*Nucleic Acids Research, 2022* **1**
*https://doi.org/10.1093/nar/gkac084*

# CellDART: cell type inference by domain adaptation of single-cell and spatial transcriptomic data

**Sungwoo Bae[1,2,†], Kwon Joong Na[3,4,†], Jaemoon Koh[5], Dong Soo Lee[1,2,6], Hongyoon Choi[2,6,*] and Young Tae Kim[3,4,*]**

[1]Department of Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, Republic of Korea, [2]Department of Nuclear Medicine, Seoul National University Hospital, Seoul, Republic of Korea, [3]Department of Thoracic and Cardiovascular Surgery, Seoul National University Hospital, Seoul, Republic of Korea, [4]Seoul National University Cancer Research Institute, Seoul National University College of Medicine, Seoul, Republic of Korea, [5]Department of Pathology, Seoul National University Hospital, Seoul, Republic of Korea and [6]Department of Nuclear Medicine, Seoul National University College of Medicine, Republic of Korea

Received June 19, 2021; Revised January 06, 2022; Editorial Decision January 24, 2022; Accepted January 26, 2022

## ABSTRACT

**Deciphering the cellular composition in genome-wide spatially resolved transcriptomic data is a critical task to clarify the spatial context of cells in a tissue. In this study, we developed a method, CellDART, which estimates the spatial distribution of cells defined by single-cell level data using domain adaptation of neural networks and applied it to the spatial mapping of human lung tissue. The neural network that predicts the cell proportion in a pseudospot, a virtual mixture of cells from single-cell data, is translated to decompose the cell types in each spatial barcoded region. First, CellDART was applied to a mouse brain and a human dorsolateral prefrontal cortex tissue to identify cell types with a layer-specific spatial distribution. Overall, the proposed approach showed more stable and higher accuracy with short execution time compared to other computational methods to predict the spatial location of excitatory neurons. CellDART was capable of decomposing cellular proportion in mouse hippocampus Slide-seq data. Furthermore, CellDART elucidated the cell type predominance defined by the human lung cell atlas across the lung tissue compartments and it corresponded to the known prevalent cell types. CellDART is expected to help to elucidate the spatial heterogeneity of cells and their close interactions in various tissues.**

## INTRODUCTION

Rapid progress in spatially resolved transcriptomics helped to comprehensively characterize the spatial interaction of cells in a tissue (1,2). Breakthrough technologies enabled capturing genome-wide spatial gene expression at a resolution of several cells (3) to the single-cell (4–6) and even subcellular levels (7). These methods have been used in various disease models to decipher spatial maps of genes of interest and culprit cells (8–12). Furthermore, emerging computational approaches facilitated the spatiotemporal tracking of specific cells and elucidated cell-to-cell interactions by preserving the spatial context (12–14). However, there is an inherent limitation in the spatial transcriptomic analysis that each spot or bead covers more than one cell in most cases. Even with a high-resolution technique, a small portion of several cells can be contained in the same spatial barcoded region. In addition, a tissue with a high level of heterogeneity, such as cancer, consists of a variety of cells in each small domain of the tissue (15). Thus, the identification of different cell types in each spot is a crucial task to understand the spatial context of pathophysiology using a spatially resolved transcriptome.

In this regard, recent computational tools have focused on integrating different types of transcriptomic data, particularly spatially resolved transcriptomic and single-cell RNA-sequencing (scRNA-seq) data (14,16–24). These tools have utilized the cell type signatures or variable genes defined by scRNA-seq and transferred the cell labels into spatial transcriptomic data. The majority of the approaches applied a statistical model or a matrix decomposition to infer the cell fraction in each spot (18,20–23). Meanwhile, calculating the proportion of cell types defined by scRNA-seq data from spots of spatially resolved transcriptomic

*To whom correspondence should be addressed. Tel: +82 2 2072 2802, Fax: +82 2 745 0345; Email: chy1000@snu.ac.kr
Correspondence may also be addressed to Young Tae Kim. Tel: +82 2 2072 3161, Fax: +82 2 745 0345; Email: ytkim@snu.ac.kr
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

data can be considered a domain adaptation task (25,26). A model that predicts cell fractions from the gene expression profile of a group of cells can be transferred to predict the spatial cell-type distribution.

In this paper, we suggest a method, CellDART, that implements modified adversarial discriminative domain adaptation (ADDA) (27) to infer the cell fraction in spatial transcriptomic data. The randomly selected cells from scRNA-seq data constitute a pseudospot in which the fraction of cells is known. The neural network model that extracts the cell fraction from the gene expression of a pseudospot is adapted to a different domain where spatial transcriptomic data are present. Consequently, the joint analysis of spatial and single-cell transcriptomic data elucidates the spatial cell composition and unveils the spatial heterogeneity of the cells. We utilized the proposed method to provide a resource for spatial mapping of the human lung cell atlas using the spatially resolved transcriptome of human lung tissue.

## MATERIALS AND METHODS

### Human brain cortex data

A publicly available Visium spatial transcriptomics dataset obtained from the DLPFC of postmortem neurotypical humans was downloaded from the data repository provided by the paper (28) [twelve datasets (151 673, 151 508, 151 676, 151 669, 151 674, 151 507, 151 671, 151 672, 151 670, 151 509, 151 510 and 151 670) in this study included 3639, 4384, 3460, 3661, 3673, 4226, 4110, 4015, 3498, 4789, 4634 and 3592 spots with 33 538 genes in common]. The count matrix for the tissue and brain layer information for each spot (cortical layers 1–6, white matter, and unknown) was added. Single-nucleus transcriptomic data acquired from the DLPFC of a healthy human control group ($n = 17$) were utilized for the joint analysis (29). The count matrix for 35 212 cells and 30 062 genes and the cell type annotation were included in the analysis.

### Mouse brain data

Visium spatial transcriptomics dataset for the mouse brain was downloaded from the 10X Genomics Data Repository. The 'Mouse Brain Serial Section (Sagittal-Anterior)' slide, which contains 2695 spots and 32 285 genes, was utilized for the CellDART analysis. For the joint analysis, scRNA-seq data obtained from the mouse primary visual cortex and anterior lateral motor cortex were selected. The count matrix was comprised of 23 178 cells and 45 768 genes. The layer-specific excitatory neuron types [L2/3 IT (intratelencephalic), L4, L5 IT, L5 PT (pyramidal tract), L6b, L6 CT (corticothalamic), and L6 IT] were determined based on the markers discovered in a previous study (30).

### Mouse hippocampus Slide-seq data

Slide-seq data for mouse hippocampus was obtained from the Single Cell Portal repository offered by the paper (6). The spatial data contains a count matrix for 23 264 genes across 53 173 beads. To infer the cellular composition in the Slide-seq data, scRNA-seq data of mouse hippocampus with 52 846 cells and 27 953 genes was utilized for the integration. The cell labels were determined based on the cell sub-clustering results derived from independent component analysis (31).

### Normal human lung data

Two normal lung samples were acquired from lung specimens from one patient who underwent surgical resection for lung cancer. We acquired samples and embedded them in optimal cutting temperature (OCT) compounds in the operating room and stored them at 80°C until cryosectioning. For cryosectioning, samples were equilibrated to −20°C with a cryotome (Thermo Scientific, USA). Sections were imaged and processed for spatially resolved gene expression using the Visium Spatial Transcriptomic kit (10× Genomics, USA). The protocol of this study was reviewed and approved by the institutional review board of Seoul National University (Application number: H-2009–081-1158). 'Lung 1' consists of 1591 spots and 36, 601 genes, and 'Lung 2' consists of 2683 spots and 36 601 genes. Single-cell data from the normal lung tissue of three subjects were downloaded and utilized for the integrative analysis (32). The count matrix for 65 662 cells and 26 485 genes was included in the downstream analysis. The cell types were assigned by representative marker genes and the anatomical location of the tissue. In addition, the cell types were classified based on expression profiles and anatomical locations (airway epithelial, alveoli epithelial, endothelial, muscle stromal, other stromal and immune cells) (32).

### Preprocessing spatial and single-cell datasets

All of the preprocessing steps were performed with Python (version 3.7) with the Scanpy toolkit (version 1.5.1) (33). The count matrices for both spatial and single-cell datasets were normalized with the 'scanpy.pp.normalize_total' function such that gene expression was comparable between spots or cells. For the single-cell data, the counts were log-transformed (scanpy.pp.log1p) followed by scaling (scanpy.pp.scale) and dimensionality reduction by principal component analysis (scanpy.tl.pca). Finally, the cells were represented with a *t*-distributed stochastic neighborhood embedding (t-SNE) plot (scanpy.tl.tsne and scanpy.pl.tsne) and were named based on the annotation data from publicly available datasets.

Meanwhile, the top $l$ highly expressed marker genes for each cell cluster from brain and lung samples were extracted with the Wilcoxon rank-sum test ('scanpy.tl.rank_genes_groups') based on the log-normalized count. Multiple comparison correction with the Benjamini–Hochberg method was applied, and genes were ranked by the corrected *P*-values. All of the cell type markers were pooled to form cell signature genes (Supplementary Figure S1A). The intersection between the cell signature genes and all provided genes from the spatial data was obtained. The downstream analysis was performed only with these intersecting genes.

For the next step, $k$ cells were randomly selected from the mouse or human brain and lung single-cell datasets. Random weights were given to each cell to mimic the cases in

which only the portion of the cells are contained in a spatial spot (Supplementary Figure S1A). The virtual mixture of the cells was defined as a 'pseudospot'. A total of $n$ pseudospots were generated, and the composite gene expression values were calculated for each pseudospot. Then, the log-normalized count matrices for single-cell, pseudospot, and real spot data were scaled such that the value lies between 0 and 1 in each cell or spot.

### CellDART: cell type inference with domain adaptation

The modified ADDA algorithm (27) was applied to develop a model to predict cell type proportions for each spot (Figure 1A and Supplementary Figure S1B). The training of neural networks was implemented based on Keras (version 2.3.1), TensorFlow (version 1.14.0), and scikit-learn (version 0.24.1) packages. First, a feature embedder that computes 64-dimensional embedding features from the gene expression data of either real spatial spots or pseudospots was defined. Since pseudospots and real spots are presumed to consist of similar cell types, the feature embedder was shared between the source (pseudospots) and target (real spots) datasets. The feature embedder was comprised of two fully connected layers, each of which underwent batch normalization and activation by the exponential linear unit (ELU) function. The outputs of the first layer and second layer have 1024 and 64 dimensions, respectively. Domain and source classifiers were defined such that they could predict the cell fraction in each spot and discriminate pseudospots from spots, respectively. The domain classifier consisted of two fully connected layers. The first layer with 32-dimensional output was connected to the embedded features. After batch normalization, ELU activation, and dropout, another layer to discriminate real spots from pseudospots was applied. The source classifier is directly connected to the embedded features of the feature extractor as a one-layer model connected to the feature embedder. Therefore, the feature extractor attached to either of the classifiers was named a domain or source classification model.

The loss function of the source classifier that predicts cell type proportions was defined by Kullback–Leibler divergence (KLD). KLD is decreased when the distribution of predicted and real cell type proportions is similar. Likewise, the loss function of the domain classifier was assigned as binary cross-entropy, a probability that a certain dataset is correctly allocated to the assigned domain label (pseudospots or real spots).

$$\boldsymbol{L_s} = KLD(Y_s||S(f(X_s))) = -\sum_k Y_{k,s} log\left[\frac{S(f(X_s),k)}{Y_{k,s}}\right]$$

$$\boldsymbol{L_{adv,1}} = -log\,D\left(f\left(X_p\right)\right) - log\left[1 - D(f(X_r))\right]$$

$$\boldsymbol{L_{total}} = \boldsymbol{L_s} + \alpha \boldsymbol{L_{adv,1}}$$

$$\boldsymbol{L_{adv,2}} = -log\,D(f(X_r)) - log\left[1 - D\left(f\left(X_p\right)\right)\right]$$

$X_p$: gene expression of pseudospots, $X_r$: gene expression of real spots, $k$: index of the cell type, $Y_{k,s}$: proportion of $k$th cell type in the pseudospots, $f$: feature embedder, $S$: source classification model, $D$: domain classification model, $\alpha$: weight of loss between the source and domain classifier.

An initialization and two optimization steps were implemented to train the model. The loss functions utilized during the training were summarized as the above formulae. For the initialization of weights of feature embedder and source classifier, pre-training was performed to predict cell type proportions from pseudospots. Then the optimization was performed as an adversarial domain adaptation. First, the model was trained to minimize $L_s$ as a pre-training process. As an adversarial loss, it is typical to train the model with the standard loss function with inverted labels as in the above formula. Thus, two optimization processes were applied (Supplementary Figure S1B). The networks were optimized to minimize $L_s$ and $L_{adv,1}$ with fixed weights of the domain classifier. Then, the domain classifier was trained to minimize $L_{adv,2}$ with fixed weights of the feature embedder, $f$, and the source classifier, $S$. These two processes were repeated with a training parameter of the number of iterations.

Finally, the trained model, CellDART, predicted the cell fraction in each spot from the spatial data, and the results for each cell type were spatially mapped to the tissue by the 'scanpy.pl.spatial' function. Additionally, the distribution of cell type compositions across the brain layer was represented with the 'scanpy.pl.stacked_violin' function.

### Optimal parameter selection

To search optimal parameters for CellDART, the performance of the model was evaluated in different parameter settings with DLPFC dataset as a reference. The DLPFC spatial data contain the brain layer information (layer 1 to layer 6, white matter, and unknown), and the single-cell data have ten layer-specific excitatory neuron clusters (*Ex_1_L5_6*, *Ex_2_L5*, *Ex_3_L4_5*, *Ex_4_L6*, *Ex_5_L5*, *Ex_6_L4_6*, *Ex_7_L4_6*, *Ex_8_L5_6*, *Ex_9_L5_6*, and *Ex_10_L2_4*). The layer specificity of excitatory neurons was determined by the cell types identified from the single-nucleus RNA-seq data (29) with the layer markers suggested by several studies (34–36). Receiver operating characteristic (ROC) analysis was performed to determine whether the spatial cell fraction of excitatory neuron clusters could differentiate the specific cortical layer. For example, the spatial composition of Ex_2_L5 was predicted and the specificity and sensitivity were tested for each threshold value whether it can assign the spatial spots as belonged to layer 5.

The performance of CellDART was tested by modulating the number of pseudospots ($n$) ranging from 800 to 640 000 (800, 4000, 20 000, 40 000, 80 000, 160 000, 320 000, 640 000). The performance plateaued when $n$ = 20 000 which was approximately five to ten times the number of real spots in one spatial dataset (Supplementary Figure S2A) and did not significantly improve when $n$ became larger. Also, CellDART was implemented for a different number of markers in each cell cluster ($l$) ranging from 5 to 160 (5, 10, 20, 40, 80, 160) which corresponds to the total number of markers from 109 to 1157 (109, 203, 364, 642, 1157, 1402). The performance showed a trend of increase when $l$ increased (Supplementary Figure S2A). However, the to-

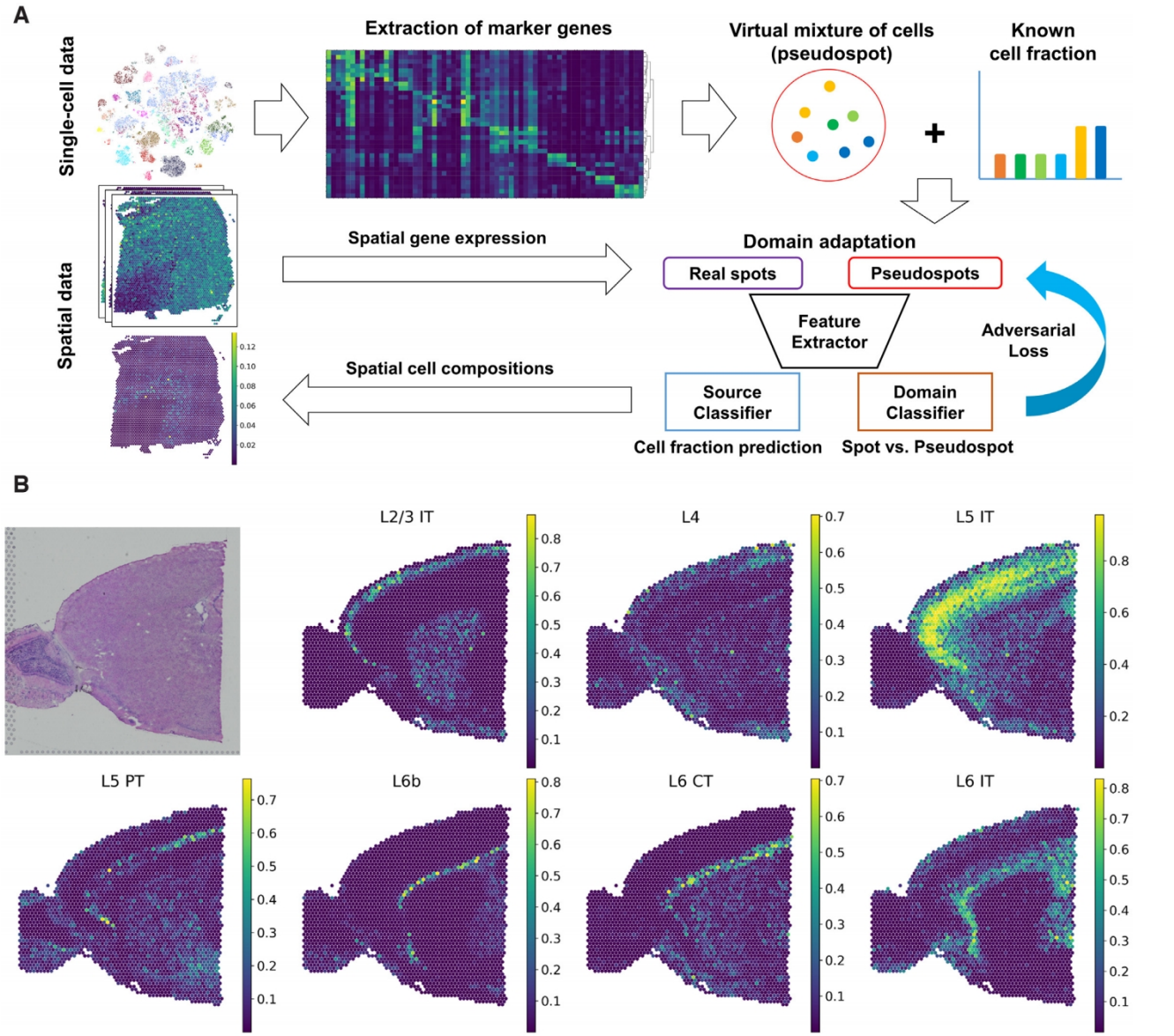


**Figure 1.** CellDART analysis in human and mouse brain tissues. (**A**) Schematic diagram for CellDART analysis. The human dorsolateral prefrontal cortex (DLPFC) dataset was preprocessed, and the marker genes for each cell cluster were extracted. The shared genes between the pooled cluster markers and the spatial transcriptomic data were selected for the downstream analysis. Then, 20 000 pseudospots were generated by randomly selecting eight cells from the single-cell data and giving them random weights. A feature extractor with a source and domain classifier was trained to estimate the cell fraction from the pseudospot and distinguish the pseudospots from the real spatial spots. First, the weights of the neural network were updated except for the domain classifier. Next, the data label for the spot and pseudospot was inverted, and only the domain classifier was updated. Finally, the trained CellDART model was applied to spatial transcriptomics data to estimate the cell proportion in each spot. (**B**) Spatial mapping of seven layer-specific mouse excitatory neurons predicted by CellDART. The figure in the top left corner shows the mouse brain tissue slide. Colormaps present the maximum and minimum values for the corresponding cell fraction.

tal time consumed for calculation gradually increased as both parameters increased (Supplementary Figure S2B). The number of cells in a pseudospot was fixed (brain Visium: $k = 8$, brain Slide-seq: $k = 2$ and lung: $k = 10$) and other parameters were tuned to obtain an optimal performance with an acceptable computation time.

Finally, the number of marker genes for each cell cluster ($l$) was set to 20 and 10 for brain and lung tissues, respectively. The number of pseudospots ($n$) was set to 20 000 and 500 000 for Visium and Slide-seq data. The iteration number was 3000, the minibatch size was 512, and the learning rate for the training domain classifier was 0.005. The loss weights between the source and domain classifiers ($\alpha$) were 1:0.6 in brain tissues and 1:1 in lung tissues.

## Comparison to other tools

For the DLPFC datasets, the performance of CellDART was compared with six other computational tools: Scanorama (16), Cell2location (20), RCTD (21), SPOT-

light (23), Seurat (version 3) (17), and DSTG (24). Briefly, Scanorama and Seurat align the single-cell and spatial datasets based on the mutual nearest neighbors (MNNs) between the two. Meanwhile, Cell2location assumes that the count matrix from spatial data follows a negative binomial distribution and can be decomposed into a linear combination of cell type signatures. RCTD postulates that count in spatial spots follow a Poisson distribution and applies the maximum likelihood method to estimate cell proportions. Next, SPOTlight utilizes non-negative matrix factorization regression to obtain cell type-specific gene signatures and applies it to infer the cellular composition of spots. Finally, DSTG generates a graph between the virtual mixture of cells and the spatial spots and adopts a graph convolutional neural network to predict spatial cell fraction from the graph structure. These six toolkits were applied, and cell density from Cell2location and cell fraction from Scanorama, RCTD, SPOTlight, Seurat and DSTG were spatially mapped to the tissue. For Cell2location analysis, we set the total number of cells, cell types, and groups of cell types in each spatial spot as 8, 9 and 5, respectively. Next, for RCTD, we selected 'full mode', which does not restrict the number of cells in each spot. In the case of SPOTlight, we randomly selected 100 cells per cell cluster and extracted the top 3000 highly variable genes for the downstream analysis. For other tools, the default parameters offered by the user guide were adopted for the downstream analysis. The ROC analysis was performed in the DLPFC dataset across 12 slides for discriminating spatial localizing patterns of layer-specific excitatory neurons. AUC values were compared between CellDART, Scanorama, Cell2location, RCTD, SPOTlight, Seurat, and DSTG. In addition, as another measure of performance, the significance of AUC was assessed based on a null hypothesis that AUC is below or same with 0.5. Wilcoxon rank-sum test was performed to evaluate whether the predicted neuron fraction in a specific layer was significantly larger than that in other locations. Multiple comparison correction was implemented based on Bonferroni method and corrected $P$-value $<0.05$ was considered significant. Finally, the percentage of significant AUCs were compared across the computational tools. The statistical analysis and visualization were implemented with scikit-learn (version 0.24.1), scipy (version 1.6.0), and matplotlib (version 3.3.4).

### Spatial mapping of lung cells to normal lung tissue data: investigation of the spatial heterogeneity of the cells

The boundary of the tissue structures in two lung samples (lung 1 and lung 2) was delineated by a pathologist on H&E staining images, and the spots were classified into 6 domains: alveolar space, bronchial epithelium, fibrous stroma, immune cluster, terminal bronchiole, and vessels. The uncertain region of the tissue was named 'unknown' and more specifically 'unknown stroma' if the corresponding region was stromal tissue. After transferring the single-cell cluster labels to the spatial data, the minimum and maximum cell fraction values across all spots were scaled to 0 and 1, respectively. The cell types were divided into six categories based on where the cells were commonly found (32) ('airway epithelium', 'alveoli epithelium', 'endothelial', 'muscle stromal' and 'other stromal'). The average scaled cell fraction in each tissue domain according to cell types was visualized with a seaborn clustermap function (version 0.11.1), and the cell type categories were color-coded and presented on the left side. For the next step, the cell types showing highly different cell fractions across the histological domains were selected, and their spatial composition was mapped to the tissue. Cell type selection was performed with the Wilcoxon rank-sum test, and Benjamini-Hochberg corrected $P$-values were computed. The cell types in each tissue domain were ranked based on a ratio of the average scaled cell fraction in a specific tissue domain to the rest of the domains. The cell types with an average scaled fraction $<0.2$ were excluded from further analysis. An adjusted $P$-value $<0.05$ was considered significant.

## RESULTS

### Decomposition of spatial cell distribution with CellDART in human and mouse brain data

The performance of CellDART was assessed in publicly available single-nucleus (also considered single-cell data) and spatial transcriptomic autopsy samples of the human dorsolateral prefrontal cortex (DLPFC), each of which was obtained from two different subject groups with no neurological disorders. Additionally, single-cell and spatial datasets acquired from the mouse brain were utilized. First, both single-cell datasets were preprocessed, and the cells were named after the annotation data provided by the original papers (29,30). The 33 and 29 annotated cell clusters from the human and mouse brains were visualized by t-distributed stochastic neighbor embedding (t-SNE) plots (Supplementary Figure S3A, B), and marker genes for each cluster were extracted (Supplementary Figure S3C, D and Supplementary Tables S1 and S2). The cell clusters showed distinct gene expression patterns represented by cell type-specific marker genes.

A specific number of cells ($k = 8$) were randomly sampled from the single-cell data with random weights to generate pseudospots (number of pseudospots = 20 000). Then, composite gene expression values were computed based on marker genes (Figure 1A). A neural network was trained to accurately decompose the pseudospots, and another network, the domain classifier, was trained to discriminate spots of real spatially resolved transcriptomes from pseudospots. During the training process, the weights of neural networks were updated to predict cell fractions and fool the domain classifier to avoid discriminating spots and pseudospots (Figure 1A). As a result, the neural network, source classifier, was trained to estimate cell fractions in both the pseudospots and the real spatial spots as an adversarial domain adaptation process.

The layer-specific excitatory neuron fraction in each spot was predicted by CellDART and spatially mapped to the tissue. In the case of mouse brain tissue, seven excitatory neurons showed spatially restricted patterns in a specific cortical layer (Figure 1B). In addition, the 10 excitatory neuron clusters in the human brain presented layer-specific distribution patterns across the six cortical layers (layers 1–6) (Figure 2A and Supplementary Figure S4A). Additionally,

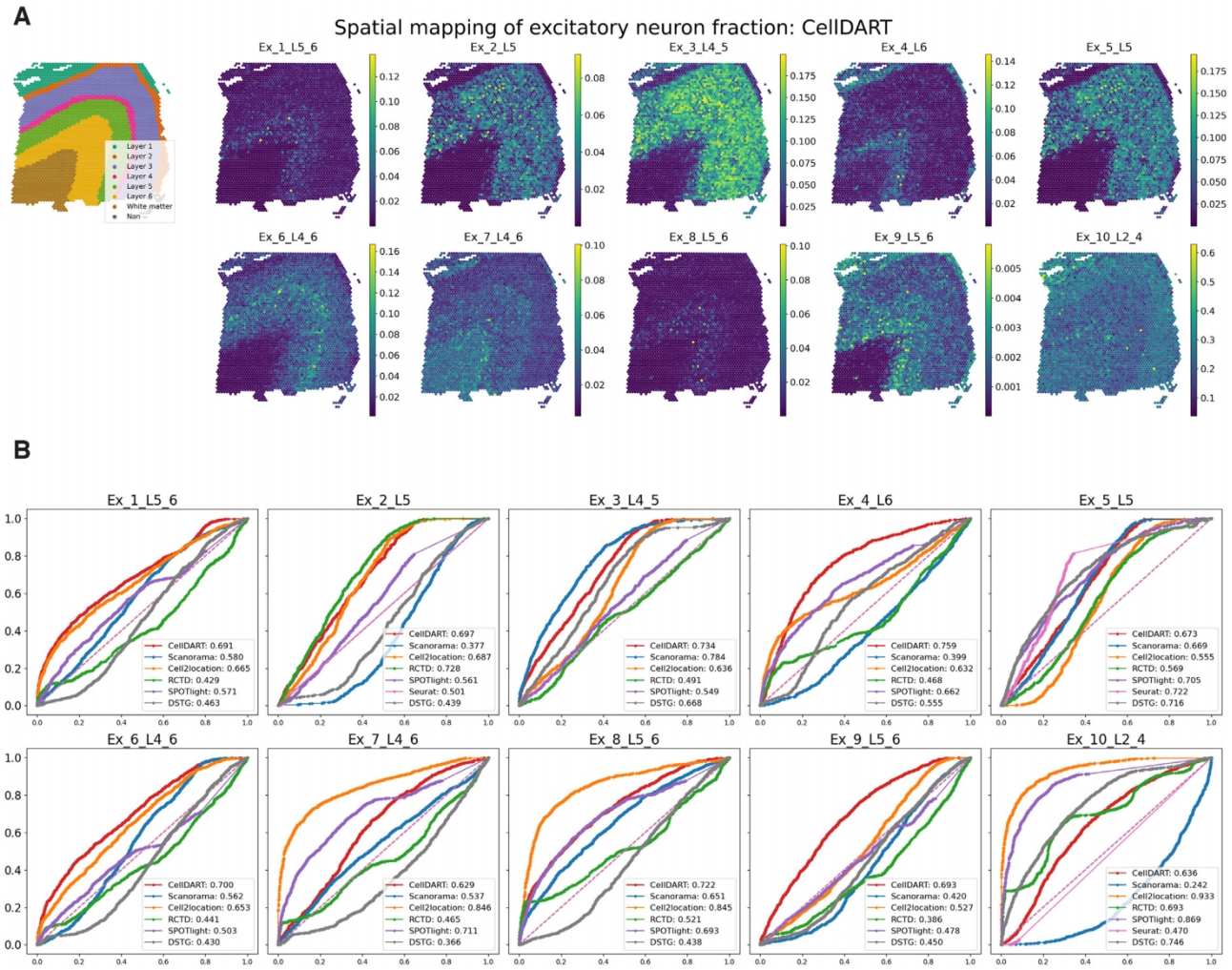


**Figure 2.** Implementation of CellDART in a human dorsolateral prefrontal cortex dataset. (**A**) Spatial mapping of 10 layer-specific excitatory neurons predicted by CellDART. The figure in the top left corner shows the layer annotation for each spatial spot. The layer consists of cortical layers 1 to 6 and white matter. 'Nan' represents the spot without the layer information. Colormaps present the maximum and minimum values for the corresponding cell fraction. (**B**) Receiver operating characteristic (ROC) analysis for predicting the layer-specific distribution of excitatory neurons. The computational tools CellDART, Scanorama, Cell2location, RCTD, SPOTlight, Seurat (version 3), and DSTG which estimate cell types in the spatial spots, were compared by means of the area under the curve (AUC). The ROC curves for CellDART, Scanorama, Cell2location, RCTD, SPOTlight, Seurat, and DSTG are color-coded, and AUC values are presented in the lower right corner of each plot. In Seurat, some cell types showed the cell fraction of 0 in all the spots; thus, the ROC curve was not visualized in those cases.

the spatial density of human non-neuronal cells was estimated with a neural network (Supplementary Figure S4B). Astrocytes were mainly located in layer 1 and layer 6, while oligodendrocyte cluster 3 (*Oligos_3*), which showed an ∼10–60 times higher cell fraction than the other two clusters (*Oligos_1* and *Oligos_2*), was predominantly localized in white matter. Endothelial cells, microglia, and macrophages were spatially distributed across the six cortical layers with low cell proportions.

Meanwhile, the impact of domain adaptation on the improvement of model performance was assessed. The feature embedder with the source classifier was trained only with pseudospots and domain adaptation was not applied. The model without domain adaptation was named 'NN_wo_da'. The cell fraction predicted by 'NN_wo_da' presented spatially localizing patterns in *Ex_2_L5* and *Ex_5_L5*; however, other cell types showed uneven spatial distributions (Supplementary Figure S5A). ROC analysis was implemented and the accuracy of CellDART in spatially localizing layer-specific excitatory neuron fraction was compared with that of NN_wo_da (Supplementary Figure S5B). In general, the performance of NN_wo_da was inferior to CellDART, which proves the necessity of domain adaptation to obtain an optimal result.

## Comparison of CellDART with other integration tools in human brain tissue

The capability of CellDART to accurately assign cell types in spatial spots was compared with that of six computational tools: Scanorama, Cell2location, RCTD, SPOTlight, Seurat version 3, and DSTG. First, the methods were employed to decipher the spatial distribution of excitatory neu-

rons and non-neuronal cells in the DLPFC 151673 dataset and the results were qualitatively assessed.

Scanorama showed a few excitatory neurons of cortical layer-specific distribution patterns, whereas *Ex_2_L5*, *Ex_4_L6*, *Ex_9_L5_6* and *Ex_10_L2_4* excitatory neurons were distributed differently from the known cortical distribution (Supplementary Figure S6A, B). Astrocytes and oligodendrocytes did not show consistent cell distribution patterns across the cell subtypes. Endothelial cells, microglia, and macrophages were predominantly localized in layer 1, layer 6, and the white matter according to the Scanorama analysis (Supplementary Figure S6C). In the case of Cell2location, excitatory neurons showed layer-specific distribution patterns except for *Ex_5_L5* and *Ex_9_L5_6* where the distribution was uneven and not restricted to layer 5 and layers 5–6, respectively. Overall, non-neuronal cells showed a layer-specific localization pattern; however, *Astros_2* and *Astros_3* exhibited heterogeneous patterns in the same layer (Supplementary Figure S7). Next, for RCTD, a few excitatory neurons (*Ex_2_L5* and *Ex_10_L2_4*) exhibited a high cell fraction in the corresponding cortical layer of a known layer specificity; however, other excitatory neurons presented heterogeneous patterns of distribution (Supplementary Figure S8A, B). Additionally, in the non-neuronal cells, the spatial distribution was relatively uneven and not layer-specific except for three oligodendrocyte cell clusters (Supplementary Figure S8C). SPOTlight exhibited spatially restricted patterns of distribution for some excitatory neurons and non-neuronal cells. However, *Ex_1_L5_6*, *Ex_2_L5*, *Ex_3_L4_5*, *Ex_6_L4_6,* *Ex_9_L5_6*, *Astros_2*, *Astros_3*, and *Oligos_3* were not layer-specific or did not show even distribution in the same layer (Supplementary Figure S9). Seurat version 3 was successful in discriminating layer-specific distribution patterns for a few cell types (*Ex_5_L5*, *Oligos_1* and *Oligos_3*). In other cell types, the cell fraction was predicted to be 0 in all the spots or spatial distribution did not exhibit layer-specific localization patterns (Supplementary Figure S10). Finally, DSTG could predict spatial patterns of a few excitatory neuron types (*Ex_3_L4_5*, *Ex_5_L5* and *Ex_10_L2_4*). However, the spatial distribution in the rest of the cell types was not localized to the corresponding layer and showed uneven patterns (Supplementary Figure S11).

Receiver operating characteristic (ROC) curve analysis was implemented in 151673 tissue to quantitatively compare the performance of the seven different tools in predicting the layer-specific distribution of excitatory neurons (Figure 2B). The spatial spots in the DLPFC data were classified into a specific layer, layer 1 to layer 6, white matter, or unknown, with manual annotation data (28) based on the tissue morphology and marker genes (34). The ROC curves for 10 excitatory neurons revealed that CellDART has overall good prediction accuracy, with an area under the curve (AUC) ranging from 0.629 in *Ex_7_L4-6* to 0.759 in *Ex_4_L6*. In the case of Cell2location and SPOTlight, a few cell types presented AUC over 0.800 (Cell2location: *Ex_7_L4_6*, *Ex_8_L5_6* and *Ex_10_L2_4*; SPOTlight: *Ex_10_L2_4*) (Supplementary Figure S12A). However, some of the cell types presented low accuracy with AUC below 0.600. Scanorama, RCTD, and DSTG exhibited relatively low discriminative accuracy in several cell types with AUCs <0.500 and comparable AUCs with CellDART for a few cell types. Meanwhile, Seurat could compute the spatial cell fraction in 3 out of 10 cell types, and among those, two of them showed AUC <0.600. The confidence interval of the AUC was generated by bootstrapping, and the results were compared among the seven methods (Supplementary Figure S12B). In general, CellDART showed robust performance in predicting layer-specific localization patterns across all excitatory neurons.

Next, the performance of CellDART was further validated across 12 Visium datasets acquired from human DLPFC. The AUC values were evaluated for each layer-specific excitatory neuron type (Figure 3A) and all cell types (Figure 3B). CellDART showed significantly superior performance to other computational methods except for Cell2location in which median AUC values were similar (CellDART: 0.674 and Cell2location: 0.679). Additional comparison was performed to evaluate how stably the layer specificity can be predicted for each cell type. The percentage of significant AUC of ROC curves was calculated with a null hypothesis of the AUC $\leq$0.5. CellDART (106/120) presented a higher number of the significant AUC of ROC than Cell2location (100/120) as well as other methods (Figure 3C). Meanwhile, the total running time per slide (slide 151673) was compared between the tools and CellDART was ~20 times faster than Cell2location (CellDART: 212 s and Cell2location: 4479 s). The running time of CellDART was the shortest among the methods (Figure 3D). In summary, CellDART exhibited the most stable performance on predicting layer specificity of cell types with the shortest computational time for predicting distribution patterns of brain cells.

### Application of CellDART in Slide-seq data

The capability of CellDART was evaluated in Slide-seq data to decompose cellular proportion in each spatial bead. CellDART could precisely localize the cells, especially for CA1, CA2, CA3 and dentate principal cells, and entorhinal cortex, which are known to be restricted in specific anatomical locations (Figure 4) (37). Also, astrocytes are found to have high density in stratum oriens and stratum radiatum and oligodendrocytes in corpus callosum as previously reported (38–40). Although the entorhinal cortex is not included in the view, the corresponding cell type showed a high proportion in the cortical layers of the slide. It is partly explained by the overlap of transcriptional signatures across cortical neurons in different regions (41). In short, CellDART can be applied to spatial transcriptome with different spatial resolutions.

### Discovery of spatial heterogeneity of human lung tissue with CellDART

CellDART was further applied to normal lung spatial transcriptomic data. Human lung tissue was obtained from one patient who underwent lobectomy for surgical resection of lung cancer. The two normal lung tissues were dissected far from the tumor and pathologically confirmed to have no tumor cells (Supplementary Figure S13). The demographic

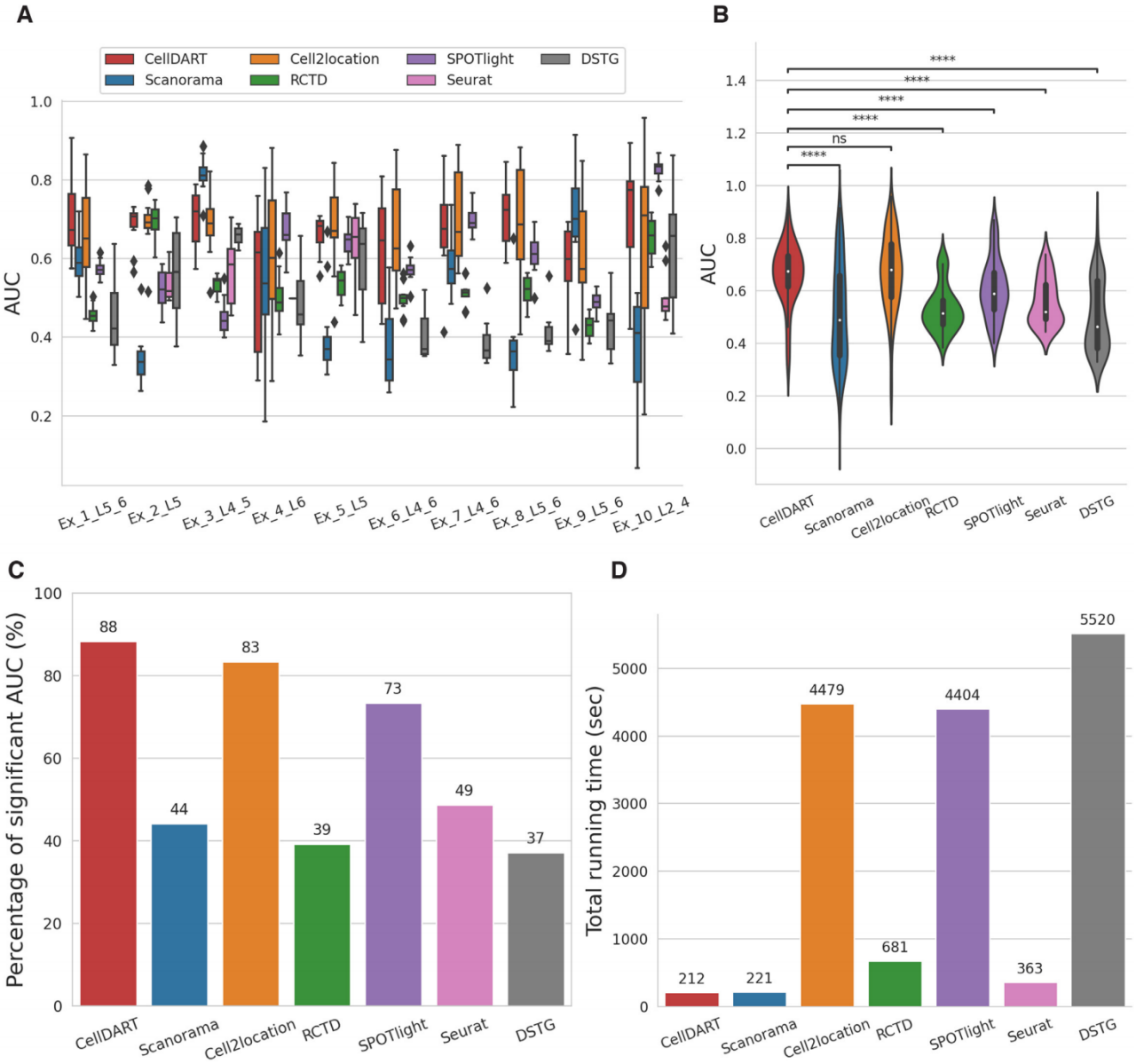


**Figure 3.** Validation of CellDART across 12 tissue slides of DLPFC. (A, B) Boxplots for comparison of performance in seven computational tools. The cellular composition of 10 layer-specific excitatory neuron types was predicted by seven different approaches. Then, the layer classifying accuracy measured using AUC was compared between the methods. (**A**) The boxplot for each cell type shows the AUC values across the 12 tissue slides of DLPFC. In Seurat, some cell types showed the cell fraction of 0 in all the spots and AUC values were not plotted in those cases. (**B**) The AUC values for the 10 neuron types across the 12 slides were pooled and their distribution in seven toolkits was visualized with violinplots combined with boxplots. Wilcoxon signed-rank test was performed between AUC values from CellDART and those from other computational tools. Statistical significance was visualized on top of the violinplots (ns: not significant, ****$P$-value $< 10^{-4}$). (**C**) The percentage of significant AUC values of ROC curves (statistical testing for AUC $> 0.5$) were compared across the seven tools and visualized with a barplot. (**D**) The total running time per tissue slide (Slide 151673) was calculated for the seven tools and visualized with a barplot. The number on top of each barplot represents the computing time (second) of each tool.

features of the patient are summarized in Supplementary Table S3. The publicly available human lung cell atlas data were used for spatial mapping of lung cell types using CellDART. They consisted of scRNA-seq from three human normal lung tissues (32). The single-cell data were embedded in low-dimensional space by a t-SNE plot, and 57 cell clusters showed discrete gene expression patterns (Supplementary Figure S14A). The marker genes selected in each cell cluster were pooled and utilized in the downstream analysis (Supplementary Figure S14B and Supplementary Table S4). After the generation of pseudospots, CellDART was trained to assign the proportion of cells in the real spatial spots. The tissue slides from two spatial datasets (lung 1 and lung 2) were manually segmented, and each spot was classified into seven categories: alveolar space, bronchial epithelium, fibrous stroma, immune cluster, terminal bronchiole, vessels, and unknown region (Figure 5A, B). In addition, the cell types were classified into five categories based

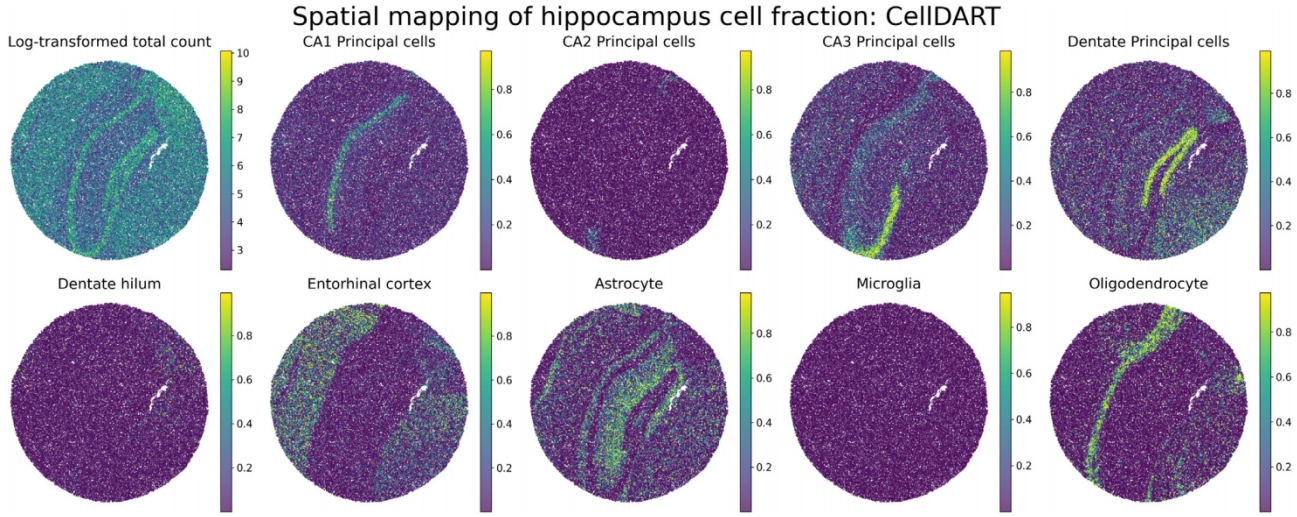


**Figure 4.** Application of CellDART in mouse hippocampus Slide-seq data. Spatial mapping of 9 hippocampal cell fractions on the Slide-seq data. The figure in the top left corner shows the log-transformed total count in each bead. Colormaps present the maximum and minimum values for the corresponding log-transformed count or cell fraction.

on a previous study (32). An average scaled cell proportion of spots in the same tissue domain was calculated, and the values were expressed with heatmaps.

In both the lung 1 and lung 2 datasets, each cell type showed different distribution patterns across the segmented tissue domains (Figure 5C, D). Among the 'airway epithelial' cells (blue color on the left side of the heatmaps), proximal ciliated, ciliated, mucous, and club were mainly localized in the bronchial epithelium or terminal bronchiole. 'Alveoli epithelial' cells (orange color) were localized in the alveolar space of lung 2 data. 'Muscle stromal' cells (red color) were mainly distributed in the unknown stroma in the lung 1 data and terminal bronchioles or vessels in the lung 2 data. 'Immune' cells (brown color), particularly B-cells, monocytes, and dendritic cells, were predominantly located in the immune cluster tissue domain. Finally, 'endothelial' and 'other stromal' cells did not present spatially localized patterns of distribution.

For the next step, cell types that showed highly different cell fractions across the tissue domains were selected (Supplementary Table S5). The top 7 cell types were ranked by the ratio of the scaled cell fraction in the specific tissue domain compared to the other domains and were mapped to the tissue (Figure 6A, B). The cell types with an average scaled cell fraction in the domain $<0.2$ were excluded. For lung 1 tissue, 'proximal basal' and 'proximal ciliated' cells, which were previously described as 'airway epithelial' cells (blue color), were predominantly distributed in the bronchial epithelium or terminal bronchiole tissue domain (Figure 6A). Most 'immune' cell types (brown color) were localized in the immune cluster domain except for 'classical monocytes', which are commonly found in the bronchial epithelium. In lung 2 tissue, 'ciliated', 'proximal ciliated', 'club', and 'mucous' cells, which are included in 'airway epithelial' (blue color) cells, were mainly located in the terminal bronchiole domain (Figure 6B). 'Capillary intermediate 2' included in the 'endothelial' cell type (green color) was localized in the alveolar space domain, while another endothelial cell type, 'artery', was mainly located in the fibrous stroma and vessels. Additionally, 'alveolar epithelial type 2' in the 'alveoli epithelial' (orange color) was predominantly distributed in the alveolar space domain. In summary, CellDART could precisely localize the spatial distribution of heterogeneous cell types in normal lung tissue.

## DISCUSSION

CellDART, which adapts the domain of single-cell and spatial transcriptomic data, could be flexibly applied to brain and lung tissues to decompose the spatial distribution of various cell types. The suggested approach was capable of accurately predicting the layer-specific localization of excitatory neurons and non-neuronal cells in the brain. Additionally, CellDART was superior to other computational tools, Scanorama, Cell2location, RCTD, SPOTlight, Seurat and DSTG in spatially localizing multiple excitatory neuron subtypes. Compared to Cell2location which showed similar performance in terms of AUC, CellDART was 20 times shorter in calculation time and it showed stable performance considering the number of significant AUCs over 0.5 to predict layer-specific neurons. However, no method showed a totally higher performance across all cell types, though the overall performance was higher in CellDART or Cell2location. In this regard, evaluation of various tools is required to assure the predicted location results, especially in the spatial datasets with no ground truth. Notably, the short running time of CellDART will serve as an advantage in the case where the multiple tools have to be tested as well as analyses to be performed in multiple spatial transcriptomic data. Besides, CellDART could be applied to Slide-seq data to estimate spatial cell compositions. Finally, our domain adaptation method deciphered the spatial distribution patterns of various lung cells across the different tissue compartments.

Since tissues consist of variable cells, spatial mapping of various cell types is crucial to understanding functions and

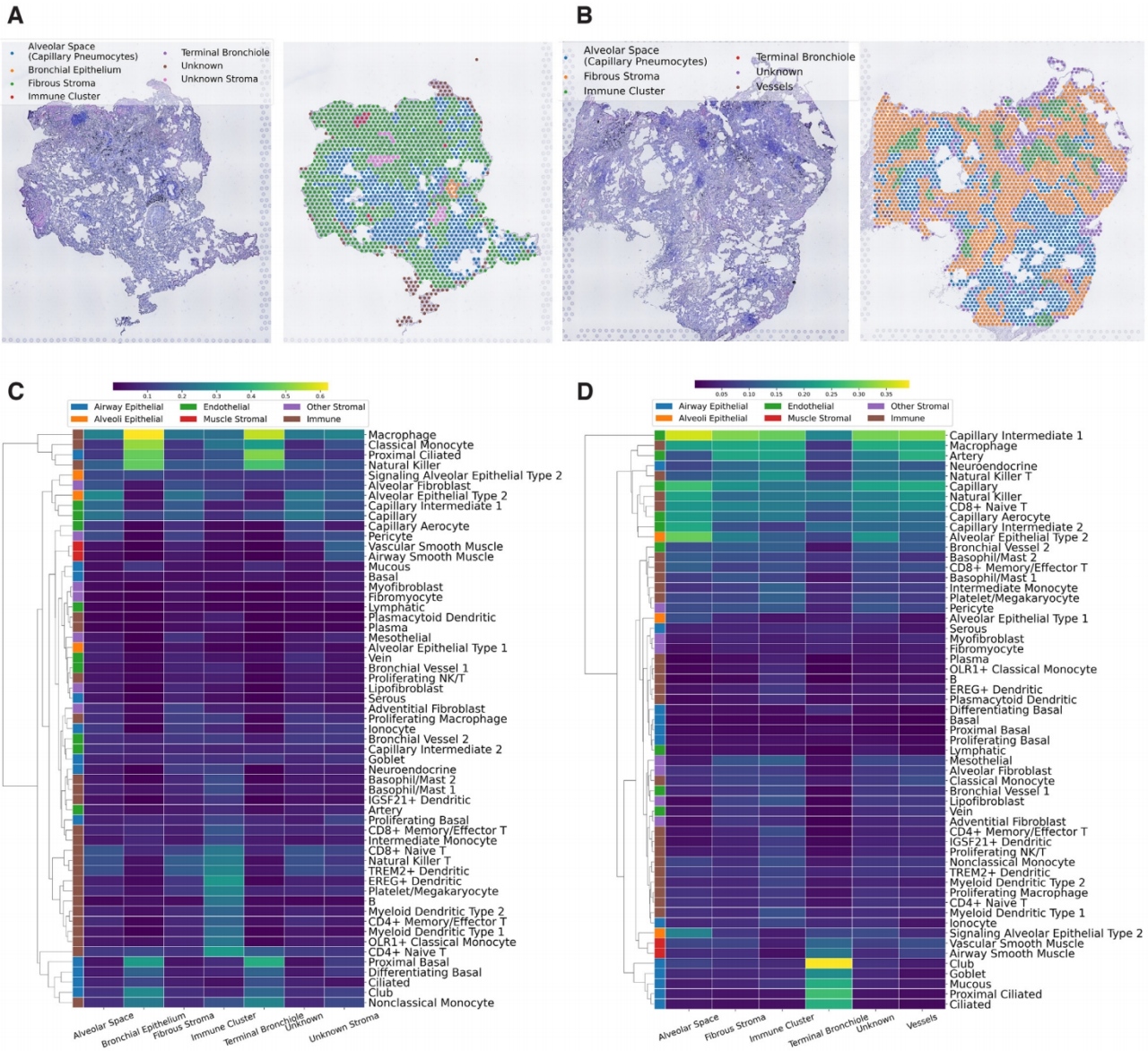


**Figure 5.** Application of CellDART in human lung data to decipher the tissue microenvironment. (**A**, **B**) Segmentation of the histological structures in normal lung tissue (**A**) 1 and (**B**) 2. The tissue was divided into six domains: alveolar space (capillary pneumocyte), bronchial epithelium, fibrous stroma, immune cluster, terminal bronchiole, and vessels. Uncertain stromal tissue was defined as 'unknown stroma', and the other unspecified areas were defined as 'unknown'. (**C**, **D**) Heatmaps for the average scaled cell fraction in each histological domain of (**C**) lung 1 and (**D**) 2 tissues. The cell types were classified into six categories ('alveolar epithelial', 'alveoli epithelial', 'endothelial', 'muscle stromal', 'other stromal' and 'immune') based on the original paper of the human lung cell atlas (32) and color-coded on the left side of the heatmaps. Additionally, hierarchical clustering was performed based on cell fraction profiles across the tissue compartment to visualize the similarity between the cell types.

pathophysiology. As examples of brain and lung tissues in our results, deciphering layer-specific cell types and tissue compartments associated with specific cell types could be a resource to understand the underlying biology. Furthermore, although a simple approach using a priori cell type markers of specific cells could be used to understand brief patterns of cell types, tissues with complex and heterogeneous cell types such as the lung require spatial mapping of precisely defined cell types based on single-cell level studies.

CellDART could be adopted for spatial transcriptomic data to portray the cellular landscape of the tissue by preserving the spatial context. In the brain, many cell types have their own regional identities, and the heterogeneous cells in each location shape distinct functional characteristics (42–44). Therefore, it is crucial to precisely decompose brain cell types in spatial transcriptomic data to comprehensively analyze the spatial crosstalk among cells. In the mouse brain tissue, layer-specific excitatory neurons revealed localized patterns of distribution across the cortical layers (Figure 1B). Also, when applied to Slide-seq data, CellDART was successful in mapping various hippocampal cell types into anatomically relevant positions (Figure 4). Additionally, in the validation study with human DLPFC tissue, not only layer-specific excitatory neurons

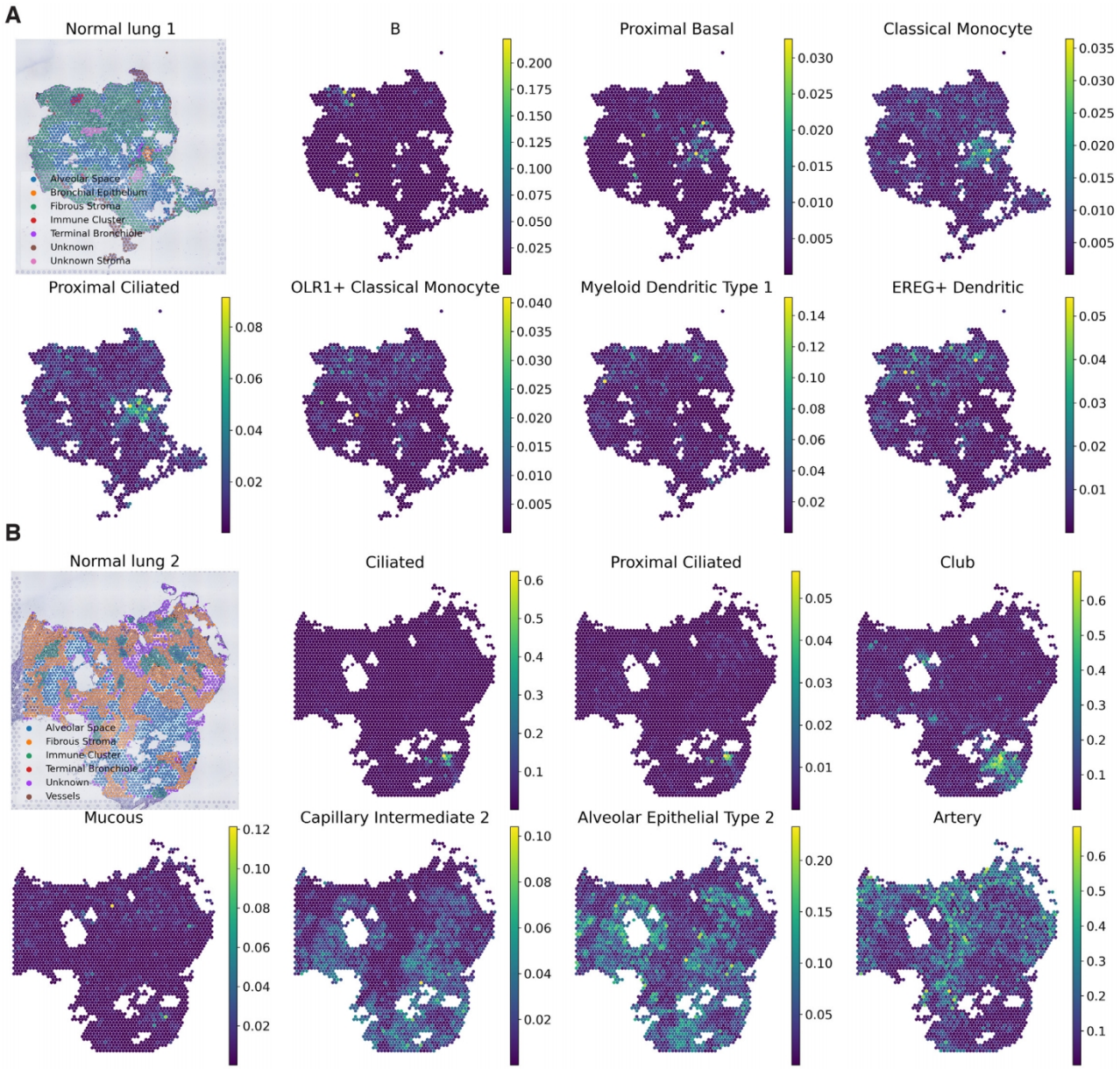


**Figure 6.** Spatial compositions of tissue compartment-specific cell types in the human lung. (**A**, **B**) Spatial mapping of the lung cell fraction for the top 7 cell types predominant in a specific tissue compartment of the (**A**) lung 1 and (**B**) 2 datasets. The cell types were ranked based on the average scaled cell fraction in a specific tissue domain compared to the other domains. The figure in the top left corner shows the classification of the tissue domain for each spatial spot. Colormaps present the maximum and minimum values for the corresponding cell fraction.

but also glial cells such as astrocytes and oligodendrocytes showed spatially restricted patterns (Figure 2 and Supplementary Figure S4B). Three astrocyte subtypes, Astros_1, Astros_2, and Astros_3, were predominantly located in L1 and L6. It has been reported that astrocytes form cortical layer-specific morphological and gene expression features (42,45,46); however, the abundance of excitatory neurons in the mid cortical layers may have masked the presence of diverse astrocyte populations. Meanwhile, one of the oligodendrocyte subtypes, Oligos_3, which presented a higher absolute cell fraction than the other subtypes, was localized in the white matter. This finding is in line with a previous study showing that oligodendrocytes are highly restricted in white matter compared to gray matter (39). On the other hand, the spatial distribution of glial cells did not match the known localization patterns or presented heterogeneous distribution in Scanorama, Cell2location, and DSTG (Supplementary Figures S6, S7 and S11). In the case of RCTD, SPOTlight, and Seurat, oligodendrocytes showed strong localization patterns in white matter; however, other glial cells did not present a layer-specific distribution (Supplementary Figures S8-S10).

Our suggested method, CellDART, was further applied to human lung tissues where a mixture of various cells was present across the tissue compartments (32). The cell types that exhibited a high proportion in a specific tissue domain

corresponded with a previous paper presenting the cell type predominance in the lung compartments (Figures 5 and 6). Notably, cell types that are included in the same category as divided by airway and alveoli epithelial, endothelial, muscle stromal, other stromal, and immune cells presented similar spatial localization patterns (Figure 5C, D). Next, the spatial overlap between cell types from different categories was investigated. The two alveolar epithelial cell types ('alveolar epithelial type 1, 2') presented similar spatial correlation patterns with 'capillary' and 'capillary aerocyte' included in endothelial cell types (Supplementary Figure S15). It is in line with the previous finding that those cell types are co-localized at the alveoli structure (32). On the contrary, 'macrophage' which is the major immune cell population in the normal lung, showed distinct spatial distribution compared with alveolar epithelial cells or endothelial cells (Supplementary Figure S15). Different patterns of correlation of immune cells in two different tissues were possibly due to immune cells not having compartmentalized patterns of distribution and showing low cellular proportion in the normal lung tissue (47).

Meanwhile, when the top 7 highly localized cell types in each tissue compartment were listed, the selected cell types were shared between the lung 1 and lung 2 tissues (Supplementary Table S5). More specifically, for the alveolar space tissue domain, three alveolar epithelial cells ('alveolar epithelial type 1, 2' and 'signaling alveolar epithelial type 2') and two capillary cells ('capillary intermediate 2 and 'capillary aerocyte') overlapped in both tissues. The cell types were also shared in the fibrous stroma ('fibromyocyte', 'mesothelial' and 'myofibroblast'), immune cluster ('B', 'OLR1 + classical monocyte', 'EREG + dendritic', and 'plasmacytoid dendritic'), and terminal bronchiole ('proximal basal', 'proximal ciliated', 'differentiating basal' and 'ciliated') tissue domains. In short, CellDART can accurately assign prevalent cell types in the tissue compartments and is reproducible across replicates of the tissue. Considering the heterogeneous cell types in the lung, our resource of spatially resolved cell types derived from human lung tissue data provides the spatial distribution of cell types and may be used as controls to analyze pathologic patterns of various lung diseases.

There are several important issues to consider before applying CellDART to transfer cell labels. First, the density of cells may vary in the different regions of the tissue. In our method, the sampled number of cells in a pseudospot is fixed during the training; however, the domain adaptation process aligns the pseudospot to the spatial data, and the impact of spatial cell density variance on the result may be attenuated. In addition, the small population of cells in the spatial data may be neglected during the prediction of the cell proportion. The proportion of those cell types can be masked due to other predominant cell types in the spatial spots. In that case, CellDART can be implemented for the corresponding subpopulation of cells by extracting the marker genes for the subclusters. Lastly, CellDART adopts a shared feature embedder between pseudospots and real spots in the adversarial domain adaptation. There have been several deep learning-based domain adaptation approaches that have differences in feature embedders as well as training methods (48). Notably, our approach used the shared feature embedder considering the similar features representing cell types between pseudospots and real spots. Though CellDART showed robust performance with stable training results, there could be room for improvement by finding optimized domain adaptation methods among various recently developed approaches.

In conclusion, CellDART is capable of estimating the spatial cell compositions in complex tissues with high levels of heterogeneity by aligning the domain of single-cell and spatial transcriptomics data. The suggested method may help elucidate the spatial interaction of various cells in close proximity and track the cell-level transcriptomic changes while preserving the spatial context.

## DATA AVAILABILITY

Seven publicly available datasets were utilized in this study. First, mouse brain spatial data was acquired from https://support.10xgenomics.com/spatial-gene-expression/datasets and single-cell data from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE115746. Second, human DLPFC spatial data was downloaded from http://research.libd.org/spatialLIBD/ and single-cell data from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE144136. Third, mouse hippocampus Slide-seq data was downloaded from https://singlecell.broadinstitute.org/single_cell/study/SCP815 and single-cell data from http://dropviz.org/. Lastly, human normal lung single-cell data was obtained from https://www.synapse.org/#!Synapse:syn21041850.

The human lung spatial data was acquired from the normal lung tissue of a lung cancer patient in Seoul National University Hospital with informed consent. The data was uploaded in Gene Expression Omnibus (GSE172416).

Python source code and R wrapper function for CellDART are uploaded on https://github.com/mexchy1000/CellDART.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

*Author's contributions:* H.C. and Y.T.K. conceived the study. S.B., K.J.N. and H.C. designed the experiments. K.J.N. and Y.T.K. mainly contributed to acquiring spatially resolved transcriptomic data. H.C. initially developed a new algorithm and S.B. modified and optimized the analytic method. H.C., D.S.L. and Y.T.K. devised the data analysis. S.B. mainly conducted the data analysis. J.K. conducted histologic image analysis. All authors contributed to the interpretation of the data and wrote the paper.

## FUNDING

National Research Foundation of Korea grant funded by the Korean government [NRF-2020M3A9B6037195, NRF-2020M3A9B6038086, NRF- 2017M3C7A1048079,

NRF-2020R1A2C2101069]. Funding for open access charge: National Research Foundation of Korea grant funded by the Korean government [NRF-2020M3A9B6037195, NRF-2020M3A9B6038086, NRF-2017M3C7A1048079, NRF-2020R1A2C2101069].
*Conflict of interest statement.* None declared.

## REFERENCES

1. Waylen,L.N., Nim,H.T., Martelotto,L.G. and Ramialison,M. (2020) From whole-mount to single-cell spatial assessment of gene expression in 3D. *Commun. Biol.*, **3**, 602.
2. Larsson,L., Frisen,J. and Lundeberg,J. (2021) Spatially resolved transcriptomics adds a new dimension to genomics. *Nat. Methods*, **18**, 15–18.
3. Stahl,P.L., Salmen,F., Vickovic,S., Lundmark,A., Navarro,J.F., Magnusson,J., Giacomello,S., Asp,M., Westholm,J.O., Huss,M. *et al.* (2016) Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, **353**, 78–82.
4. Rodriques,S.G., Stickels,R.R., Goeva,A., Martin,C.A., Murray,E., Vanderburg,C.R., Welch,J., Chen,L.M., Chen,F. and Macosko,E.Z. (2019) Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, **363**, 1463–1467.
5. Vickovic,S., Eraslan,G., Salmen,F., Klughammer,J., Stenbeck,L., Schapiro,D., Aijo,T., Bonneau,R., Bergenstrahle,L., Navarro,J.F. *et al.* (2019) High-definition spatial transcriptomics for in situ tissue profiling. *Nat. Methods*, **16**, 987–990.
6. Stickels,R.R., Murray,E., Kumar,P., Li,J., Marshall,J.L., Di Bella,D.J., Arlotta,P., Macosko,E.Z. and Chen,F. (2021) Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.*, **39**, 313–319.
7. Cho,C.S., Xi,J., Si,Y., Park,S.R., Hsu,J.E., Kim,M., Jun,G., Kang,H.M. and Lee,J.H. (2021) Microscopic examination of spatial transcriptome using Seq-Scope. *Cell*, **184**, 3559–3572.
8. Berglund,E., Maaskola,J., Schultz,N., Friedrich,S., Marklund,M., Bergenstrahle,J., Tarish,F., Tanoglidi,A., Vickovic,S., Larsson,L. *et al.* (2018) Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat. Commun.*, **9**, 2419.
9. Thrane,K., Eriksson,H., Maaskola,J., Hansson,J. and Lundeberg,J. (2018) Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. *Cancer Res.*, **78**, 5970–5979.
10. Carlberg,K., Korotkova,M., Larsson,L., Catrina,A.I., Stahl,P.L. and Malmstrom,V. (2019) Exploring inflammatory signatures in arthritic joint biopsies with spatial transcriptomics. *Sci. Rep.*, **9**, 18975.
11. Chen,W.T., Lu,A., Craessaerts,K., Pavie,B., Sala Frigerio,C., Corthout,N., Qian,X., Lalakova,J., Kuhnemund,M., Voytyuk,I. *et al.* (2020) Spatial transcriptomics and in situ sequencing to study alzheimer's disease. *Cell*, **182**, 976–991.
12. Ji,A.L., Rubin,A.J., Thrane,K., Jiang,S., Reynolds,D.L., Meyers,R.M., Guo,M.G., George,B.M., Mollbrink,A., Bergenstrahle,J. *et al.* (2020) Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell*, **182**, 497–514.
13. Pham,D.T., Tan,X., Xu,J., Grice,L.F., Lam,P.Y., Raghubar,A., Vukovic,J., Ruitenberg,M.J. and Nguyen,Q.H. (2020) stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. bioRxiv doi: https://doi.org/10.1101/2020.05.31.125658, 31 May 2020, preprint: not peer reviewed.
14. Dries,R., Zhu,Q., Dong,R., Eng,C.H.L., Li,H., Liu,K., Fu,Y., Zhao,T., Sarkar,A. and Bao,F. (2021) Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol.*, **22**, 78.
15. Dagogo-Jack,I. and Shaw,A.T. (2018) Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.*, **15**, 81–94.
16. Hie,B., Bryson,B. and Berger,B. (2019) Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nat. Biotechnol.*, **37**, 685–691.
17. Stuart,T., Butler,A., Hoffman,P., Hafemeister,C., Papalexi,E., Mauck,W.M. 3rd, Hao,Y., Stoeckius,M., Smibert,P. and Satija,R. (2019) Comprehensive integration of single-cell data. *Cell*, **177**, 1888–1902.
18. Welch,J.D., Kozareva,V., Ferreira,A., Vanderburg,C., Martin,C. and Macosko,E.Z. (2019) Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, **177**, 1873–1887.
19. Andersson,A., Bergenstrahle,J., Asp,M., Bergenstrahle,L., Jurek,A., Fernandez Navarro,J. and Lundeberg,J. (2020) Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun. Biol.*, **3**, 565.
20. Kleshchevnikov,V., Shmatko,A., Dann,E., Aivazidis,A., King,H.W., Li,T., Elmentaite,R., Lomakin,A., Kedlian,V., Gayoso,A. *et al.* (2022) Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.*, https://doi.org/10.1038/s41587-021-01139-4.
21. Cable,D.M., Murray,E., Zou,L.S., Goeva,A., Macosko,E.Z., Chen,F. and Irizarry,R.A. (2021) Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat. Biotechnol.*, https://doi.org/10.1038/s41587-021-00830-w.
22. Dong,R. and Yuan,G.C. (2021) SpatialDWLS: accurate deconvolution of spatial transcriptomic data. *Genome Biol.*, **22**, 145.
23. Elosua-Bayes,M., Nieto,P., Mereu,E., Gut,I. and Heyn,H. (2021) SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res.*, **49**, e50.
24. Song,Q. and Su,J. (2021) DSTG: deconvoluting spatial transcriptomics data through graph-based artificial intelligence. *Brief Bioinform.*, **22**, bbaa414.
25. Dai Yang,K., Belyaeva,A., Venkatachalapathy,S., Damodaran,K., Katcoff,A., Radhakrishnan,A., Shivashankar,G.V. and Uhler,C. (2021) Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nat. Commun.*, **12**, 31.
26. Zhou,X., Chai,H., Zeng,Y., Zhao,H. and Yang,Y. (2021) scAdapt: virtual adversarial domain adaptation network for single cell RNA-seq data classification across platforms and species. *Brief Bioinform.*, **22**, bbab281.
27. Tzeng,E., Hoffman,J., Saenko,K. and Darrell,T. (2017) In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 7167–7176.
28. Maynard,K.R., Collado-Torres,L., Weber,L.M., Uytingco,C., Barry,B.K., Williams,S.R., Catallini,J.L. 2nd, Tran,M.N., Besich,Z., Tippani,M. *et al.* (2021) Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.*, **24**, 425–436.
29. Nagy,C., Maitra,M., Tanti,A., Suderman,M., Theroux,J.F., Davoli,M.A., Perlman,K., Yerko,V., Wang,Y.C., Tripathy,S.J. *et al.* (2020) Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nat. Neurosci.*, **23**, 771–781.
30. Tasic,B., Yao,Z., Graybuck,L.T., Smith,K.A., Nguyen,T.N., Bertagnolli,D., Goldy,J., Garren,E., Economo,M.N., Viswanathan,S. *et al.* (2018) Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, **563**, 72–78.
31. Saunders,A., Macosko,E.Z., Wysoker,A., Goldman,M., Krienen,F.M., de Rivera,H., Bien,E., Baum,M., Bortolin,L., Wang,S. *et al.* (2018) Molecular diversity and specializations among the cells of the adult mouse brain. *Cell*, **174**, 1015–1030.
32. Travaglini,K.J., Nabhan,A.N., Penland,L., Sinha,R., Gillich,A., Sit,R.V., Chang,S., Conley,S.D., Mori,Y., Seita,J. *et al.* (2020) A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature*, **587**, 619–625.
33. Wolf,F.A., Angerer,P. and Theis,F.J. (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, **19**, 15.
34. Hawrylycz,M.J., Lein,E.S., Guillozet-Bongaarts,A.L., Shen,E.H., Ng,L., Miller,J.A., van de Lagemaat,L.N., Smith,K.A., Ebbert,A., Riley,Z.L. *et al.* (2012) An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, **489**, 391–399.
35. Lake,B.B., Ai,R., Kaeser,G.E., Salathia,N.S., Yung,Y.C., Liu,R., Wildberg,A., Gao,D., Fung,H.L., Chen,S. *et al.* (2016) Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*, **352**, 1586–1590.
36. Habib,N., Avraham-Davidi,I., Basu,A., Burks,T., Shekhar,K., Hofree,M., Choudhury,S.R., Aguet,F., Gelfand,E., Ardlie,K. *et al.* (2017) Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods*, **14**, 955–958.

37. Cembrowski,M.S., Wang,L., Sugino,K., Shields,B.C. and Spruston,N. (2016) Hipposeq: a comprehensive RNA-seq database of gene expression in hippocampal principal neurons. *Elife*, **5**, e14997.
38. Ero,C., Gewaltig,M.O., Keller,D. and Markram,H. (2018) A cell atlas for the mouse brain. *Front Neuroinform*, **12**, 84.
39. Valerio-Gomes,B., Guimaraes,D.M., Szczupak,D. and Lent,R. (2018) The absolute number of oligodendrocytes in the adult mouse brain. *Front Neuroanat*, **12**, 90.
40. Refaeli,R., Doron,A., Benmelech-Chovav,A., Groysman,M., Kreisel,T., Loewenstein,Y. and Goshen,I. (2021) Features of hippocampal astrocytic domains and their spatial relation to excitatory and inhibitory neurons. *Glia*, **69**, 2378–2390.
41. Yao,Z., van Velthoven,C.T.J., Nguyen,T.N., Goldy,J., Sedeno-Cortes,A.E., Baftizadeh,F., Bertagnolli,D., Casper,T., Chiang,M., Crichton,K. *et al.* (2021) A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell*, **184**, 3222–3241.
42. Bayraktar,O.A., Bartels,T., Holmqvist,S., Kleshchevnikov,V., Martirosyan,A., Polioudakis,D., Ben Haim,L., Young,A.M.H., Batiuk,M.Y., Prakash,K. *et al.* (2020) Astrocyte layers in the mammalian cerebral cortex revealed by a single-cell in situ transcriptomic map. *Nat. Neurosci.*, **23**, 500–509.
43. Zonouzi,M., Berger,D., Jokhi,V., Kedaigle,A., Lichtman,J. and Arlotta,P. (2019) Individual oligodendrocytes show bias for inhibitory axons in the neocortex. *Cell Rep.*, **27**, 2799–2808.
44. Herrero-Navarro,A., Puche-Aroca,L., Moreno-Juan,V., Sempere-Ferrandez,A., Espinosa,A., Susin,R., Torres-Masjoan,L., Leyva-Diaz,E., Karow,M., Figueres-Onate,M. *et al.* (2021) Astrocytes and neurons share region-specific transcriptional signatures that confer regional identity to neuronal reprogramming. *Sci. Adv.*, **7**, eabe8978.
45. Lanjakornsiripan,D., Pior,B.J., Kawaguchi,D., Furutachi,S., Tahara,T., Katsuyama,Y., Suzuki,Y., Fukazawa,Y. and Gotoh,Y. (2018) Layer-specific morphological and molecular differences in neocortical astrocytes and their dependence on neuronal layers. *Nat. Commun.*, **9**, 1623.
46. Batiuk,M.Y., Martirosyan,A., Wahis,J., de Vin,F., Marneffe,C., Kusserow,C., Koeppen,J., Viana,J.F., Oliveira,J.F., Voet,T. *et al.* (2020) Identification of region-specific astrocyte subtypes at single cell resolution. *Nat. Commun.*, **11**, 1220.
47. Kim,N., Kim,H.K., Lee,K., Hong,Y., Cho,J.H., Choi,J.W., Lee,J.I., Suh,Y.L., Ku,B.M., Eum,H.H. *et al.* (2020) Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat. Commun.*, **11**, 2285.
48. Wang,M. and Deng,W.H. (2018) Deep visual domain adaptation: a survey. *Neurocomputing*, **312**, 135–153.