Briefings in Bioinformatics, 22(5), 2021, 1–13

https://doi.org/10.1093/bib/bbaa414
Method Review

DSTG: deconvoluting spatial transcriptomics data through graph-based artificial intelligence

Qianqian Song and Jing Su

Corresponding authors: Jing Su, Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN 46202, USA. Tel.: +1-678-642-7428. E-mail: su1@iu.edu; Qianqian Song, Department of Cancer Biology, Wake Forest School of Medicine, Winston Salem, NC 27157, USA. Tel.: +1-336-926-4972. E-mail: qsong@wakehealth.edu

Abstract

Recent development of spatial transcriptomics (ST) is capable of associating spatial information at different spots in the tissue section with RNA abundance of cells within each spot, which is particularly important to understand tissue cytoarchitectures and functions. However, for such ST data, since a spot is usually larger than an individual cell, gene expressions measured at each spot are from a mixture of cells with heterogenous cell types. Therefore, ST data at each spot needs to be disentangled so as to reveal the cell compositions at that spatial spot. In this study, we propose a novel method, named deconvoluting spatial transcriptomics data through graph-based convolutional networks (DSTG), to accurately deconvolute the observed gene expressions at each spot and recover its cell constitutions, thus achieving high-level segmentation and revealing spatial architecture of cellular heterogeneity within tissues. DSTG not only demonstrates superior performance on synthetic spatial data generated from different protocols, but also effectively identifies spatial compositions of cells in mouse cortex layer, hippocampus slice and pancreatic tumor tissues. In conclusion, DSTG accurately uncovers the cell states and subpopulations based on spatial localization. DSTG is available as a ready-to-use open source software (https://github.com/Su-informatics-lab/DSTG) for precise interrogation of spatial organizations and functions in tissues.

Key words: spatial transcriptomics; deconvolution; graph-based artificial intelligence; single-cell RNA-seq

Introduction

Cells of different types are spatially and structurally organized within tissues to perform their functions. Uncovering the complex spatial architecture of heterogenous tissue is significant for understanding the cellular mechanisms and functions in diseases. The fast advance of single-cell RNA sequencing technologies (scRNA-seq) attracts the attention to elucidate the heterogenous cell formation [1–4] and trace the lineage relationship within tissue [5–7]. Unfortunately, due to the lack of spatial information, scRNA-seq is incapable of identifying the structural organization of heterogeneous cells within a complex tissue. Therefore, as the complementary to scRNA-seq, spatially resolved transcriptomic profiling methods [8–10] have been introduced. To reveal the spatial cytoarchitectures within tissues, sequencing-based high-throughput spatial transcriptomics (ST) technologies [11–14], such as 10X Genomics Visium [8] and Slide-seq [15, 16], use spatially indexed barcodes with RNA sequencing that allows quantitative analysis of the transcriptome with spatial resolution in individual tissue sections.

Qianqian Song is a postdoctoral research fellow in the Department of Cancer Biology, Wake Forest School of Medicine, NC, USA. Her major is bioinformatics and her recent work focuses on developing graph-based tools for single-cell genomics and spatial transcriptomics data.
Jing Su is an assistant professor in the Department of Biostatistics, Indiana University School of Medicine, IN, USA. His research focuses on graph artificial intelligence and machine learning in biomedical informatics and precision health.
Submitted: 26 October 2020; **Received (in revised form):** 8 December 2020

Emerging ST technologies are able to spatially index transcripts and measure expression profiles, advancing our understanding of precise tissue architectures. However, the resolution of ST data is far lower than single-cell level. Transcripts captured at a specific location by a 'spot' [8] or a 'bead' [15, 16] is usually composed of a mixture of heterogenous cells. For example, Visium, one of the microarray-based ST techniques developed by 10X Genomics, uses spots of 50 μm diameter, with each spot covering 10–20 cells in average, which varies depending on the tissue histology [17]. Even for the Slide-seq [15, 16] that quantifies gene expression with high resolution (10 microns), one pixel may still be overlapped with multiple cells. As a result, the measured gene expressions at a 'spot' reflect a mixture of cells. Therefore, uncovering the cell compositions within each spot of the ST data is critical for investigating tissue's molecular and cellular architecture at high resolution.

To address this problem, very few tailored approaches have been developed yet. SPOTlight [18] is a deconvolution algorithm using nonnegative matrix factorization regression and nonnegative least squares, which has been applied to ST data [16] successfully. Specifically, SPOTlight incorporates the reference scRNA-seq data to identify cell type-specific topic profiles, which is further used to deconvolute spatial spots. This method leverages scRNA-seq data for the identification of cell states and subpopulations to deconvolute the ST data, showing that leveraging well-characterized scRNA-seq data will aid and facilitate the exploration of spatial datasets. A major limit of this ST deconvolution method is that the intrinsic topological information of cell type constitutions within spots, which provides crucial information about the relations between the observed gene expression patterns and associated cell types at spots, cannot be effectively learned and utilized.

In recent years, graph convolutional networks (GCN) [19] have demonstrated promising capability in utilizing such intrinsic topological information of data to improve model performance. The topological relations inside the data, such as similarity between samples, can be represented as graphs. Through learning the shared kernel used in spectral graph convolution across all nodes in a graph, a semi-supervised GCN model captures local graph structures as well as node features and incorporates both information as latent space representation. GCN [19] and its variants [20, 21] have been applied to different scenarios successfully, including cancer patient subtyping using real-world evidence [22], protein prediction [23] and drug design [24], as well as single cells and diseases [25–29]. These works shows that, through effectively learning and leveraging the latent representation and topological relations among data, GCN models are able to significantly improve learning performance.

In this work, we have developed a novel graph-based artificial intelligence (AI) model, deconvoluting spatial transcriptomics data through graph-based convolutional networks (DSTG), for reliable and accurate decomposition of cell mixtures in the spatially resolved transcriptomics data. Based on the well-characterized scRNA-seq dataset, DSTG is able to learn the precise composition of ST data using semi-supervised GCN. The performance of DSTG has been validated on synthetic ST data, as well as on different experimental ST datasets with well-defined structures including mouse cortex layer, hippocampus tissue and pancreatic tumor tissues. In addition, we provided the implementation software of DSTG as a ready-to-use Python package, which is compatible with current ST profiling datasets for accurate cell type deconvolution.

Materials and methods

Variable gene selection

For the scRNA-seq data, we first identified genes that exhibited the most variability across different cell types using the analysis of variance. The top 2000 most variable gene features in the scRNA-seq data are selected according to adjusted *P*-values with Bonferroni correction. Then, we used the scRNA-seq data with the top variable genes to generate the pseudo-ST data with synthetic mixtures of cells with known cell compositions. For simplicity and illustration, we consistently used the term 'spot' to represent the synthetic cell mixture of the pseudo-ST data as well as a spot or a bead of real-ST data.

Pseudo-ST data

Real-ST generated by ST assay captures the gene expression at each spot that covers a heterogeneous cell mixture. These cell mixtures can be mimicked by and constructed from scRNA-seq data from the same tissue. Specifically, to mimic the cell mixture at a spot, we selected two to eight cells from the scRNA-seq datasets and combined their transcriptomic profiles as the pseudo-ST data. The number of selected cells is similar to the spatial resolution of real-ST data. Herein, the exact proportions of cell types at each pseudo-ST spot are available since the identities of the selected cells are known. In order to better mimic the real-ST spot data, if the total Unique Molecular Identifier (UMI) counts of the resulting pseudo-ST data exceed that of the real-ST data, we downsampled it accordingly. Therefore, this pseudo-ST data resembles the real-ST data that is obtained from the same tissue. To further ensure that DSTG leverages the similarity between pseudo- and real-ST data, we learned a link graph to connect similar spots between pseudo-ST and real-ST, which is used as the input graph of DSTG.

Link graph

For both pseudo-ST data and real-ST data, we first performed the data normalization: the raw UMI counts of a gene in a cell are first divided by the total counts for that cell (library size normalization), then multiplied by a size factor of 10 000, and finally log transformed with one added. The normalized data are then subjected to standardized transformation, i.e.

$$x_{g,i} = \frac{x^0_{g,i} - \overline{x^0_g}}{\rho_g},$$

where $x^0_{g,i}$ is the normalized counts of gene g and spot i, $\overline{x^0_g}$ is the mean of $x^0_{g,i}$ over all spots and ρ_g is the SD of $x^0_{g,i}$. Thus, $x_{g,i}$ is the standardized gene expression.

After the data standardization, we then built a link graph incorporating pseudo-ST and real-ST data for the DSTG method. The built graph is $G = (V, E)$, with $N = \mid V \mid$ nodes denoting the spatial spots and E representing the edges. A is the adjacent matrix in terms of this graph. Here, we applied the dimension reduction of the pseudo-ST and real-ST data by canonical correlation analysis [30–33], and then identified the mutual near neighbors [23] in the space of reduced dimension.

First, with the pseudo-ST data and the real-ST data represented as ${X_{\text{pseudo}}}^{m \times n_p}$ and ${X_{\text{real}}}^{m \times n_r}$, where m is the number of variable genes, and n_p and n_r are the respective number of spots, we projected these two data into a lower S dimension space by

canonical correlation vectors μ_s of n_p dimension and v_S of n_r dimension, where $s = 1, \cdots, S$, to maximize

$$\mu_s{}^T \left(X_{\text{pseudo}}{}^{m\times n_p}\right)^T X_{\text{real}}{}^{m\times n_r} v_s,$$

subjecting to the constraints $\|\mu_s\|_2^2 \le 1$ and $\|v_s\|_2^2 \le 1$. To identify the canonical correlation vector pairs, we used singular value decomposition and got the S canonical correlation vector pairs with the S largest eigenvalues. Each pair of μ_s and v_s projects the original data $X_{\text{pseudo}}{}^{m\times n_p}$ and $X_{\text{real}}{}^{m\times n_r}$ to the sth dimension of the low-dimension space. For DSTG, we took S as 20 for the reduced dimension space.

Second, in the low-dimension space, we identified the mutual nearest neighbors among spots from pseudo-ST and real-ST data. Specifically, if spot i is in one of nearest neighbor of spot j by k-nearest neighbor (KNN, default k is 200), meanwhile spot j is in one of nearest neighbor of spot i by KNN, then spot i and spot j are mutual nearest neighbors. In this way, we built the link graph between the pseudo-ST data and the real-ST data. To further utilize the information of real-ST data in the DSTG model, we also identified the mutual nearest neighbors within the real-ST data itself. To this end, the final link graph is built and represented by the adjacent matrix A. That is, if spot i and the other spot j are mutual nearest neighbors, $A_{ij} = 1$, otherwise $A_{ij} = 0$. This graph captures the intrinsic topological structure of spot similarity between all spots.

DSTG method

We utilized the GCN on the link graph $G = (V, E)$ for the identification and prediction of the compositions of different cell types in the ST data. Each spot is viewed as a node. The cell mixtures in the pseudo-ST data are generated with known compositions. The goal of DSTG is to predict the cell type compositions of the real-ST data by using not only the features of each spot, but also the graph information leveraging the pseudo-ST data and real-ST data, which is characterized as the above adjacent matrix A. Explicitly, the DSTG method takes two inputs. One input is the spot similarity graph structure learned above (see the section Link graph). The other is the data matrix of combined pseudo-ST and real-ST data. As denoted above, with the pseudo-ST data and the real-ST data represented as $X_{\text{pseudo}}{}^{m\times n_p}$ and $X_{\text{real}}{}^{m\times n_r}$, where m is the number of variable genes, whereas n_p and n_r are the corresponding number of spots, the input data matrix is shown as

$$X = \left[X_{\text{pseudo}} \;\; X_{\text{real}}\right] \in R^{m\times N},$$

where $N = n_p + n_r$.

Herein, with these two inputs, i.e. X and A, the DSTG is constructed with multiple convolutional layers. For efficient training of DSTG, the adjacent matrix A is modified and normalized as

$$\tilde{A} = \check{D}^{-1/2}\hat{A}\check{D}^{-1/2},$$

where $\hat{A} = A + I$, I is the identity matrix and $\check{D}$ is the diagonal degree matrix of $\hat{A}$.

Specifically, each graph convolutional layer is defined as

$$H^{(l+1)} = f\left(H^{(l)}, \tilde{A}\right) = \sigma\left(\tilde{A}H^{(l)}W^{(l)}\right) = \text{ReLU}\left(\tilde{A}H^{(l)}W^{(l)}\right),$$

where $H^{(l)}$ is the input from the previous layer, $W^{(l)}$ is the weight matrix of the lth layer, $\sigma(\cdot) = \text{ReLU}(\cdot)$ is the nonlinear activation function and the input layer $H^{(0)} = X$. The composition of a specific cell type f at a pseudo-spot i is represented as $y_{i,f} \in Y_p$, where $i \in \{1, \cdots, n_p\}$ and cell type $f \in \{1, \cdots, F\}$, F represents the total number of different cell types and $Y_p \in R^{n_p\times F}$ represents the known cell compositions at all spots from the pseudo-ST data.

Specifically, for a three-layer DSTG with F distinct cell types, the forward propagation is realized as

$$\hat{Y} = f\left(X, A^H\right) = \text{softmax}\left(\tilde{A}\,\text{ReLU}\left(\tilde{A}X^T W^{(0)}\right)W^{(1)}\right),$$

where $W^{(0)} \in R^{m\times h}$ is the input-to-hidden weight matrix projecting the input data with m variable genes into an h dimension hidden layer, ReLU stands for the rectified linear unit activation function, $W^{(1)} \in R^{h\times F}$ is a hidden-to-output weight matrix and $\hat{Y} = \begin{bmatrix}\hat{Y}_p \\ \hat{Y}_r\end{bmatrix} \in R^{N\times F}$ is composed of two components: $\hat{Y}_p \in R^{n_p\times F}$ represents the predicted proportions of different cell types at pseudo-ST spots and $\hat{Y}_r \in R^{n_r\times F}$ represents the prediction of cell compositions at real-ST spots. The softmax activation function below is used as the activation function in the output layer that learns the cell type proportions,

$$\text{softmax}(\cdot) = \frac{\exp(\cdot)}{\sum \exp(\cdot)}.$$

The evaluation function is defined as the cross-entropy at pseudo-ST spots, i.e.

$$\mathcal{L} = -\sum_{i=1}^{n_p}\sum_{f=1}^{F} y_{i,f}\ln\left(\hat{y}_{i,f}\right),$$

where $y_{i,f} \in Y_p$ and $\hat{y}_{i,f} \in \hat{Y}_p$. The goal of this semi-supervised learning is to minimize the cross-entropy $\mathcal{L}$ between the known cell compositions Y_p and the predicted cell proportions $\hat{Y}_p$. During the propagation of each layer, the model will reduce the cross-entropy error on the training data. After training, we had

$$\hat{Y} = \text{GCN}(X, A) \in R^{N\times F}.$$

Note that $\hat{Y} = \begin{bmatrix}\hat{Y}_p \\ \hat{Y}_r\end{bmatrix}$. Thus, cell compositions of real-ST spots are predicted as $\hat{Y}_r$.

When applying the DSTG model, we randomly split the pseudo-ST data as training (80%), test (10%) and validation set (10%), whereas the real-ST data is unlabeled and will be predicted. For the ST data in this study, we trained three-layer DSTG models for a maximum of 200 epochs using the Adaptive Moment Estimation (Adam) algorithm [34] with a learning rate of 0.01 and early stopping with a window size of 10. For the dimension of the latent layer, we screened options of 32, 64, 128, 256, 528 and 1024 dimensions and selected the optimal one.

For the evaluation metrics, we used the Jensen–Shannon divergence (JSD) score, which is a symmetrized and smoothed version of the Kullback–Leibler divergence. At a spot i, there are discrete probability distribution of composition $P^i = (p_1^i, p_2^i, \cdots, p_C^i)$ as ground truth as well as predicted distribution of composition $Q^i = (q_1^i, q_2^i, \cdots, q_C^i)$, where C indicates the number of cell types. Here, $\sum_{k=1}^{C} p_k^i = 1$ and $\sum_{k=1}^{C} q_k^j = 1$. The JSD score at

spot i is defined as

$$\mathrm{JSD}\left(P^i \| Q^i\right) = \frac{1}{2}\sum_{k\in\{1,\cdots,C\}} p_k^i \log\left(\frac{p_k^i}{\left(p_k^i + q_k^i\right)/2}\right) + \frac{1}{2}\sum_{k\in\{1,\cdots,C\}} q_k^i \log\left(\frac{q_k^i}{\left(p_k^i + q_k^i\right)/2}\right).$$

In this way, the JSD values can be calculated for all spots. Then, the quantiles of JSD values across all spots are used as our evaluation metrics.

Results

Overview of DSTG

Herein, we propose a novel, graph-based AI approach, namely DSTG, to deconvolute ST data through graph-based convolutional networks. The DSTG approach leverages scRNA-seq data to unveil the cell mixtures in the ST data (Figure 1). Our hypothesis is that the captured gene expression on a spot is contributed by a mixture of cells located on that spot. Our strategy is to use the scRNA-seq-derived synthetic ST data, called 'pseudo-ST', to predict cell compositions in real-ST data through semi-supervised learning. First, DSTG constructs the synthetic pseudo-ST data from scRNA-seq data as the learning basis of our method. Then, DSTG learns a link graph of spot mapping across the pseudo-ST data and real-ST data using shared nearest neighbors. The link graph captures the intrinsic topological similarity between spots and incorporates the pseudo-ST and real-ST data into the same graph for learning. Then, based on the link graph, semi-supervised GCN is used to learn a latent representation of both local graph structure and gene expression patterns that can explain the various cell compositions at spots. The major advantages of such similarity-based semi-supervised GCN model are as follows: (1) sensitive and efficient, since for each spot, only the features of similar spots (i.e. neighbor nodes) are used and (2) acquiring generalizable knowledge about the association between gene expression patterns and cell compositions across spots in both pseudo- and real-ST, since the weight parameters in the convolution kernel are shared by all spots. To test the performance of DSTG, we used synthetic pseudo-ST data generated with cell mixtures of known cell type compositions from peripheral blood mononuclear cell (PBMC) scRNA-seq datasets [35], in which DSTG presents superior performance than the SPOTlight method. Furthermore, DSTG is validated and applied to real tissue context from mouse cortex, hippocampus and pancreatic tissues with well-defined structures.

Performance of DSTG on benchmarking data

To evaluate the performance of DSTG, we used the synthetic ST data generated by scRNA-seq cell mixtures as ground truth. Briefly, each spot of this synthetic ST data is constructed by combining the randomly selected two to eight cells from scRNA-seq data. Such synthetic ST data not only mimics real ST data, but also provides ground truth that can be used to evaluate the DSTG's performance in identifying the proportions of different cell types within each synthetic spot. As for the evaluation metrics, we used the JSD, which is a distance metric that measures the similarity between two probability distributions. A smaller value of JSD represents a higher similarity between two distributions, thus signifies a higher accuracy of estimated cell type compositions across spots.

Specifically, we used 13 PBMC scRNA-seq datasets [35] profiled by different protocols, with well-characterized cell populations and discrete cell numbers, to generate benchmarking synthetic spatial data. For each PBMC data, we generated 10 synthetic data and applied both DSTG and SPOTlight to those 10 synthetic data for comparison. Our results show that DSTG achieves lower JSD values (mean JSD = 0.12, Figure 2A), which is significantly lower (P-value $<$2.2e−16) than SPOTlight (mean JSD = 0.24), indicating the higher accuracy of DSTG than SPOTlight across datasets generated from different experiment protocols. Notably, DSTG shows the most accuracy than SPOTlight in the CEL-Seq2 synthetic datasets. Though SPOTlight performs the best on Quartz-Seq2 datasets, DSTG still outperforms SPOTlight with lower JSD value. In addition to PBMC, to examine the performance of DSTG on other different tissues, we included eight other scRNA-seq data from different tissues and protocols to generate the benchmarking synthetic data (Figure 2B). Then, we compared DSTG with SPOTlight based on the synthetic data from these eight additional scRNA-seq data. As shown in Figure 2B, the predicted results of DSTG still outperform SPOTlight using the JSD evaluation metric. Notably, DSTG achieves the mean JSD values of 0.016 and 0.087 for the Smart-Seq2 and single nucleus RNA sequencing (snRNA-Seq) datasets, respectively, which are better than the ones achieved by SPOTlight (0.19 and 0.24). These consistently superior performances of DSTG demonstrate the accuracy and robustness of our method.

To investigate whether DSTG is sensitive to the design of synthetic ST data, we generated discrete synthetic data with different number of spots, library sizes and variable genes, which covers the characteristics of the current and emerging ST data. For the synthetic data with different spot numbers (500–4000) (Figure 2C), we found that DSTG tends to perform better with more spots in the synthetic data, suggesting that the more the spots used, the better the model is trained. Meanwhile, the result suggests that using 1500 spots is sufficient to reach high performance in practice, as the marginal gain of performance is neglectable when using more spots. For the synthetic data with different downsampled library sizes (5000–50 000 reads per cell) (Figure 2D), DSTG shows stable accuracy at lower or higher library sizes. For the synthetic data with different number of variable genes (500–5000) (Figure 2E), DSTG demonstrates stable performance, with optimal performance reaches at 2000 variable genes.

Evaluation of different parameters of DSTG

Regarding the DSTG method, we tested hyperparameters and examined the model performance in ST data deconvolution. First, we tested different number of hidden units and evaluated their impact on DSTG's results. With the synthetic ST data generated from four single-cell protocols (Smart-Seq2, Quartz-Seq2, Chromium and inDrop), we applied DSTG with hidden units ranging from 16 to 2048 and assessed their respective accuracy (Supplementary Figure S1). The results across all cases show that DSTG consistently performs better with 32 units in hidden layer. Interestingly, too, more or too less units show less accuracy of composition prediction.

Second, we tested the impact of different number of layers on the deconvolution performance of DSTG. Here, we still used the synthetic ST data generated from the above four protocols for evaluation. We applied DSTG with 2–10 layers and assessed their respective accuracy. Supplementary Figure S2 shows the performance of DSTG with different number of layers based on the synthetic ST data. For all cases considered here, the

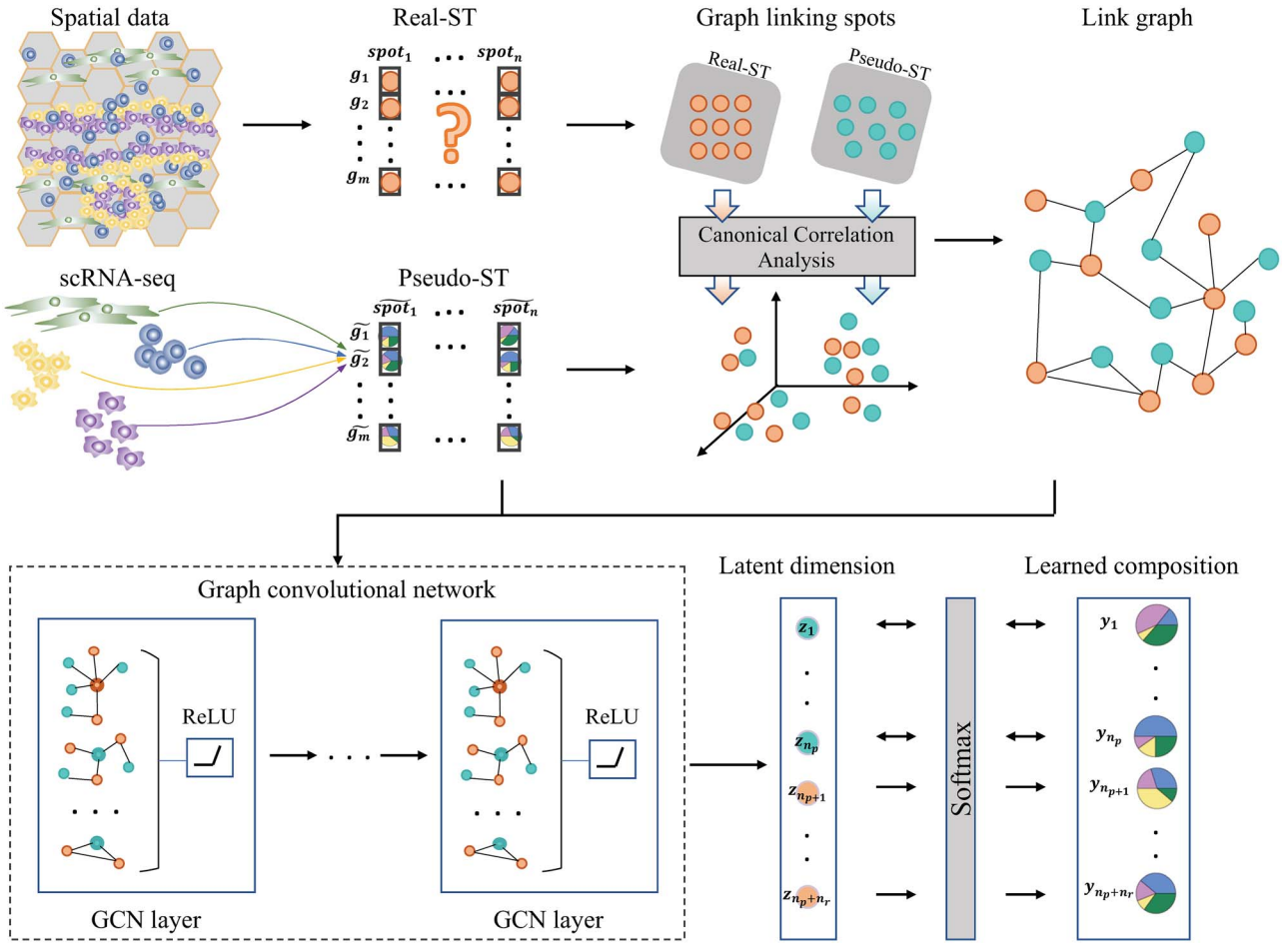

Figure 1. Schematic overview of DSTG for deconvoluting spatial transcriptomics data. Schematic representation of how DSTG deconvolutes spatial transcriptomics (ST) data using scRNA-seq profiles. DSTG first generated the pseudo-ST data with cell mixtures from scRNA-seq data. Between this pseudo- and real-ST data, DSTG identifies a link graph of spot mapping through the canonical correlation analysis. Based on the link graph, graph convolutional network is used to propagate both pseudo-ST and real-ST data into the latent layer and identify the compositions of different cell types for each spot. In this way, cell compositions of real-ST data can be predicted and learned by using pseudo-ST data.

best results are obtained with a three-layer model. Interestingly, DSTG shows less accuracy with two, four and five layers. We also found that for models deeper than six layers, the performance remains at the same level, which is often observed for GCN-based models [36, 37]. The reason is that a deep GCN model causes excessive Laplacian smoothing [38] and a large number of parameters, which lead to less distinguishable representations [39] and overfitting [37].

In addition, we also tested the impact of data preprocessing and normalization on DSTG's performance. Here, we compared ours with other normalization methods including sctransform [40] using regularized negative binomial regression, scran [41] using pooling-based normalization as well as Linnorm [42] using linear model and normality-based normalization. Specifically, synthetic ST data are preprocessed and normalized by different normalization methods. The processed data are then used as input of DSTG for comparison. In all cases (Supplementary Figure S3), DSTG shows robust performance regarding different normalization methods, with no detectible difference in accuracy (JSD value). The comparison results suggest that our preprocessing step is able to reach good performance in practice.

Finally, we investigated the impact of different graph construction approaches on the deconvolution performance of DSTG. Here, we applied four kinds of graphs including DSTG's link graph, KNN-based graph, identity graph and random graph to compare their respective accuracy on synthetic ST data. With the results (Supplementary Figure S4), we found that DSTG shows consistently higher accuracy across all synthetic ST data when using the link graph. In contrast, DSTG shows much lower accuracy as indicated by higher JSD values, when using the other three kinds of graphs, especially the identity and random graphs. These results suggest that the graph construction is critical for DSTG's performance, which also demonstrate that the DSTG's link graph is essential for accurate deconvolution.

Spatial decomposition of mouse cortex layer

To examine whether DSTG can reveal microanatomical structures in complex tissue, we used the 10X Visium ST data of cerebral cortex layer in mouse brain. This cortex layer has well-defined cytoarchitecture and thus is suitable to evaluate DSTG's performance. To deconvolute this ST data by DSTG, we used the scRNA-seq dataset profiled by the Smart-Seq2

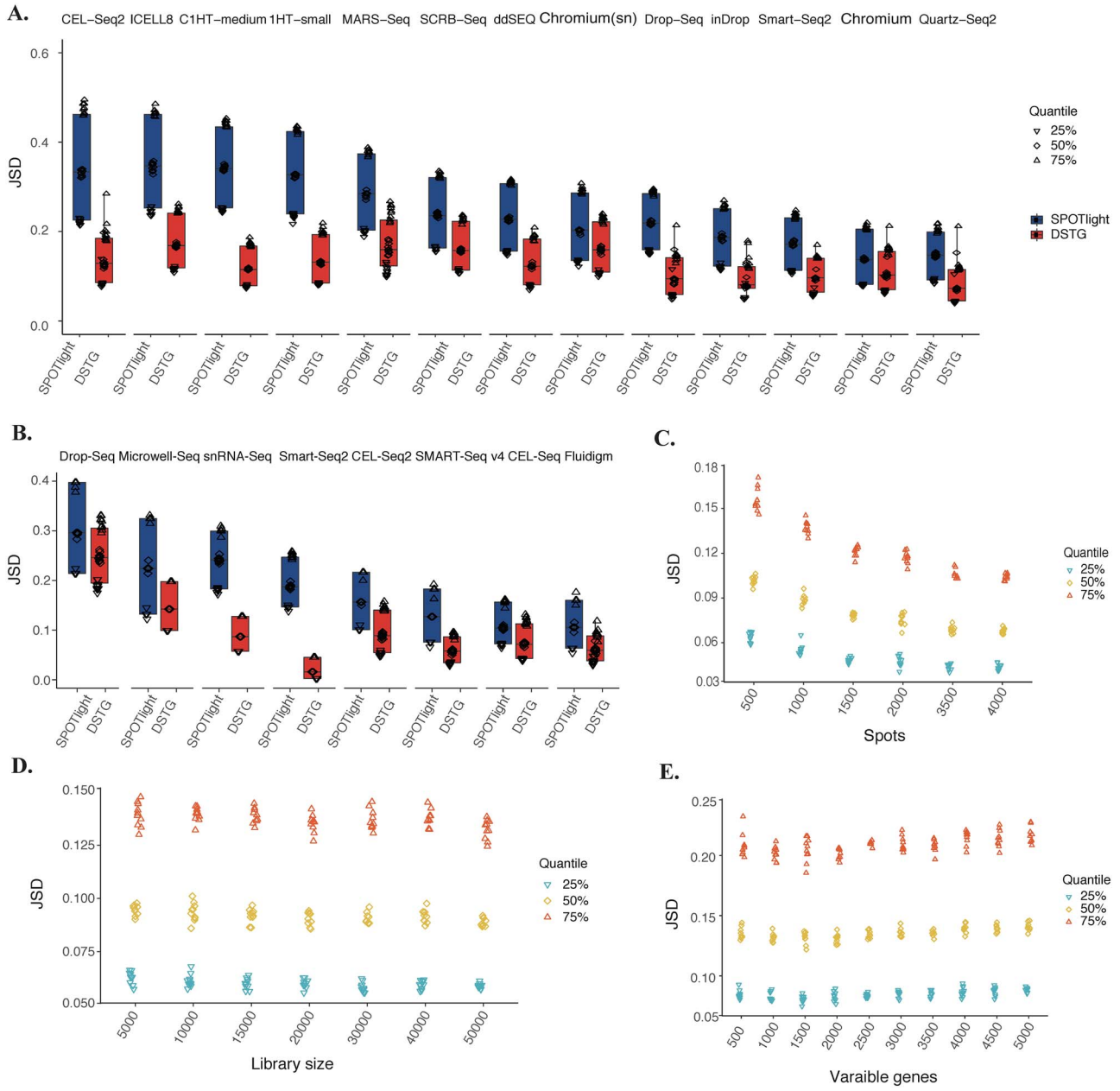

Figure 2. Performance of DSTG on benchmarking datasets. **(A)** Performance of DSTG is assessed and compared with SPOTlight by synthetic spatial data generated from 13 PBMC datasets of different scRNA-seq protocols. **(B)** Performance of DSTG is further benchmarked with SPOTlight by synthetic spatial transcriptomics data generated from the other scRNA-seq datasets from different tissues and protocols. **(C)** DSTG's performance on synthetic data with different number of spots. **(D)** DSTG's performance on synthetic data with different library depths. **(E)** DSTG's performance on synthetic data with different number of variable genes. In **(A–E)**, the y-axis represents the JSD value.

protocol from the Allen Institute, which consists of ~14 000 adult mouse cortical cell taxonomy and 22 cell types (Figure 3A). With this scRNA-seq data, the spatial deconvolution of the ST data by DSTG accurately reconstructs the architecture of brain cortex layer (Figure 3B). The identified heterogenous cell proportions of each localized spot are shown by the pie chart at the respective spot, which are confirmed by their existence in cortical areas, suggesting the high accuracy and sensitivity of the DSTG's predictions. In the spots visualized in Figure 3B, about 46% are with mixed cell types. To clearly illustrate the cell mixtures within each spot, we present two separated figures of the same data in Supplementary Figure S5. Specifically, Supplementary Figure S5A shows the identified spots with unique cell types, whereas Supplementary Figure S5B shows the identified spots with mixed cell types.

Moreover, our predicted compositions provide more detailed information about the heterogeneity of this area. Specific investigation shows the regional enrichment of each cell type based on their identified proportions. Illustrative examples are the differentially enriched neuronal subtypes including cortical layer 2/3 (L2/3), cortical layer 6 (L6b) and oligodendrocytes (oligo) (Figure 3C). The subpopulation of L2/3 is shown with high compositions in the outside liner of spots within the cortex. Spots with most L6b cells are shown with high proportions

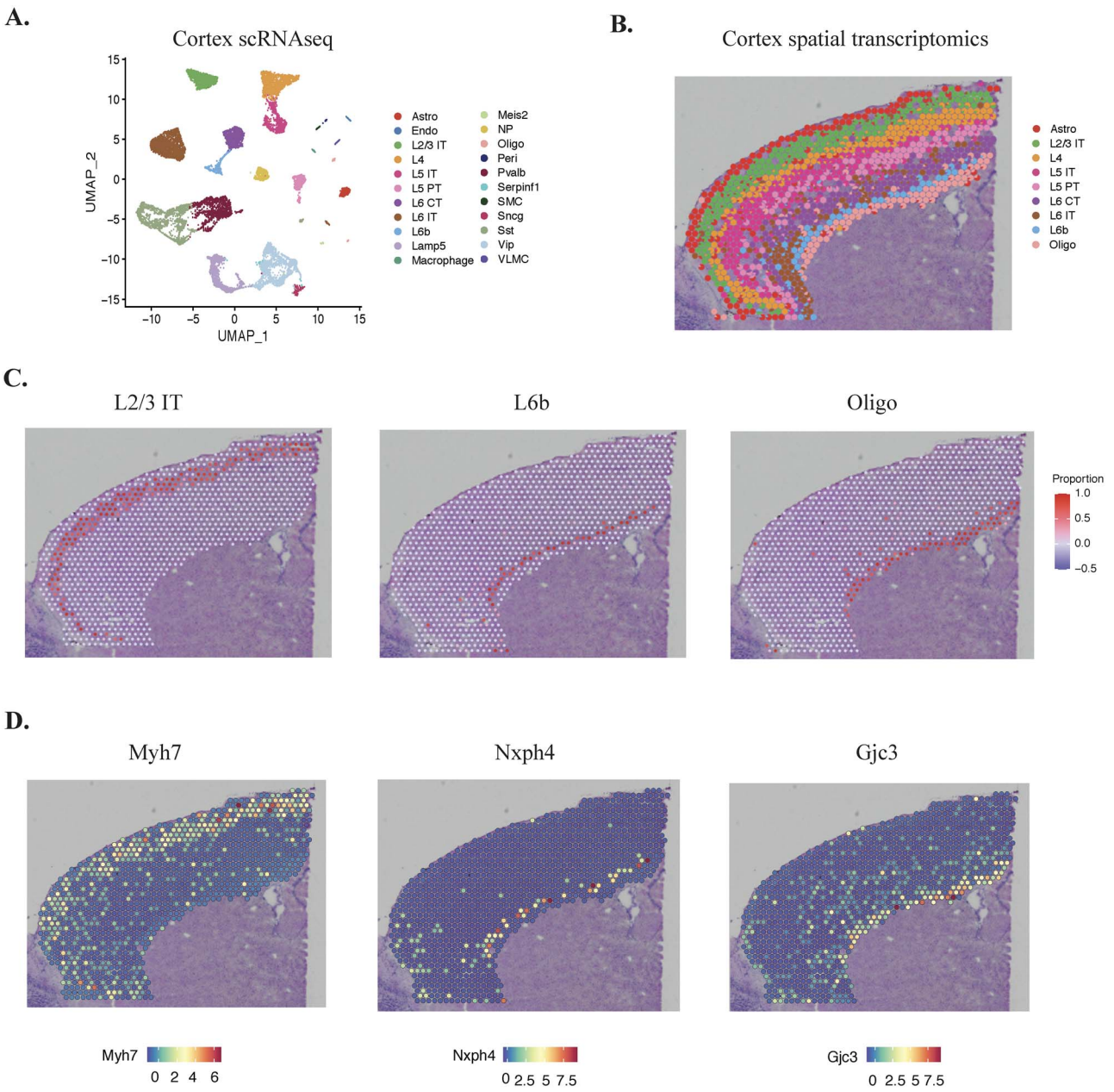

Figure 3. Spot deconvolution of mouse cortex layer. (**A**) UMAP projections of single-cell RNA-seq data from mouse cortex tissue. Different cell types are labeled and colored according to known cell annotations. (**B**) Spatial plot with pie chart shows the predicted cell compositions of each spot within the cortex layer. (**C**) Spatial plot shows the proportions of specific neuron subtypes in the spots within the captured region. Red spot indicates high proportion of the respective cell type. (**D**) Visualization of the spatial expressions of cell type specific markers in ST data. Red color indicates the high gene expression in that spot.

in the inner liner of the cortex. Towards the innermost layer, oligodendrocytes cells are mainly abundant in these spots. These data are consistent with the layered cytoarchitecture of the cortex tissue. The ability to identify the distinct spatial cellular compositions of each spot in the cortical neuronal layers indicates the accuracy and sensitivity of DSTG.

To examine whether the cell type specific genes are enriched in their corresponding spatial locations, we investigated the distribution of marker genes known to be specific to the respective cell types (Figure 3D). For example, the top expressed gene marker of L2/3 cells, Myh7, is detected with high expression in the ST data of L2/3 dominated spots, which is in line with the predicted proportions of this cell type. The top markers of L6b and Oligo, Nxph4 and Gjc3, also show high expression in their corresponding spots, respectively. Meanwhile, these genes are undetectable or detected at very low levels at other spatial spots. It is worth noting that the partial expression of cell type markers in a specific ST spot may reflect the heterogeneous composition of cell types in that spot. Interrogation of these differential genes further confirms the accurate predicted cell proportions within spots in the tissue section.

To investigate whether the pseudo-ST generated from scRNA-seq data presents similar spatial patterns with the real-ST data, we looked into the pseudo-spots from pseudo-ST data that are linked with real-spots from real-ST data in our link graph. After mapping these pseudo spots onto the spatial

contexture, we found that the compositions within these pseudo-spots show similar patterns (Supplementary Figure S5C) with the real-ST structures (Supplementary Figure S5D). That is, these pseudo-spots from pseudo-ST data recapitulate the architecture of mouse brain cortex layer. The heterogenous cell compositions within each pseudo-spot are shown as pie chart at their respective spot. Specific investigation shows that the Astro subtype mainly locates at the outside liner of the cortex structure, whereas the L2/3 cells are enriched next to the Astro cells, which is consistent with the spatial patterns of real-ST data. Pseudo-spots containing L6b cells are shown with high proportions in the inner liner of the cortex tissue. Oligodendrocyte cells within pseudo-spots remain abundant at the innermost layer. These data show that the pseudo-ST data presents coherent spatial patterns with real-ST data as illustrated in Figure 3B, which is consistent with the layered cytoarchitecture of the cortex tissue. These results also verify the accuracy and reliability of using pseudo-ST data in DSTG.

Mapping distinct cell populations of mouse hippocampus

The fast advance of ST technologies raises new challenges in deconvoluting the transcriptomics data at each spot: spots become smaller; meanwhile the total number of spots grows exponentially, but the sequencing depth at each spot becomes much lower. We demonstrated the performance of DSTG on such emerging ST data, using the recently available Slide-seq v2 [15] of mouse hippocampus tissues as an example. Comparing with the 10X Genomics' Visium platform, the bead size of the Slide-seq v2 platform is 5.5-fold smaller, thus the spatial resolution is 25–100-fold higher. Consequently, the typical median library size per bead is 550 UMIs, 100-fold lower than 10X Genomics Visium.

To deconvolute this ST data, we used the existing scRNA-seq dataset from mouse hippocampus [43], which consists of 52 846 cells with 19 cell types that are profiled by the Drop-seq protocol (Figure 4A). Based on this scRNA-seq data, DSTG's spatial decomposition of the ST data accurately identifies different cell types within the hippocampus slice (Figure 4B). The spatially localized pie charts represent the identified different cell proportions in the slice. Moreover, our predicted compositions provide more detailed information about the heterogeneity of this area. Closer investigation confirms the regional enrichment of specific cell types with their identified composition (Figure 4C). For example, oligodendrocyte cells are identified with high compositions in the middle wide strips within the hippocampus slice. In the Cornu Ammonis (CA) subfield, CA3 principal cells scatter within the slice with low proportions, but are majorly abundant in the half strip at the right of the slice. Ependymal cells present mainly at an irregular circle and the other band at the top right of the slice, of which data are consistent with the spatial structures within the mouse hippocampus.

To further evaluate our method, we selected the cell type-specific genes from scRNA-seq data and assessed their expression in the ST data. As expected, the top differentially expressed gene marker in oligodendrocyte (Figure 4D), Proteolipid Protein 1 (PLP1), expresses strictly in line with the oligodendrocyte region based on the ST data. Coiled-Coil Domain Containing 153 (CCDC153) is the top expressed gene marker in CA3 principal cells that is also detected with high expression in the ST data of CA3 principal cells. Another example is neuronal pentraxin receptor (NPTXR), the marker of ependymal, is enriched at regions with a high abundance of ependymal cells. These cell-specific genes detected in their corresponding locations further underlines the accuracy of DSTG. In summary, DSTG demonstrates accurate and reliable deconvolution capabilities on ST data generated from the latest ST platform such as the Slide-seqV2, which has much smaller spot size, much larger number of spots and much lower sequencing depth.

Deconvolution of pancreatic cancer tissue sections

Tissues in diseases such as tumors exhibit unique pathological cytoarchitectures. To further demonstrate and test the DSTG's performance in such conditions, we applied it to two ST data obtained from two tumor sections of pancreatic ductal adenocarcinoma (PDAC), i.e. PDAC-A and PDAC-B. Sample-matched scRNA-seq data (Supplementary Figure S6) generated by inDrop protocol is used to generate the pseudo-ST data, which shares similar characteristics with real-ST data (Supplementary Figure S7).

For the PDAC-A sample (Figure 5A), after DSTG's deconvolution, we observed discrete regional enrichment of cancer clones and noncancer cells. Specifically, cells of cancer clone S100 Calcium Binding Protein A4 (S100A4) and Transmembrane 4 L Six Family Member 1 (TM4SF1) are mainly identified mixed in the spots of cancerous region, which are excluded from the spots of ductal cells including the centroacinar ductal cells and the co-localized antigen-presenting ductal cells. Stroma cells are involved between the ductal cells and cancer cells, which are consistent with previous results annotated by hematoxylin and eosin (H&E) staining and brightfield imaging [13]. We also found a few proportions of hypoxic ductal cells in the spots close to the cancerous region, indicating the low oxygen environment in tumor. Further inspections of specific cell types (Figure 5B), including cancer clone cells and ductal cells, confirm their regional proportions on their identified structures. The point size and related color indicate different compositions in the spatial spots.

In the other PDAC-B sample, as shown in Figure 5C, cells of cancer clone TM4SF1 rather than cancer clone S100A4 are identified. These cancer clone TM4SF1 cells are localized preferentially in the spots of the bottom right region, distinguished from the interstitium and ductal cells. We noticed that most interstitium cells are adjacent to the cancer clone TM4SF1 cells, and some interstitium cells co-localize with the cancer cells. Ductal cells are mainly abundant in the spatial spots of the top region. These findings highlight the precise consistency with previous H&E staining results [13]. Further inspections of cancer clones and ductal cells (Figure 5D) confirm their regional compositions on their known locations. These results of DSTG are consistent with independent histological annotations, supporting its ability to identify accurate cellular compositions from the ST data of tumor tissues.

In addition, we applied SPOTlight to the real-ST data of PDAC-A and compared its results with DSTG. After deconvolution, we observed similar spatial contexture as well as different cell compositions between SPOTlight's and DSTG's results (Supplementary Figure S8A). Specifically, SPOTlight and DSTG identify different proportions of centroacinar ductal cells (Supplementary Figure S8B). Both methods discern abundant centroacinar ductal cells in the left region, which are distinguished from stroma and cancer cells. However, centroacinar ductal cells identified by DSTG are more consistent with the expression of its marker Claudin 2 (CLDN2), with higher Pearson correlation (cor = 0.739) than DSTG (cor = 0.524). We also looked into the proportions of stroma cells identified by SPOTlight and DSTG, respectively (Supplementary Figure S8C).

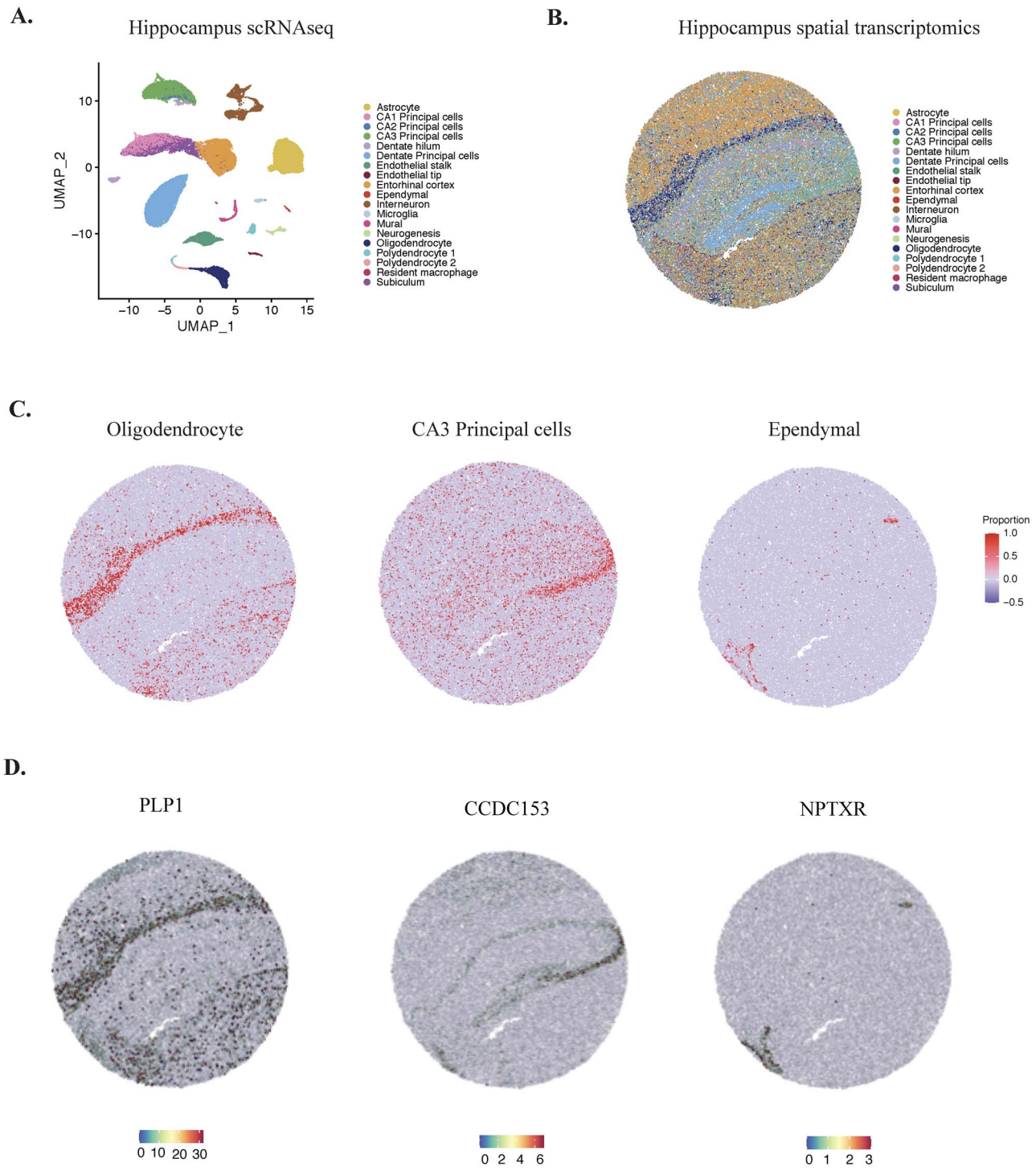

Figure 4. Spatial chart of mouse hippocampus tissue using DSTG. (**A**) Uniform manifold approximation and projection (UMAP) projections of single-cell RNA-seq data from mouse hippocampus tissue. Different cell types are labeled and colored based on known cell annotations. (**B**) Spatial plot with pie chart shows the predicted cell compositions within the captured locations in the mouse hippocampus. (**C**) Spatial plot presents the proportions of specific neuron subtypes within the captured location. Red color indicates high abundance of certain cell type in this region. (**D**) Spatial expression of cell type specific markers of the respective neuron subtypes in the ST data.

The stroma cells revealed by SPOTlight localize more prevalently within the tissue with less proportions. Moreover, the stroma compositions identified by SPOTlight are less associated with its maker Collagen alpha 1 chain type I (COL1A1) expression (cor = 0.457), whereas DSTG identified stroma proportions are far more correlated with COL1A1 expression (cor = 0.644). In addition to the ductal and stroma cells, we also compared the compositions of cancer clone S100A4 cells identified by two methods (Supplementary Figure S8D), of which the expression of Glyceraldehyde-3-Phosphate Dehydrogenase (GAPDH) maker

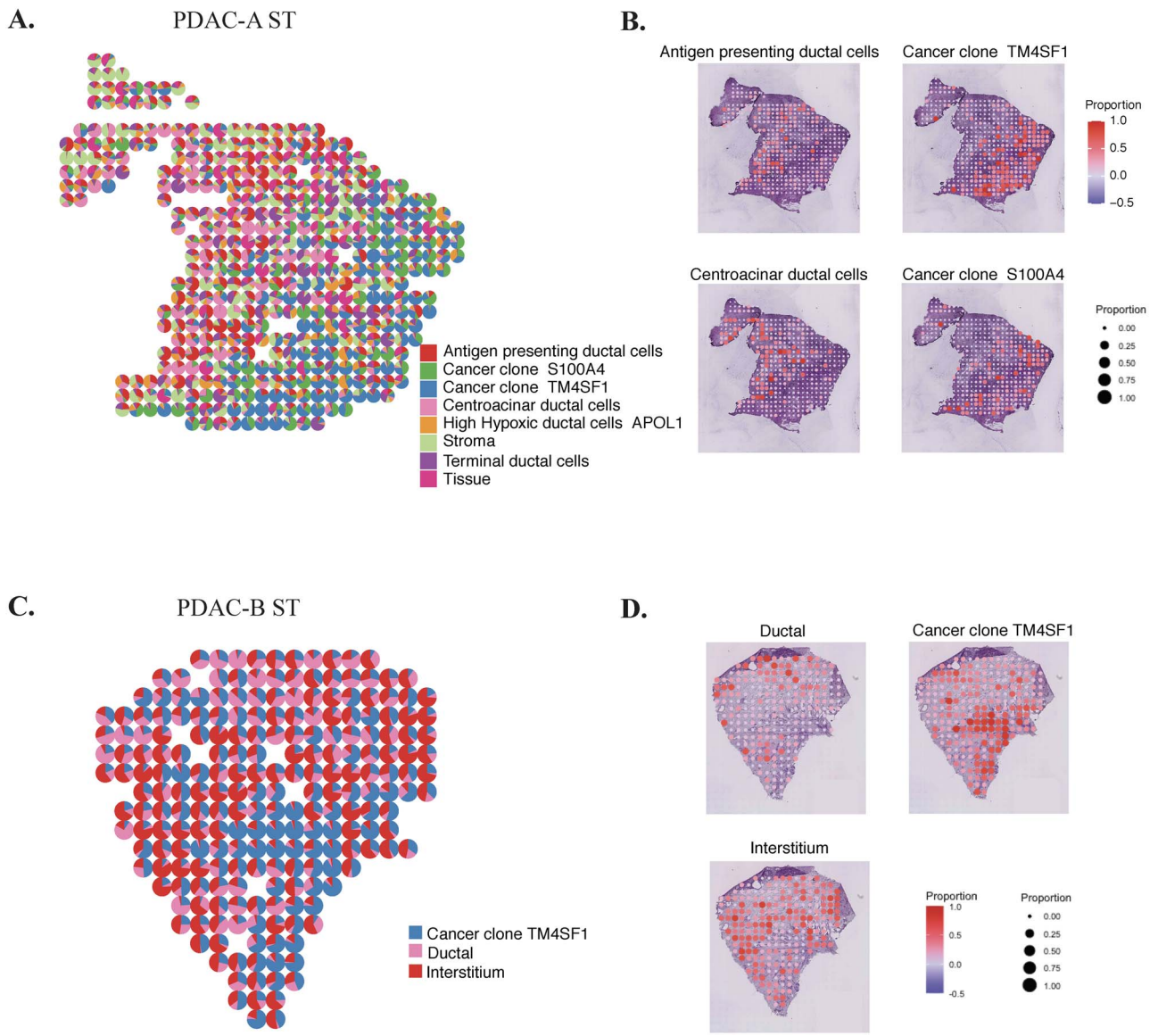

Figure 5. Mapping spatial spots across pancreatic cancer tissue. (**A**) Spatial plot shows the predicted compositions of different cell types within the captured spots of PDAC-A tissue slice. (**B**) Predicted proportions of different cell types including antigen-presenting ductal cells, centroacinar ductal cells, cancer clone TMSF1 and S100A4. (**C**) Spatial plot presents the predicted compositions of different cell types within captured spots of PDAC-B tissue slice. (**D**) Predicted proportions of different cell types including ductal cells, cancer clone TMSF1 and interstitium. Red color indicates high proportion of certain cell type.

exhibits stronger association with DSTG's prediction (cor = 0.783) than that of SPOTlight (cor = 0.511). Thus, DSTG shows superior performance than SPOTlight in accurately decomposing tissue architectures.

In conclusion, DSTG is able to detect the unique cytoarchitectures in disease tissues, distinguish the spatial distribution of tumor cells evolved from different clones and characterize cancer-specific cellular phenomena such as local hypoxia as well as antigen presenting in the tumor cell dominating regions.

Discussion

ST provides unprecedented opportunities to study tissue heterogeneity and cell spatial organization [44–46]. However, the resolution of ST is less than the single-cell level. As single spot in ST data may cover heterogenous cell types, our DSTG method aims to determine the proportions of different cell types and states across spots where genes are reliably identified. In this study, we present the DSTG method for performing cell type deconvolution in ST data using the GCN. DSTG is evaluated by benchmarking synthetic data generated from PBMC and other tissues, in which DSTG demonstrates excellent accuracy between the predicted cell mixtures and the actual cell composition. DSTG is also shown to achieve high consistency with H&E staining observations on ST data from complex tissues including mouse cortex, hippocampus and human pancreatic tumor slices.

For the pseudo-ST data used in DSTG, we tested that pseudo-ST shares similar characteristics with real-ST data through investigating their respective distributions of gene expression using the PDAC tissue sample (Supplementary Figure S7A). Through Kolmogorov–Smirnov test, we found that the distribution of DSTG's pseudo-ST data is not significantly differed from the distribution of real-ST data (P-value = 0.419). In contrast, the distribution of sample-matched scRNA-seq data is

significantly differed from that of real-ST data (P-value = 4.89e−14, Supplementary Figure S7B). This observation suggests that pseudo-ST data shares similar characteristics with real-ST data, even more than its matched scRNA-seq data. Moreover, we looked into the Euclidian distance between the pseudo-ST and real-ST data (Supplementary Figure S7C). Comparing with the distances between scRNA-seq data and real-ST data (red histogram), the distances between pseudo-ST and real-ST data (blue histogram) are significantly closer with P-value <2.2e-16, again showing that pseudo-ST is more similar with the real-ST data than its matched scRNA-seq data. These results confirm that the pseudo-ST data used in DSTG shares similar characteristics with the real-ST data.

As SPOTlight is also used to deconvolute the ST data, we compared DSTG with SPOTlight and found DSTG consistently outperforms SPOTlight on benchmarking synthetic data. From a technical perspective, DSTG provides some major advantages. First, DSTG simultaneously utilizes variable genes and graphical structures through a nonlinear propagation in each layer, which is appropriate for learning the cellular composition due to the heteroskedastic and discrete nature of ST data. Second, DSTG identifies the respective weights of different cell types in the pseudo-ST data generated from scRNA-seq data, which can be effectively leveraged to learn the cell compositions in the real-ST data. Third, as the sequencing depth of spatial data is expected to increase, DSTG has been shown to perform better to interrogate the cell distribution quantitatively in such ST data.

In addition to the successful results, there are several aspects that DSTG can be improved. First, as an AI model, DSTG shows not only the merits of its kind, but also some limitations including the black-box nature of AI models [47–49], which can be addressed through downstream analysis that can ameliorate some of the problems and bring insights into the learned cellular compositions. Second, as a graph model, improving the built graph can further boost the model performance. Our link graph based on mutual nearest neighbors best captures the spots' similarity in spatial data, reflecting the effective graph representation. As a fast-growing research field, new approaches of building graph are emerging, of which we will test and adapt in future versions of DSTG.

Our DSTG method paves the way for inferring functional relationships between heterogenous cell subpopulations based on their composition and co-localization in the tissue spots. This includes intercellular communication across neighboring spots, which opens up future possibilities of studying the complete interactome in a spatially resolved manner. Moreover, as the precise composition of tissue may vary from one individual patient to the other, the spatial composition of cellular subpopulations can be of prognostic value for patients in the future. We anticipate that the spatial deconvolution using DSTG will contribute to future patient prognosis and pathological assessments. Overall, DSTG demonstrates as a robust and accurate tool to determine cell type locations and precise compositions of spatial spots, which provides an unbiased perspective and investigation into the spatial organization of distinct cellular populations in tissue.

Data availability

All scRNA-seq datasets are downloaded from their public accessions. The first benchmarking PBMC scRNA-seq datasets [35] in Figure 2A are generated by 13 different protocols, including C1HT-medium, ICELL8, MARS-Seq, Chromium (sn), CEL-Seq2, gmcSCRB-Seq, C1HT-small, inDrop, Drop-Seq, ddSEQ, Smart-Seq2, Chromium and Quartz-Seq2. These 13 PBMC datasets are publicly available through the Gene Expression Omnibus (GEO) (GSE133549). Cell types with too small sizes are ignored to avoid misleading information. To evaluate the impact of synthetic data (Figure 2C–E), we used the Smart-Seq2 PBMC dataset to generate discrete synthetic data with different number of spots and variable genes, as well as different sequencing depths.

For the second benchmarking in Figure 2B, the Drop-Seq data is downloaded from the Short Read Archive under accession number SRP073767 [2]. The Microwell-Seq data is profiled using the Microwell-Seq protocol [50] that can be downloaded from the Mouse Cell Atlas. The snRNA-Seq data is profiled from the entorhinal cortex from human brains of Alzheimer's disease, yielding a total of 13 214 high-quality nuclei [51] using the single-nucleus RNA-seq protocol, which can be downloaded from GEO (accession number: GSE138852). The Smart-Seq2 data is profiled using the Smart-Seq2 platform [52], which is profiled from melanoma tumor and downloaded from GEO (accession number: GSE72056). The CEL-Seq2 data is obtained from human cadaveric pancreata using the CEL-Seq2 protocol (accession number: GSE85241) that consists of 2122 cells and 18 915 genes [53]. The SMART-Seq v4 data is downloaded from the database of Genotypes and Phenotypes (dbGAP) (accession number: phs001790) [54], which is generated using SMART-Seq v4 platform. This dataset contains 16 024 genes and 14 055 cells, from 34 cell types in the middle temporal gyrus of human cerebral cortex. The CEL-Seq data is obtained from three human lung adenocarcinoma cell lines using the CEL-Seq platform that consists of 570 cells and 12 627 genes [55], which can be downloaded from GEO (accession number: GSE117617). The Fluidigm data is profiled using Fluidigm C1 platform with 11 778 cells and 3803 genes [56]. We downloaded this data from GEO (accession number: GSE81608).

For the applications of DSTG in real-ST data, we selected the well-annotated scRNA-seq data of the same tissue type as ground truth and generate synthetic pseudo-ST data to train the model. We listed the cell types with sufficient cell numbers of each scRNA-seq data in Supplementary tables. Specifically, to deconvolute the mouse cortex ST, we used the scRNA-seq dataset profiled by the Smart-Seq2 protocol from the Allen Institute, which consists of ∼14 000 adult mouse cortical cell taxonomy and 22 cell types to generate the pseudo-ST data as training data (Supplementary Table S1). To deconvolute the ST data of mouse hippocampus tissue, we used the existing scRNA-seq dataset from mouse hippocampus [43], which consists of 52 846 cells with 19 cell types that are profiled by the Drop-seq protocol (Supplementary Table S2). Regarding the deconvolution of ST data in pancreatic tissues, we used their matched scRNA-seq data with adequate cells profiled by inDrop protocol (Supplementary Tables S3 and S4).

Code availability

All the functions mentioned above were implemented as a Python software, which can be downloaded at https://github.com/Su-informatics-lab/DSTG.

Key Points

- We have developed a novel semi-supervised GCN model, named DSTG, to accurately deconvolute the observed gene expressions at each spot of ST data and recover its cell constitutions.

- DSTG demonstrates high accuracy and robust performance across experimental protocols, ST platforms and tissues from different organs.
- DSTG is available as a ready-to-use open source software for precise interrogation of spatial organizations and functions in tissues.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

Acknowledgments

The authors acknowledge the DEMON high-performance computing (HPC) cluster, the Texas Advanced Computing Center at The University of Texas at Austin (http://www.tacc.utexas.edu), and the Extreme Science and Engineering Discovery Environment (which is supported by National Science Foundation grant number ACI-1548562) for providing HPC resources that have contributed to the research results reported within this paper.

Funding

This work was supported by the Indiana University Precision Health Initiative to J.S.

Authors' Contributions

Q.S. and J.S. developed the structure and arguments for the paper and wrote the manuscript. All the authors reviewed and approved the final manuscript.

References

1. Song Q, Su J, Miller LD, *et al*. scLM: automatic detection of consensus gene clusters across multiple single-cell datasets. *bioRxiv* 2020. doi: 10.1101/055822.
2. Zheng GXY, Terry JM, Belgrader P, *et al*. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 2017;**8**(1):14049.
3. Schaum N, Karkanias J, Neff NF, *et al*. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* 2018;**562**(7727):367–72.
4. Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat Rev Immunol* 2018;**18**(1):35–45.
5. Song Q, Hawkins GA, Wudel L, *et al*. Dissecting intratumoral myeloid cell plasticity by single cell RNA-seq. *Cancer Med* 2019;**8**(6):3072–85.
6. Bendall SC, Davis KL, Amir el AD, *et al*. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* 2014;**157**(3):714–25.
7. Trapnell C, Cacchiarelli D, Grimsby J, *et al*. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 2014;**32**(4):381–6.
8. Ståhl PL, Salmén F, Vickovic S, *et al*. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 2016;**353**(6294):78–82.
9. Lee JH, Daugharthy ER, Scheiman J, *et al*. Highly multiplexed subcellular RNA sequencing in situ. *Science* 2014;**343**(6177):1360–3.
10. Chen KH, Boettiger AN, Moffitt JR, *et al*. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 2015;**348**(6233):aaa6090.
11. Asp M, Giacomello S, Larsson L, *et al*. A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. *Cell* 2019;**179**(7):1647–60.e19.
12. Maynard KR, Collado-Torres L, Weber LM, *et al*. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *bioRxiv* 2020. doi: 10.1101/969931.
13. Moncada R, Barkley D, Wagner F, *et al*. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat Biotechnol* 2020;**38**(3):333–42.
14. Maniatis S, Äijö T, Vickovic S, *et al*. Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science* 2019;**364**(6435):89–93.
15. Stickels RR, Murray E, Kumar P, *et al*. Sensitive spatial genome wide expression profiling at cellular resolution. *bioRxiv* 2020. doi: 10.1101/989806.
16. Rodriques SG, Stickels RR, Goeva A, *et al*. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* 2019;**363**(6434):1463–7.
17. Saiselet M, Rodrigues-Vitória J, Craciun L, *et al*. Transcriptional output, cell types densities and normalization in spatial transcriptomics. *bioRxiv* 2018;503870. doi: https://doi.org/10.1101/503870.
18. Elosua M, Nieto P, Mereu E, *et al*. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *bioRxiv* 2020. doi: 10.1101/131334.
19. Kipf T, Welling M. Semi-supervised classification with graph convolutional networks. In: *ICLR*. Toulon, France: International Conference on Learning Representations (ICLR), 2017.
20. Veličković P, Cucurull G, Casanova A, *et al*. Graph attention networks. *arXiv* 2017; preprint arXiv:1710.10903.
21. Defferrard M, Bresson X, Vandergheynst P (eds). Convolutional neural networks on graphs with fast localized spectral filtering. In: *Advances in Neural Information Processing Systems*. Barcelona, Spain, 2016.
22. Fang C, Xu D, Su J, *et al*. DeePaN: a deep patient graph convolutional network integrating clinico-genomic evidence to stratify lung cancers benefiting from immunotherapy. *medRxiv* 2020;19011437. doi: https://doi.org/10.1101/19011437.
23. Haghverdi L, Lun ATL, Morgan MD, *et al*. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* 2018;**36**(5):421–7.
24. Song Y, Zheng S, Niu Z, *et al*. (eds). Communicative representation learning on attributed molecular graphs. In: *IJCAI*, 2020; pp. 2831–38.
25. Zhao T, Hu Y, Valsdottir LR, *et al*. Identifying drug–target interactions based on graph convolutional network and deep neural network. *Brief Bioinform* 2020; bbaa044.
26. Zeng Y, Zhou X, Rao J, *et al*. Accurately clustering single-cell RNA-seq data by capturing structural relations between cells through graph convolutional network. *bioRxiv* 2020. doi: 10.1101/278804.
27. Yuan Y, Bar-Joseph Z. GCNG: graph convolutional networks for inferring cell-cell interactions. *bioRxiv* 2019. doi: 10.1101/887133.

28. Li J, Zhang S, Liu T, *et al*. Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction. *Bioinformatics* 2020;**36**(8):2538–46.
29. Song Q, Su J, Zhang W. scGCN: a graph convolutional networks algorithm for knowledge transfer in single cell omics. *bioRxiv* 2020. doi: 10.1101/295535.
30. Kettenring JR. Canonical analysis of several sets of variables. *Biometrika* 1971;**58**(3):433–51.
31. Nielsen AA. Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE Trans Image Process* 2002;**11**(3):293–305.
32. Hardoon DR, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: an overview with application to learning methods. *Neural Comput* 2004;**16**(12):2639–64.
33. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 2009;**10**(3):515–34.
34. Kingma DP, Ba JA. A method for stochastic optimization. *arXiv 2014* 2019;**434**: preprint arXiv:1412.6980.
35. Mereu E, Lafzi A, Moutinho C, *et al*. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat Biotechnol* 2020;**38**(6):747–55.
36. Hamilton W, Ying Z, Leskovec J (eds). Inductive representation learning on large graphs. In: *Advances in Neural Information Processing Systems*. Long Beach, CA, USA, 2017.
37. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *arXiv* 2016; preprint arXiv: 1609.02907.
38. Taubin G. A signal processing approach to fair surface design. In: *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques: Association for Computing Machinery*. Los Angeles, CA, USA, 1995. p. 351–8.
39. Li Q, Han Z, Wu X-M. Deeper insights into graph convolutional networks for semi-supervised. *Learning* 2018. arXiv:1801.07606.
40. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* 2019;**20**(1):296.
41. Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* 2016;**17**(1):75.
42. Yip SH, Wang P, Kocher JA, *et al*. Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Res* 2017;**45**(22):e179.
43. Saunders A, Macosko EZ, Wysoker A, *et al*. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* 2018;**174**(4):1015–30.e16.
44. Giacomello S, Salmén F, Terebieniec BK, *et al*. Spatially resolved transcriptome profiling in model plant species. *Nature Plants* 2017;**3**(6):17061.
45. Berglund E, Maaskola J, Schultz N, *et al*. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat Commun* 2018;**9**(1): 1–13.
46. Thrane K, Eriksson H, Maaskola J, *et al*. Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. *Cancer Res* 2018;**78**(20):5970–9.
47. Ghosh A, Kandasamy D. Interpretable artificial intelligence: why and when. *AJR Am J Roentgenol* 2020;**214**(5): 1137–8.
48. Moore JH, Boland MR, Camara PG, *et al*. Preparing next-generation scientists for biomedical big data: artificial intelligence approaches. *Per Med* 2019;**16**(3): 247–57.
49. Filipp FV. Opportunities for artificial intelligence in advancing precision medicine. *Curr Genet Med Rep* 2019;**7**(4): 208–13.
50. Han X, Wang R, Zhou Y, *et al*. Mapping the mouse cell atlas by Microwell-seq. *Cell* 2018;**172**(5):1091–107.e17.
51. Grubman A, Chew G, Ouyang JF, *et al*. A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat Neurosci* 2019;**22**(12):2087–97.
52. Tirosh I, Izar B, Prakadan SM, *et al*. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 2016;**352**(6282):189–96.
53. Muraro MJ, Dharmadhikari G, Grun D, *et al*. A single-cell transcriptome atlas of the human pancreas. *Cell Syst* 2016;**3**(4):385–94.e3.
54. Hodge RD, Bakken TE, Miller JA, *et al*. Conserved cell types with divergent features in human versus mouse cortex. *Nature* 2019;**573**(7772):61–8.
55. Tian L, Su S, Dong X, *et al*. scPipe: a flexible R/Bioconductor preprocessing pipeline for single-cell RNA-sequencing data. *PLoS Comput Biol* 2018;**14**(8):e1006361.
56. Xin Y, Kim J, Okamoto H, *et al*. RNA sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metab* 2016;**24**(4):608–15.