two new spatial datasets. The first included paired scRNA-Seq data from a familial IPF lung sample, and the other consists of unpaired data from an IPF lung sample (Methods).

Paired data of familial IPF lung sample We first identified in the scRNA-seq data, cell types with a large fraction of cells exhibiting senescent. For this, we used a list of 68 senescent marker genes (Methods). Within each cell type, we separated the cells into senescent and non-senescent cells (Fig. 4a, b). For this familial IPF lung sample, the ratio of senescent cells to non-senescent cells is low. For most cell types, we observed very few senescent cells. For others, we found more. For example, for mast cells, T cell lineage, and airway epithelium, we identified 14%, 13%, and 17%, respectively. We thus focused on these three cell types. for these we had 24, 193, and 3 senescent cells for mast cells, T cell lineage, and airway epithelium, respectively. Next, we manually annotated the regions where senescent cells from different cell types are collocated (Fig. 4b, c). For these regions, we computed differentially expressed genes (DEG) w.r.t. the rest of the tissue. As expected, given the way we selected these regions, we found among the top ranked DEG IGFBP4 and IGFBP7 (t-test p-values are 1.1e−11 and 7.2e−07 respectively), which are both senescent marker genes (Fig. 4d). We next performed gene set enrichment analysis (GSEA) with this ranked gene list and a gene set of 340 senescent markers (which is a superset of the 68 senescent marker genes set we used for re-annotation, Additional file 2), we confirmed that cellular senescence is enriched—with p-value $= 0.006002$; FDR $= 0.006002$, and the normalized enrichment score is 1.726—in the annotated region (Fig. 4d). The leading-edge subset of genes in this analysis comprised IGFBP4, IGFBP7, FGF7, THBS1, IGF1, IGFBP6, IL6, SERPINE2, PIM1, ALDH1A3, SERPINE1, COL1A2, ANGPTL4, CYP1B1, and PLAU. While IGFBP4 and IGFBP7 belong to the initial set of 68 senescent marker genes, the remaining genes are part of the larger set of 340 senescent marker genes. Of particular note, IGFBP4 and IGFBP7 are SASP factors that have been identified as key components needed for triggering senescence in young mesenchymal stem cells (MSC) [33]. The pro-senescent effects of IGFBP4 and IGFBP7 are reversed by single or simultaneous immunodepletion of either proteins from the conditioned medium (CM) from senescent cells [33]. According to a previous study, prolonged IGF1 treatment leads to the establishment of a premature senescence phenotype characterized by a unique senescence network signature [34]. Combined IGF1/TXNIP-induced premature senescence can be associated with a typical secretory inflammatory phenotype that is mediated by STAT3/IL-1A signaling [34].

Inferring cell-cell interactions driving senescence

We also looked at the cell type neighborhood of senescent cells. These are summarized in Fig. 5a. We observe that senescent cells are often close to non-senescent cells of the same type (e.g., senescent T cells to non-senescent T cells) which can explain why some cell types have a much higher percentage of senescent cells than others.

Utilizing the CellPhoneDB [35], we further identified the ligand-receptor (LR) pairs involved in the cell-cell interactions within the neighborhood of senescent cells (i.e.,

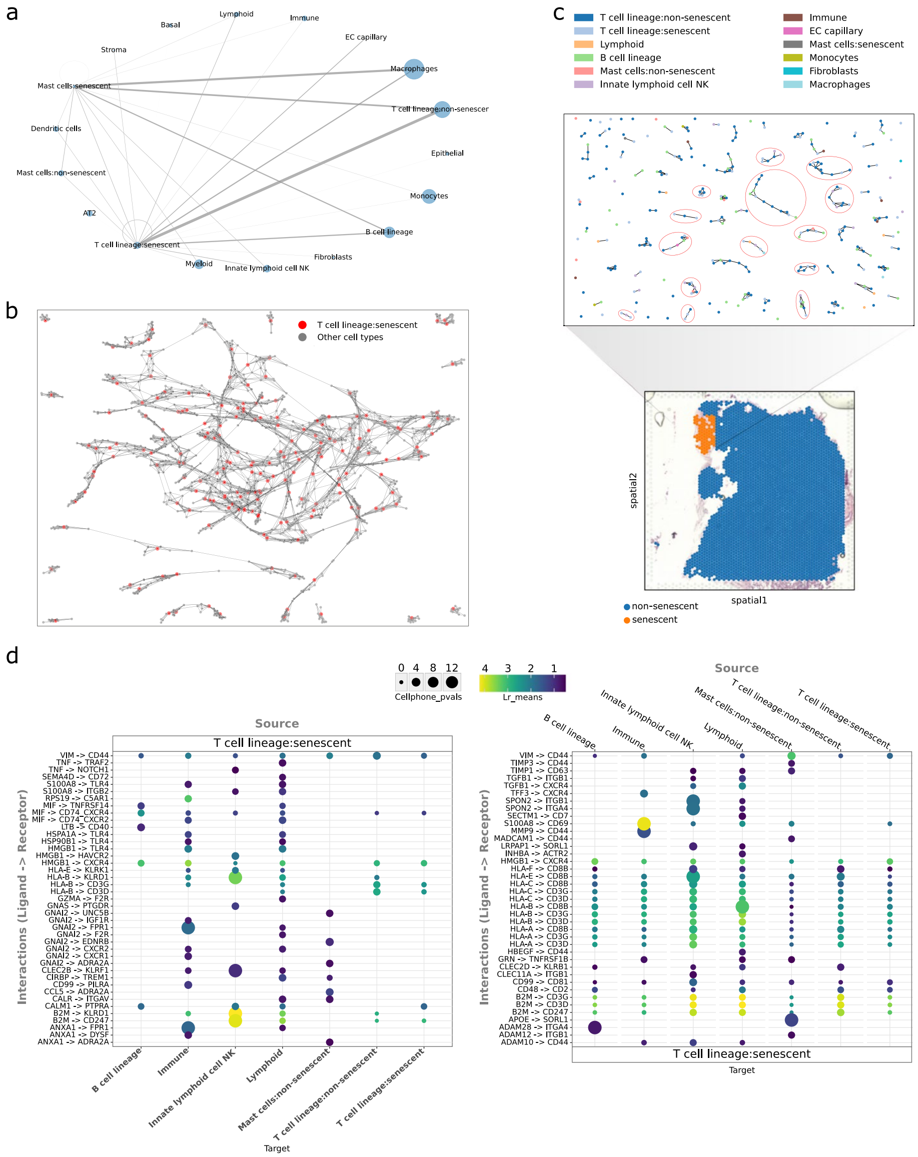

Fig. 5 Analysis of senescent cell-cell communication in the upper lobe of the familial IPF lung sample. **a** The graph summarizes the spatial neighborhood of senescent mast cells and T cell lineage. Nodes represent cell types, and edges indicate direct neighboring relations in physical proximity. The size of each node corresponds to the number of cells within a cell type, while the width of the edges represent the number of neighboring cells of a specific cell type (i.e., the total node degree per neighboring cell type). Edges representing a small number of neighbors are omitted. As can be seen, senescent cells are close to both non-senescent cells within the same cell types and senescent cells belonging to different cell types. **b** Cell-cell spatial neighborhood of senescent cells for the T cell lineage. The validity of this neighborhood graph is assessed in Additional file 1: B (Supplementary Analysis). **c** The subgraph of the cell-cell neighborhood depicted in panel (**b**), specifically showing the cells located in the senescent region (colored orange). **d** The results from CellphoneDB display the co-expressed ligand-receptor pairs between senescent cells of the T cell lineage and all other cells within the subgraph illustrated in panel (**c**)

within the graph G') (Fig. 5d). We observed that 11 senescent markers, namely B2M, CALR, CCL5, CD44, HMGB1, IGF1R, MIF, TNF, VIM, MMP9, and TNFRSF1B, were significantly overrepresented in the list of ligands and receptors identified by CellPhoneDB (hypergeometric test p-value $= 0.00072$). Among the LR pairs involved in senescent-to-senescent cell-cell communication (i.e., between senescent T cells), most of the pairs include senescent marker genes. The other remaining LR pairs involve the HLA gene family (which is essential for T cell activation). For example, HLA-E acts as an inhibitory signal for NK and CD8 T cells-and depletion of HLA-E renders senescent cells susceptible to elimination by both NK and CD8 T cells [36]. Another LR pair involves S100A8, which increases with age, inducing inflammation and cellular senescence-like phenotypes in oviduct epithelial cells [37, 38].

Unpaired data from IPF lung sample To demonstrate the general utility of the method for unpaired data, we performed the same analysis as described for the paired data mentioned above for another spatial dataset we profiled, this time without matched scRNA-Seq (Methods). Using a scRNA-seq dataset of an IPF lung sample, we were still able to identify several of the same senescence cell types as in the paired dataset, including T cells and mast cells. There were 300 assigned senescent cells out of the total 3747 T cells and 11 assigned senescent mast cells out of the total 249 mast cells. We also observed high fraction of senescence cells for other cell types including for fibroblasts (290 out of the total 461 fibroblast cells) and 2-smooth muscle (8 out of 21).

We again observed that senescence cells co-localized in the same regions (Fig. 6a). While T cells tended to be distributed throughout the tissue, there is a high fraction of senescent cells co-localized with fibroblasts and mast cells (Fig. 6a). Fibroblasts and 2-smooth muscle cells co-localized in specific regions, with a total of four overlapping regions as depicted in Fig. 6a. Since senescent cells tend to co-localize with other cells of the same type, most senescent fibroblast cells and 2-smooth muscle cells also co-localized (except for the region in the upper left corner of the tissue, which exhibited only senescent fibroblast cells). These observations of senescent spatial distribution align with previous studies suggesting that senescent cells have the potential to influence neighboring cells through processes collectively referred to as the senescence-associated secretory phenotype [39].

Figure 6b, c illustrate the physical proximity among cells of different cell types. Similar to the paired data of the familial IPF lung sample, senescent cells are closely clustered together and near cells of the same type. As shown in Fig. 6b, the senescent T cells are adjacent to other T cells, mast cells, and macrophages. The cell-to-cell spatial neighborhood graph, with nodes representing senescent T cells and their immediate neighbors, is depicted in Fig. 6c. The validity of this neighborhood graph is assessed in Additional file 1: B (Supplementary Analysis). For a more specific focus on senescent fibroblasts, a cell-to-cell neighborhood graph can be found in Fig. S4.

Discussion

In this study, we introduced a novel method for integrating single-cell and spatial transcriptomics, addressing the simultaneous tasks of cell type deconvolution and spatial reconstruction. The challenge of spatial reconstruction lies in the non-linear relationship between gene expression profiles of single-cells and the spatial transcriptomics data [16] as well as the inherent uncertainty in high-resolution mapping. However, by

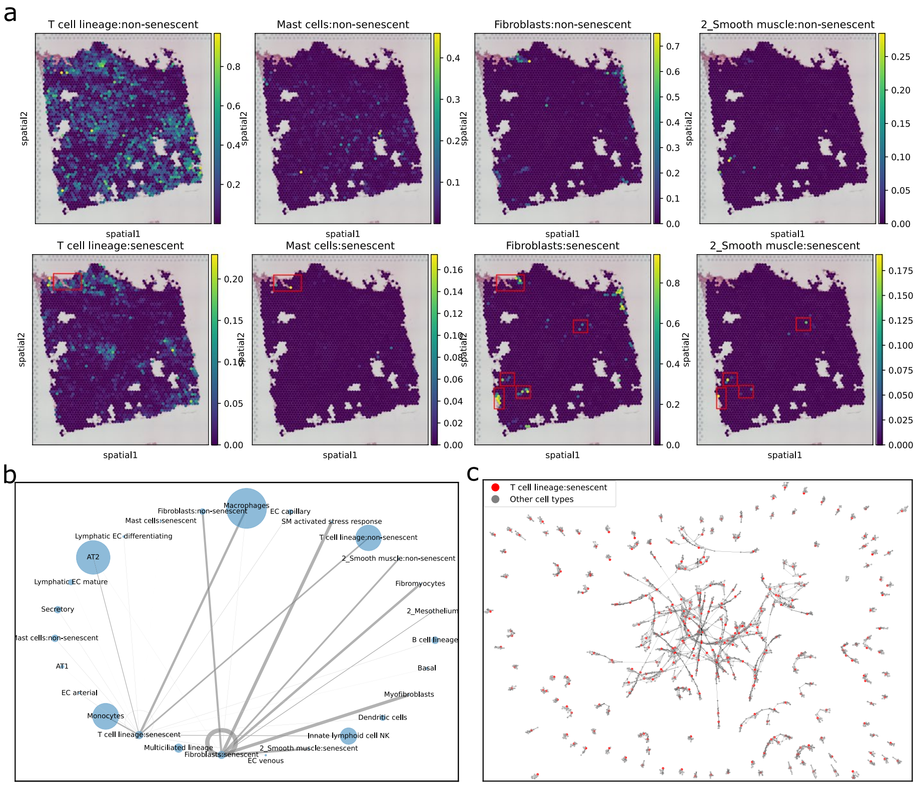

Fig. 6 The analysis of senescent cell-cell neighborhood for unpaired IPF dataset. **a** Spatial distribution of senescent and non-senescent cells for T cells, mast cells, fibroblasts, and 2-smooth muscle. Red rectangles indicate regions where senescent cells of multiple types are co-located. **b** The graph summarizes the spatial neighborhood of senescent fibroblasts and T cells. Nodes represent cell types, and edges represent direct neighboring relations in physical proximity. The size of the nodes corresponds to the number of cells within a cell type, and the width of the edges corresponds to the number of neighboring cells of a specific cell type (i.e., the total node degree per neighboring cell type). Edges representing a small number of neighbors are omitted. The graph demonstrates that senescent cells are neighbors to non-senescent cells within the same cell types, as well as to senescent cells belonging to different cell types. **c** The cell-cell spatial neighborhood graph of senescent cells of the T cell lineage. The validity of this neighborhood graph is assessed in Additional file 1: B (Supplementary Analysis)

incorporating internal references from cell type deconvolution, we can modulate and enhance the precision of this task.

Our method, scDOT, was shown to efficiently and accurately assign individual cells to their spatial origins using synthetic data. By combining OT and deconvolution scDOT improves on all prior methods we compared to. We also used scDOT to study and analyze new paired and unpaired spatial transcriptomics data from IPF and familial IPF lungs. We observed that senescent cells tend to co-localize in specific regions and are in close proximity to cells of the same type. While the distribution of senescent T cells appears sparse in both datasets, we noted a denser population of senescent fibroblast cells in the IPF lung compared to the familial IPF lung, which can be explained by the paracrine senescence and is consistent with previous studies indicating that senescent fibroblasts contribute to the pathogenesis of IPF through various mechanisms [40–42].

The integration of single-cell and spatial transcriptomics has been a topic of interest in recent years [1], with a number of multiview learning approaches suggested [43]. A crucial aspect of this integration is assessing the similarity of gene expression levels between cells and spatial spots. Unlike prior methods that utilized optimal transport, which rely on fixed cost matrices to represent the dissimilarity between cells and spots, scDOT utilizes a differentiable optimization layer in a deep declarative network to dynamically learn the cost matrix [24]. This use of optimal transport can be formulated as a domain adaptation problem, and the learned cost matrix holds potential for further applications involving mass transportation between the two modalities of other types of data.

Comparative studies and benchmarks exist for cell type deconvolution in spatial transcriptomics data [2, 44, 45]. Since there is no universal evaluation metric that applies to all scenarios, comparisons among methods depend on datasets and evaluation metrics used, such as root mean square error and Lin's concordance correlation coefficient, which may not consistently correlate [10, 44]. In our paper, we compared our method with recent approaches representing computational techniques like deep learning, probabilistic modeling, and optimal transport. While the performance of these methods may vary, certain high-performance methods, particularly Tangram [26], have been reported [2, 44, 45]. We note that the normalization of our synthetic data 2 precludes comparisons to methods utilizing count matrices as input, such as Stereoscope [8] and Cell2Location [5].

Limitations of scDOT include its reliance on corresponding scRNA-Seq data, which may not be fully accessible or directly applicable to the spatial region under investigation. This can result in inaccuracies in mapping cells to spots. One approach to mitigate this issue is through the utilization of unbalanced optimal transport, a variant of the problem that does not necessitate equal masses for the two distributions. Exploring this avenue is part of our future research agenda. Additionally, scDOT encounters challenges when faced with large spots containing 25 or more cells, especially of some of these cells are rare.

An important component of our biological analysis focused on IPF and familial IPF lung tissue was the identification of senescent cells. Evaluating cellular senescence poses challenges as there are various approaches, such as assessing senescent gene markers or morphological features of senescent cells. Additionally, different cell types or diseases may require distinct sets of senescent markers due to the complex nature of the senescence process. In our study, we employed a combined list of senescent markers and categorized cells within each cell type as either senescent or non-senescent. However, senescent states can exist on a continuum, ranging from non-senescence to primary senescence, and different markers may be associated with primary and secondary senescence. Still, using scDOT, we were able to identify cell-cell spatial neighborhood, which can aid in assessing senescent cells in close physical proximity. It also allowed us to explore how senescent cells reorganize and impact their environment and nearby cells. Cells neighboring senescent cells can transition into a secondary senescent state. Hence, the influence of senescence can be approached as a diffusion problem within a network, where cells reach a senescent state through contact with senescent neighbors. This network-based diffusion approach, relying on the spatial mapping of individual cells to their origins, holds promise for fruitful future investigations.

scDOT is implemented in PyTorch and is available for download from https://github.com/namtk/scDOT. The datasets used in this study are available as part of the Cellular Senescence Network [46] at https://data.sennetconsortium.org/ under identifiers SNT283.VJLG.263, SNT863.DDDW.723, SNT362.WCQB.995, SNT597.VKCR.938, and SNT964.RXQC.298.

Conclusions

In conclusion, we presented scDOT, a novel method that integrates single-cell and spatial transcriptomics, improving upon existing approaches by incorporating cell type deconvolution and optimal transport. Our study demonstrates scDOT's capability to accurately map individual cells to spatial origins, providing insights into cellular senescence in IPF and familial IPF lung tissue. By analyzing senescent cell proximity and interactions, we highlighted the potential of scDOT in uncovering the spatial dynamics of senescence. While scDOT faces challenges, such as reliance on scRNA-Seq data and limitations in mapping cells within large spatial spots, its application offers promising avenues for further research into senescence and disease progression.

Methods

Datasets

To investigate the effects of the proposed method that combines cell-type deconvolution and spatial reconstruction, we collected both synthetic and real data. Since there is no immediate method to assess the performance of cell-type deconvolution and spatial reconstruction tasks on real data, we generated two simulation datasets to evaluate and benchmark scDOT as well as other related methods against the ground truth. It is important to note that, for benchmarking the deconvolution task, methods designed for spatial reconstruction can be utilized. However, for benchmarking the reconstruction task, methods solely designed for cell-type deconvolution cannot be used, as inferring the fine-grained mapping γ of individual cells from a coarse-grained mapping P of cell clusters poses a challenging inverse problem, even though inferring the cell type proportion P from the coupling matrix γ is straightforward ($P = \gamma \times C$).

Synthetic datasets

Synthetic dataset 1 The synthetic data 1 is generated based on Gaussian process (GP) by assuming that the nearby spots have similar proportions of cell types as well as gene expressions [7]. Here, we used scRNA-seq data of an IPF lung tissue and projected the cells from this data onto grids, which represent the spatial coordinates obtained from a different IPF lung sample's upper lobe lung slice. Thus, scRNA-seq data is real while spatial locations are synthetic for this dataset. See supporting methods for more details.

Synthetic dataset 2 For the synthetic data 2, we conducted simulations using gene expression data from individual cells obtained through multiplex error-robust fluorescence in situ hybridization (MERFISH) in the mouse medial preoptic area (MPOA) [11, 47]. By aggregating the gene expression information of cells within spatially contiguous pixels, we created a representation of the spatial organization. See supporting methods for more details.

Real datasets

Preparation and data collection of single-cell RNA sequencing and spatial transcriptomics Tissue samples were obtained by the Human Tissue Biorepository at The Ohio State University from the explanted lungs of patients diagnosed with idiopathic pulmonary fibrosis (IPF) and familial IPF after a Total Transplant Care Protocol informed consent and research authorization from the patient. The tissue biorepository operates in accordance with NCI and ISBER Best Practices for Repositories.

For single-cell RNA sequencing (scRNA-seq): Samples of 15 g of upper and lower lobe lung parenchyma tissue were washed with PBS, minced finely with scalpels, and digested using an enzyme cocktail (1 mg/mL of Liberase DL, DNase I, DMEM) for 2 h at 37 °C with rocking. Cell suspension was filtered through a serial filter of 300 μm, 100 μm, and 70 μm strainers. After straining, the cell suspension was centrifuged at 500g for 7 min, the supernatant was removed, and 1x RBC lysis buffer was added to the pellet and incubated at 4 °C for 7 min and then filtered through a 40-μm strainer to remove the agglomerated dead cells. Finally, cell number and viability were determined using a countess automatic cell counter (Invitrogen). Whole lung cell suspension was loaded on the Chromium Controller, according to 10x Genomics protocol. $3'$ Gene Expression libraries were sequenced on Illumina sequencer with read lengths of 28 cycles Read 1, 10 cycles i7 index, 10 cycles i5 index, 90 cycles Read 2. ScRNA-seq data was extracted from the raw sequencing data using Cell Ranger (version 7.1.0, 10x Genomics).

For spatial transcriptomics: Tissue sections of $\leq 6.5 \times 6.5$ mm from the upper and lower lobe of lung parenchyma were used for spatial analysis. After collection, samples were fixed for 24 h in 10% neutral buffered formalin and embedded in paraffin (wax) to create a formalin-fixed paraffin-embedded (FFPE) block. Sections of 5 μm were then cut from the FFPE blocks onto Visium slides (10x Genomics) and processed according to the manufacturer’s protocol. Scan of H&E staining was performed with EVOSTM M7000 microscope (Invitrogen) using a 10x objective. FFPE libraries were prepared according to 10x Genomics protocol and sequenced on Illumina sequencer to a read depth of at least 25k reads/spot, with read lengths of 28 cycles Read 1, 10 cycles i7 index, 10 cycles i5 index, 50 cycles Read 2. Spatial transcriptomics data was extracted from the raw sequencing data using Space Ranger (version 2.0.0, 10x Genomics).

Paired familial IPF lung dataset We obtained two paired datasets of single-cell and spatial transcriptomics from a patient with familial IPF lung, one of which is from the upper lobe slice and the other from the lower lobe slice. The upper lobe pair contains 6762 cells and 3336 spots while the lower lobe pair contains 6173 cells and 2246 spots. For each of these two paired datasets, we preprocessed the data by (1) removing lowly expressed genes of both two data modalities, keeping genes that have at least 10 counts, and (2) removing cells with low counts, keeping cells that have at least 500 counts and 500 genes expressed, then (3) obtaining the common gene sets for both modalities by taking the intersection of the two gene sets.

The cell type annotations were transferred from the Lung Cell Atlas (HLCA) using scArches and FastGenomics platform.

To re-annotate cells that reflect senescent states, we utilized a list of 68 senescent markers (Additional file 2) and then calculated the average expression of the marker genes across all cells. Next, senescent cells were identified as having a higher than 95 percentile of average expression of the marker genes.

Unpaired IPF lung dataset To demonstrate the general utility of the method even for non-paired data, we obtained an unpaired scRNA-seq and spatial transcriptomics dataset from two different IPF patients. While the preparation for spatial transcriptomics is the same as for the paired data, the preparation for single-cell RNA sequencing is described as follows. Single-cell sequencing of human lung tissue was performed as previously described [48, 49]. In short, human lung tissue (IPF) was homogenized, and 4 g of tissue were digested by dispase/collagenase (collagenase: 0.1U/mL, dispase: 0.8 U/mL, Roche) for 1 h at 37 °C. Samples were successively filtered through nylon filters (100 μm and 20 μm) followed by a Percoll gradient. Single epithelial cell suspensions were loaded onto a Chromium Single-Cell Chip (Chromium™ Single Cell 3′ Reagent Kit, v2 Chemistry) to obtain single-cell 3′ libraries for sequencing. cDNA obtained after droplet reverse transcription was amplified for 14 cycles and analyzed using Agilent Bioanalyzer. The barcoded libraries were sequenced using Illumina NextSeq-500 through the University of Pittsburgh Genomics Core Sequencing Facility, aiming for 100,000 reads per cell and capturing 10,000 per library.

The single-cell data contains 25,260 cells, and the spatial data, which consists of an upper lobe lung slide, contains 3412 spots. The preprocessing, cell type annotation, and senescence re-annotation were carried out following the same procedures as for the paired familial IPF lung dataset.

It is important to note that, since the familial IPF lung datasets are paired, the coordinates of cells after spatial reconstruction represent the actual tissue coordinates. However, for the unpaired IPF lung dataset, the inferred cell coordinates do not directly reflect the actual tissue coordinates. Instead, they serve as an intermediate step to infer the relative spatial relationships among cells.

Cell type deconvolution

For gene expression, cell type deconvolution can be formulated as a nonnegative least squares (NNLS) problem, where the goal is to estimate the relative abundances of different cell types by solving for the nonnegative coefficients of a linear combination of their respective gene expression profiles. Specifically, a multicellular resolution spatial transcriptomics profile $Y \in \mathbb{R}^{m \times p}$ of p genes across m spots each of which contains transcripts from multiple cells can be represented as $Y = PS$ in which $P \in \mathbb{R}^{m \times c}$ is the cell type proportions to be estimated and $S \in \mathbb{R}^{c \times p}$ is the signature matrix consisting of known gene expression profiles for each cell type of the total c cell types. We solved for P the following nonnegative least squares problem:

$$P^* \in \underset{P \geq 0}{\arg\min} \|Y - PS\|_F \tag{1}$$

There are several solvers available for solving a NNLS problem, including Lawson-Hanson's active set method [50]. Here, we used projected gradient descent [51].

Mapping single cell to spatial images

The spatial reconstruction task involves assigning cells from scRNA-seq data to a predicted corresponding location in a tissue sample. Note that such assignment, implicitly, also provides deconvolution of the spot data assuming that the cell types for cells in the scRNA-Seq data are known. Here, we formulate this as an optimal transport (OT) problem from scRNA-seq dataset $X \in \mathbb{R}^{n \times p}$ of p genes across n cells to spatial transcriptomics dataset $Y \in \mathbb{R}^{m \times p}$ of p genes across m. OT is commonly used to model the coupling between two probability distribution. In our case, we use it to model the transport of gene expression from one dataset to another in an optimal way. By solving the optimal transport problem, it is possible to estimate the optimal coupling and quantify the degree of similarity between the datasets. Formulating the spatial reconstruction task as an optimal transport problem involves constructing a cost matrix $M \in \mathbb{R}_{+}^{m \times n}$ representing distances between cells of X and spots of Y. Here, we used cosine distance $d_{cos}(X_{i,:}, Y_{j,:}) = 1 - \frac{\langle X_{i,:}, Y_{j,:} \rangle}{\|X_{i,:}\| \cdot \|Y_{j,:}\|}$, which is scale-invariant and can account for differences in measurement sensitivity between the two technologies. Furthermore, scale-invariant cosine dissimilarity is well-suited for handling the fact that expression of a spot in the spatial transcriptomics dataset is the mixture or sum of multiple cells in the scRNA-seq dataset. Specifically, the coupling matrix $\gamma \in \mathbb{R}_{+}^{m \times n}$ is solved for obtaining the optimal transport as follows:

$$\begin{aligned} \gamma^* \in \arg\min_{\gamma \in \mathbb{R}_{+}^{m \times n}} & \textstyle\sum_{i,j} \gamma_{i,j} M_{i,j} + \lambda \Omega(\gamma) \\ \text{s.t. } & \gamma 1 = m; \gamma^T 1 = n; \gamma \geq 0 \end{aligned} \tag{2}$$

where $M \in \mathbb{R}_{+}^{m \times n}$ is the cost matrix defining the cost of moving gene expression from cell a_i to spot b_j and $\Omega(\gamma) = \sum_{i,j} \gamma_{i,j} \log\left(\gamma_{i,j}\right)$ is an entropic regularization term [52]. The entropic regularization version of optimal transport can be solved by Sinkhorn-Knopp's alternative projection algorithm [52]. In other words, this minimization process aims to match cells with similar expression profiles to spots with similar transcriptomic characteristics, measured by cosine similarity, thereby capturing the underlying biological relationships between the two datasets.

It is important to note that we utilized the entropy regularization version of optimal transport, resulting in a probabilistic mapping between cells and spots. This probabilistic coupling, represented by the left-stochastic matrix γ, indicates the likelihood of a specific cell being associated with a particular spot. This probabilistic coupling offers computational efficiency and eliminates assumptions about the number of cells in a spot, including cases where a cell may reside on the boundary of two spots.

Combination of deconvolution and mapping

OT for spatial and scRNA-Seq data is challenging since spatial data is often sparse leading to less dependable inferred individual cell-spot pairs. We thus further extend OT by incorporating the deconvolution result, which, as mentioned above, maps a group of cells to a group of spots. As a result, scDOT integrates the two mentioned data modalities, single-cell and spatial transcriptomics, by simultaneously solving the deconvolution and OT problems. Specifically, given paired data modalities X and Y, representing gene expression profiles of a scRNA-seq and a spatial transcriptomics data respectively, scDOT simultaneously solves the deconvolution problem of estimating cell type fractions, P, of c cell types across m spots, and the spatial reconstruction problem of mapping n cells to their m spatial origins, resulting in a coupling matrix γ. These two solutions are constrained by the relation $\gamma \times C = P$, where C is a binary matrix representing the cell type of each cell, encoded as a one-hot vector of size $1 \times c$ across the total n cells. The two results, P and γ, are computed simultaneously in an iterative manner in order to improve each other's accuracy. The problem is then formed as a bi-level optimization where the deconvolution and the spatial reconstruction are two inner optimization problems nested inside the outer optimization that reflects the relation $\gamma \times C = P$. See supporting methods for more details.

The integration of cell type deconvolution and spatial mapping can be formulated as a bi-level optimization problem. This problem consists of two inner optimization problems: the deconvolution and the spatial reconstruction, which are nested inside the outer optimization problem representing the relationship between $\gamma \times C$ and P. The deconvolution problem is formulated as a NNLS optimization, while the spatial reconstruction problem is formulated as an OT problem:

$$\begin{aligned} &\min_{\gamma^*,P^*} \|\gamma^* C - P^*\|_F \\ &\quad \text{s.t. } P^* \in \arg\min_{P\in\mathbb{R}_+^{m\times c}} \|Y - PS\|_F \\ &\qquad \boldsymbol{\gamma^*} \in \arg\min_{\gamma\in\mathbb{R}_+^{m\times n}} \sum_{i,j} \gamma_{i,j} M_{i,j} + \lambda\Omega(\boldsymbol{\gamma}) \\ &\qquad\quad \text{s.t. } \boldsymbol{\gamma}1 = m;\ \boldsymbol{\gamma}^T 1 = n;\ \boldsymbol{\gamma} \geq 0 \end{aligned} \tag{3}$$

This bilevel optimization can be rendered differentiable, enabling efficient computation of backpropagation using the implicit differentiation theorem. It can be implemented by training a declarative neural network [24] in an end-to-end manner, allowing for iterative learning of the deconvolution and spatial reconstruction results until convergence is reached.

The optimization procedure relies on gradient descent facilitated by backpropagation. Each iteration involves updates calculated using the implicit differentiation theorem. Specifically, during the forward pass of the optimal transport layer, the Sinkhorn-Knopp solver [52] is employed, while the backward pass entails dealing with the gradient of the argmin operator, particularly focusing on computing the Jacobian-vector product of the adjoint variable, which presents a significant challenge. As a result, the time complexity of scDOT is contingent upon the number of iterations, typically $O(Tn)$, with a convergence rate of $O(1/T)$, where T represents the iterations and n

denotes the number of samples, akin to a conventional gradient descent algorithm. The hyperparameter tuning process is outlined in detail in Additional file 1 (A.1).

Inference of cell-cell spatial neighborhood graph

Utilizing the coupling matrix learned from optimal transport, we employed manifold alignment [43, 53] to project the single-cell data X and spatial coordinates $Z \in \mathbb{R}^{m\times 2}$ of spatial transcriptomics data onto a common nonlinear subspace. This subspace preserves the correspondence between cells and spots as well as the intrinsic similarity within each dataset. Consequently, in the common subspace, cells are represented in terms of both gene expression and spatial coordinates. Subsequently, we constructed a k-nearest neighbor graph (k-NNG) based on this new representation, which consists of the new coordinates in the common subspace for each cell. This allowed us to obtain the cell-cell spatial neighborhood graph (in our experiments, we set $k = 10$). The projections f and g resulting from manifold alignment serve as minimizers of the following optimization problem, which can be formulated as a generalized eigenvalue problem:

$$\begin{aligned} f^*, g^* = &\arg\min_{f,g} (1-\mu) \sum_{i=1}^{m}\sum_{j=1}^{n} \left\| f(x_i) - g\left(z_j\right) \right\|_2^2 \gamma^{i,j} \\ &+ \mu \sum_{i=1}^{m}\sum_{j=1}^{n} \left\| f(x_i) - f\left(x_j\right) \right\|_2^2 W_X^{i,j} \\ &+ \mu \sum_{i=1}^{m}\sum_{j=1}^{n} \left\| g(x_i) - g\left(z_j\right) \right\|_2^2 W_Z^{i,j} \end{aligned} \tag{4}$$

where W_X and W_Z are adjacent matrices of kNN graphs of X and Z, respectively.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03426-0.

Additional file 1. Supplementary methods, Supplementary analysis, Supplementary figures S1-S5.

Additional file 2. List of senescent marker genes.

Additional file 3. Review history.

Peer review information

Andrew Cosgrove was the primary editor of this article at Genome Biology and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The reivew history is available as Additional file 3.

Authors' contributions

N.D.N., J.L.-M., and Z.B.-J. conceptualized the study. N.D.N., J.L.-M., and Z.B.-J. designed the algorithm and methodology. P.J. and N.D.N. generated the synthetic datasets. E.H., M.K., O.E., and Z.B.-J. generated combined senescent marker gene list. L.R., T.K., J.A.O.-R., M.R., and A.L.M. generated the spatial transcriptomics data. L.R., T.K., J.A.O.-R., M.B., M.K., O.E., M.R., and A.L.M. generated the single-cell transcriptomics data. N.D.N. developed the software. N.D.N, J.L.-M., and Z.B.-J. performed the synthetic data analysis. All authors performed the analyses for senescent cells. N.D.N., L.R., P.J., J.L.-M., and Z.B.-J. wrote the manuscript. All authors read and approved the final manuscript.

Funding

The work in this article was partially supported by National Institutes of Health (NIH) grants OT2OD026682 and 1U24CA268108 to Z.B.-J and 1U54AG075931 to I.R., M.K., A.L.M., and T.F. This work used Bridges-2 GPU at the Pittsburgh Supercomputing Center through allocation CIS230034P and HMCMUTC from the Advanced Cyberinfrastructure

Coordination Ecosystem: Services and Support (ACCESS) program, which is supported by National Science Foundation grants 2138259, 2138286, 2138307, 2137603, and 2138296.

Data availability

The synthetic datasets generated and analyzed during the current study are available at Zenodo [54].

The scDOT package and data processing code are available at https://github.com/namtk/scDOT and Zenodo [55]. The synthetic data and the scDOT package are released with MIT license.

Declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 3 August 2023 Accepted: 22 October 2024
Published online: 08 November 2024

References

1. Longo SK, Guo MG, Ji AL, Khavari PA. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. Nat Rev Genet. 2021;22(10):627–44.
2. Li H, Zhou J, Li Z, Chen S, Liao X, Zhang B, et al. A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics. Nat Commun. 2023;14(1):1548.
3. Elosua-Bayes M, Nieto P, Mereu E, Gut I, Heyn H. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. Nucleic Acids Res. 2021;49(9):e50–e50.
4. Ma Y, Zhou X. Spatially informed cell-type deconvolution for spatial transcriptomics. Nat Biotechnol. 2022;40(9):1349–59.
5. Kleshchevnikov V, Shmatko A, Dann E, Aivazidis A, King HW, Li T, et al. Cell 2location maps fine-grained cell types in spatial transcriptomics. Nat Biotechnol. 2022;40(5):661–71.
6. Cable DM, Murray E, Zou LS, Goeva A, Macosko EZ, Chen F, et al. Robust decomposition of cell type mixtures in spatial transcriptomics. Nat Biotechnol. 2022;40(4):517–26.
7. Lopez R, Li B, Keren-Shaul H, Boyeau P, Kedmi M, Pilzer D, et al. DestVI identifies continuums of cell types in spatial transcriptomics data. Nat Biotechnol. 2022;40(9):1360–9.
8. Andersson A, Bergenstråhle J, Asp M, Bergenstråhle L, Jurek A, Fernández Navarro J, et al. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. Commun Biol. 2020;3(1):565.
9. Dong R, Yuan GC. SpatialDWLS: accurate deconvolution of spatial transcriptomic data. Genome Biol. 2021;22(1):145.
10. Chen Y, Wang Y, Chen Y, Cheng Y, Wei Y, Li Y, et al. Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis. Nat Commun. 2022;13(1):6735.
11. Miller BF, Huang F, Atta L, Sahoo A, Fan J. Reference-free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data. Nat Commun. 2022;13(1):2339.
12. Zhang Y, Lin X, Yao Z, Sun D, Lin X, Wang X, Yang C, Song J. Deconvolution algorithms for inference of the cell-type composition of the spatial transcriptome. Comput Struct Biotechnol J. 2023;21:176–84.
13. Overholtzer M. Senescent cells feed on their neighbours. Nature Publishing Group UK London; 2019.
14. Fafián-Labora JA, O'Loghlen A. Classical and nonclassical intercellular communication in senescence and ageing. Trends Cell Biol. 2020;30(8):628–39.
15. Gasek NS, Kuchel GA, Kirkland JL, Xu M. Strategies for targeting senescent cells in human disease. Nat Aging. 2021;1(10):870–9.
16. Walker BL, Cang Z, Ren H, Bourgain-Chang E, Nie Q. Deciphering tissue structure and function using spatial transcriptomics. Commun Biol. 2022;5(1):220.
17. Karaiskos N, Wahle P, Alles J, Boltengagen A, Ayoub S, Kipar C, et al. The Drosophila embryo at single-cell transcriptome resolution. Science. 2017;358(6360):194–9.
18. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36(5):411–20.
19. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888–902.
20. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-cell multi-omic integration compares and contrasts features of brain cell identity. Cell. 2019;177(7):1873–87.
21. Peyré G, Cuturi M, et al. Computational optimal transport: With applications to data science. Found Trends® Mach Learn. 2019;11(5–6):355–607.
22. Villani C, et al. Optimal transport: old and new, vol. 338. Springer; 2009.
23. Colson B, Marcotte P, Savard G. An overview of bilevel optimization. Ann Oper Res. 2007;153:235–56.
24. Gould S, Hartley R, Campbell D. Deep declarative networks. IEEE Trans Pattern Anal Mach Intel. 2021;44(8):3988–4004.
25. Janesick A, Shelansky R, Gottscho AD, Wagner F, Williams SR, Rouault M, et al. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. Nat Commun. 2023;14(1):8353.
26. Biancalani T, Scalia G, Buffoni L, Avasthi R, Lu Z, Sanger A, et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. Nat Methods. 2021;18(11):1352–62.

27. Nitzan M, Karaiskos N, Friedman N, Rajewsky N. Gene expression cartography. Nature. 2019;576(7785):132–7.
28. Hewitt RJ, Lloyd CM. Regulation of immune responses by the airway epithelial cell landscape. Nat Rev Immunol. 2021;21(6):347–62.
29. Deprez M, Zaragosi LE, Truchi M, Becavin C, Ruiz García S, Arguel MJ, et al. A single-cell atlas of the human healthy airways. Am J Respir Crit Care Med. 2020;202(12):1636–45.
30. Long AF, Stearns T. A not-so-simple twist of fate. Dev Cell. 2021;56(4):402–4.
31. Peyser R, MacDonnell S, Gao Y, Cheng L, Kim Y, Kaplan T, et al. Defining the activated fibroblast population in lung fibrosis using single-cell sequencing. Am J Respir Cell Mol Biol. 2019;61(1):74–85.
32. Kim N, Vu TH. Parabronchial smooth muscle cells and alveolar myofibroblasts in lung development. Birth Defects Res C Embryo Today Rev. 2006;78(1):80–9.
33. Severino V, Alessio N, Farina A, Sandomenico A, Cipollaro M, Peluso G, et al. Insulin-like growth factor binding proteins 4 and 7 released by senescent cells promote premature senescence in mesenchymal stem cells. Cell Death Dis. 2013;4(11):e911–e911.
34. Nagaraj K, Sarfstein R, Laron Z, Werner H. Long-term IGF1 stimulation leads to cellular senescence via functional interaction with the thioredoxin-interacting protein, TXNIP. Cells. 2022;11(20):3260.
35. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. Cell PhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. Nat Protoc. 2020;15(4):1484–506.
36. Pereira BI, Devine OP, Vukmanovic-Stejic M, Chambers ES, Subramanian P, Patel N, et al. Senescent cells evade immune clearance via HLA-E-mediated NK and CD8+ T cell inhibition. Nat Commun. 2019;10(1):2387.
37. Nakamura Y, Iwata H, Kuwayama T, Shirasuna K. S100A8, which increases with age, induces cellular senescence-like changes in bovine oviduct epithelial cells. Am J Reprod Immunol. 2019;82(3):e13163.
38. Feng X, Wang L, Zhou R, Zhou R, Chen L, Peng H, et al. Senescent immune cells accumulation promotes brown adipose tissue dysfunction during aging. Nat Commun. 2023;14(1):3208.
39. Mehdizadeh M, Aguilar M, Thorin E, Ferbeyre G, Nattel S. The role of cellular senescence in cardiac disease: basic biology and clinical relevance. Nat Rev Cardiol. 2022;19(4):250–64.
40. Álvarez D, Cárdenes N, Sellarés J, Bueno M, Corey C, Hanumanthu VS, et al. IPF lung fibroblasts have a senescent phenotype. Am J Physiol Lung Cell Mol Physiol. 2017;313(6):L1164–73.
41. Waters DW, Blokland KE, Pathinayake PS, Burgess JK, Mutsaers SE, Prele CM, et al. Fibroblast senescence in the pathology of idiopathic pulmonary fibrosis. Am J Physiol Lung Cell Mol Physiol. 2018;315(2):L162–72.
42. Lin Y, Xu Z. Fibroblast senescence in idiopathic pulmonary fibrosis. Front Cell Dev Biol. 2020;8:593283.
43. Nguyen ND, Wang D. Multiview learning for understanding functional multiomics. PLoS Comput Biol. 2020;16(4):e1007677.
44. Li B, Zhang W, Guo C, Xu H, Li L, Fang M, et al. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. Nat Methods. 2022;19(6):662–70.
45. Yan L, Sun X. Benchmarking and integration of methods for deconvoluting spatial transcriptomic data. Bioinformatics. 2023;39(1):btac805.
46. SenNet Consortium. NIH SenNet Consortium to map senescent cells throughout the human lifespan to understand physiological health. Nat Aging. 2022;2(21):1090–100.
47. Moffitt JR, Bambah-Mukku D, Eichhorn SW, Vaughn E, Shekhar K, Perez JD, et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. Science. 2018;362(6416):eaau5324.
48. Valenzi E, Bulik M, Tabib T, Morse C, Sembrat J, Bittar HT, et al. Single-cell analysis reveals fibroblast heterogeneity and myofibroblasts in systemic sclerosis-associated interstitial lung disease. Ann Rheum Dis. 2019;78(10):1379–87.
49. Heinzelmann K, Hu Q, Hu Y, Dobrinskikh E, Ansari M, Melo-Narváez MC, et al. Single-cell RNA sequencing identifies G-protein coupled receptor 87 as a basal cell marker expressed in distal honeycomb cysts in idiopathic pulmonary fibrosis. Eur Respir J. 2022;59(6):2102373.
50. Lawson CL, Hanson RJ. Solving least squares problems. Classics in applied mathematics, vol. 15. SIAM; 1995. https://dblp.org/rec/books/daglib/0000911.bib.
51. Lin CJ. Projected gradient methods for nonnegative matrix factorization. Neural Comput. 2007;19(10):2756–79.
52. Cuturi M. Sinkhorn distances: lightspeed computation of optimal transport. In: Burges CJC, Bottou L, Ghahramani Z, Weinberger KQ. Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Lake Tahoe; 2019. p. 2292–300. https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.
53. Wang C, Mahadevan S. A general framework for manifold alignment. In: Manifold Learning and Its Applications, Papers from the 2009 (AAAI) Fall Symposium. Arlington: AAAI; 2009. http://aaai.org/ocs/index.php/FSS/FSS09/paper/view/852.
54. Nguyen. synthetic data for scDOT. Zenodo; 2024. https://doi.org/10.5281/zenodo.13869290.
55. Nguyen ND. namtk/scDOT: 0.0.1. Zenodo; 2024. https://doi.org/10.5281/zenodo.13869260.

Publisher's Note