Joint dimension reduction and clustering analysis for single-cell RNA-seq and spatial transcriptomics data

Wei Liu[1,2], Xu Liao[2], Xiang Zhou[3*], Xingjie Shi[1*], Jin Liu[2*]

[1] Academy of Statistics and Interdisciplinary Sciences, East China Normal University

[2] Centre for Quantitative Medicine, Health Services & Systems Research, Duke-NUS Medical School

[3] Department of Biostatistics, University of Michigan

Abstract

Dimension reduction and (spatial) clustering are two key steps for the analysis of both single-cell RNA-sequencing (scRNA-seq) and spatial transcriptomics data collected from different platforms. Most existing methods perform dimension reduction and (spatial) clustering sequentially, treating them as two consecutive stages in tandem analysis. However, the low-dimensional embeddings estimated in the dimension reduction step may not necessarily be relevant to the class labels inferred in the clustering step and thus may impair the performance of the clustering and other downstream analysis. Here, we develop a computation method, DR-SC, to perform both dimension reduction and (spatial) clustering jointly in a unified framework. Joint analysis in DR-SC ensures accurate (spatial) clustering results and effective extraction of biologically informative low-dimensional features. Importantly, DR-SC is not only applicable for cell type clustering in scRNA-seq studies but also applicable for spatial clustering in spatial transcriptimics that characterizes the spatial organization of the tissue by segregating it into multiple tissue structures. For spatial transcriptoimcs analysis, DR-SC relies on an underlying latent hidden Markov random field model to encourage the spatial smoothness of the detected spatial cluster boundaries. We also develop an efficient expectation-maximization algorithm based on an iterative conditional mode. DR-SC is not only scalable to large sample sizes, but is also capable of optimizing the spatial smoothness parameter in a data-driven manner. Comprehensive simulations show that DR-SC outperforms existing clustering methods such as Seurat and spatial clustering methods such as BayesSpace and SpaGCN and extracts more biologically relevant features compared to the conventional dimension reduction methods such as PCA and scVI. Using 16 benchmark scRNA-seq datasets, we demonstrate that the low-dimensional embeddings and class labels estimated from DR-SC lead to improved trajectory inference. In addition, analyzing three published scRNA-seq and spatial transcriptomics data in three platforms, we show DR-SC can improve both the spatial and non-spatial clustering performance, resolving a

*Correspondence should be addressed to Xiang Zhou (xzhousph@umich.edu), Xingjie Shi(xjshi@fem.ecnu.edu.cn) and Jin Liu (jin.liu@duke-nus.edu.sg)

low-dimensional representation with improved visualization, and facilitate the downstream analysis such as trajectory inference.

Keywords: dimension reduction; clustering; expectation-maximization algorithm hidden Markov random field; spatial transcriptomics; scRNA-Seq

Introduction

Single-cell RNA sequencing (scRNA-seq) is a set of widely applied technologies that profile the transcriptome of individual cells on a large scale and could reveal cell subpopulations within a tissue [1, 2]. Spatial transcriptomics are a set of recently developed technologies that allow for simultaneous characterization of the expression profiles on multiple tissue locations while retaining their location information. scRNA-seq technologies include full-length transcript sequencing approaches (e.g., Smart-seq2 [3] and MATQ-seq [4]) and $3'/5'$-end transcript sequencing technologies (e.g., Drop-seq [5] and STRT-seq [6]). While spatial transcriptomics technologies include earlier fluorescence in situ hybridization (FISH)-based approaches (e.g., seqFISH [7] and MERFISH [8]), sequencing-based techniques (e.g., 10x Visium [9] and Slide-seq [10]), among others. Both scRNA-seq and spatially resolved transcriptomic technologies have provided unprecedented new opportunities for characterizing the cell type heterogeneity within a tissue, investigating the spatial gene expression patterns [11, 12], exploring the transcriptomic landscape of the tissue, as well as the characterization of the spatial distribution of cell types on the tissue across multiple tissue types [13–15].

In the analysis of both scRNA-seq and spatial transcriptomics datasets, dimension reduction and (spatial) clustering are two key analytic steps that are critical for many downstream analyses such as cell lineage analysis and differential expression analysis. Specifically, due to the curse of dimensionality, dimension reduction methods are usually applied to transform the original noisy expression matrix in either scRNA-seq or spatial transcriptomics into a low-dimensional representation before performing (spatial) clustering analysis [16–18]. In existing literature, many dimension reduction methods have been developed and common methods include principal component analysis (PCA), weighted PCA (WPCA) [19], t-distributed stochastic neighbor embedding (tSNE) [20], uniform manifold approximation and projection (UMAP) [21], etc. Among them, PCA is a well-recognized approach that is routinely used in many software for both scRNA-seq and spatial transcriptomics analyses [16, 22] and enjoys many desirable features such as simplicity, computational efficiency, and relative accuracy. For example, Seurat [23], SpaGCN [24], BayesSpace [25] and SC-MEB [26] all first extract the top PCs from the high-dimensional expression matrix and then perform (spatial) clustering analysis. WPCA is a variation of PCA that imposes different weights on different genes to upweight the potentially informative genes [19] in the presence of heteroscedastic noises. While fitting with PCA and WPCA are generally automatic without the requirement of parameter tuning, tSNE and UMAP are two widely used nonlinear dimension reduction methods that rely relatively heavily on manually tuned parameters for optimized performance [18, 22]. In addition to these generic methods, several dimension reduction methods have been developed to account for either the count nature and/or dropout events of scRNA-seq data, e.g., ZIFA [27] ZINB-WaVE [28], and scVI [29].

After obtaining the low-dimensional representation with dimension reduction, (spatial) clustering analyses are then carried out. Clustering in scRNA-seq data aims to identify cell types and cluster cells into distinct cell types. While spatial clustering in spatial transcriptomics aims to use the spatial transcriptomic information to cluster the spatial locations on the tissue

into multiple spatial clusters, effectively segmenting the entire tissue into multiple tissue structures or domains. Cell type clustering facilitates the understanding of the cell type composition of a tissue with potentially heterogeneous cell types. While spatial clustering facilitates the characterization of the tissue structure and is a key step towards understanding the spatial and functional organization of the tissue. Common clustering methods for scRNAseq analysis include k-means [30] and Gaussian mixture model (GMM) [31]. Common spatial clustering methods for spatial transcriptomics analysis include the graph convolutional network (GCN)-based approach SpaGCN [24], the hidden Markov random field model implemented in the Giotto package [32], BayesSpace [25], and SC-MEB [26], all of which promote smoothness of cluster assignment in neighboring tissue locations. By performing dimension reduction and (spatial) clustering sequentially, the estimated low-dimensional embeddings and class labels could be used for many types of downstream analyses, such as cell lineage analysis [33–36] and differential gene expression analysis [37].

The majority of existing methods for dimension reduction and (spatial) clustering have been considered as tandem analysis by first performing dimension reduction for expression matrix followed by (spatial) clustering analysis on the estimated low-dimensional embeddings [18], as shown in Fig. 1a. Performing dimension reduction and (spatial) clustering in two sequential analytic steps is not ideal for two important reasons. First, these tandem methods optimize distinct loss functions for dimension reduction and (spatial) clustering separately, and the two loss functions may not be consistent with each for achieving optimal (spatial) cluster allocation [38]. PCA aims to retain as much variance as possible in as few PCs as possible, whereas spatial clustering aims to either minimize within-cluster variances or maximize between-cluster variances. Second, the dimension-reduction step in the tandem methods does not consider the uncertainty in obtaining low-dimensional features. Consequently, the extracted low-dimensional components are effectively treated as error free in the spatial clustering analysis, which is not desirable. Recognizing the drawbacks of tandem analysis, several recent methods have been developed in other research areas for performing joint dimension reduction and clustering analysis. For example, an ad-hoc remedy would iteratively perform two steps of analyses: estimate low-dimensional embeddings by applying supervised dimension reduction together with the inferred latent class labels (dimension-reduction step); infer class labels using the estimated embeddings (clustering step). These simple procedures echo some recent literature in self-supervised learning [39, 40], where deep neural networks combined with simple classifiers were utilized to perform unsupervised clustering for image data. To some extent, joint methods perform self-learning to classify all spots and obtain latent features iteratively. However, it is still challenging to unify existing methods, combining both dimension-reduction and spatial-clustering steps, in a self-learning manner.

Here, we propose a unified and principled method to both estimate low-dimensional embeddings relevant to latent class labels and, in the case of spatial transcriptomics analysis, further leverage these embeddings with spatial information to perform spatial clustering using a hidden Markov random field (HMRF). The proposed method is built on a hierarchical model with two layers as shown in Fig. 1b: the first layer relates gene expression with low-dimensional embeddings as a Dimension-Reduction step, while the second layer relates latent embeddings with cluster labels, and, if necessary, spatial information as a Spatial-Clustering step (DR-SC). These two layers are unified in DR-SC such that relevant features are estimated while the spatial clustering is performed simultaneously. We develop an efficient expectation-maximization (EM) algorithm based on an iterative conditional mode (ICM) [41, 42]. DR-SC is not only computationally efficient and scalable to large sample sizes, but is also capable of optimizing the smoothness parameter in the spatial clustering component. Importantly, when the smoothness

parameter is set to be zero, DR-SC directly performs clustering for scRNA-seq data with no spatial information. Unlike existing spatial clustering approaches, DC-SR can determine the number of clusters in an automatic fashion using a modified Bayesian information criterion (MBIC) [43]. Using 16 benchmark scRNA-seq datasets (Additional file 1), we demonstrate that the low-dimensional embeddings and class labels estimated from DR-SC lead to better performance in the downstream lineage analysis using Slingshot [35]. We further illustrate that DR-SC achieves higher spatial clustering accuracy, and resolves a low-dimensional representation with improved visualization using both CITE-seq (Additional file 1) and spatial transcriptomics (10x Visium and Slide-seqV2) datasets. To exemplify the utility of the estimated low-dimensional embeddings from DR-SC, we performed analysis to infer cell lineage using a seqFISH dataset from mouse embryo. The R package DR.SC is available on CRAN `https://CRAN.R-project.org/package=DR.SC`, with functions implemented for standalone analysis and Seurat [44] based pipeline analyses.

Results

DR-SC simultaneously performs dimension reduction and clustering with/without spatial information

We propose DR-SC to estimate the low-dimensional latent features while improve the clustering performance via a unified statistically principled method. DR-SC relates a two-layer hierarchical model to simultaneously perform dimension reduction via a probabilistic PCA model and promote spatial clustering using a hidden Markov random field (HMRF) based on empirical Bayes. In spatial transcriptomics datasets, we observe p-dimensional log-normalized expression vector $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^T$ for each spot $s_i \in \mathbb{R}^2$ on square or hexagonal lattices, while its class label $y_i \in \{1, \cdots, K\}$ and q-dimensional embeddings $\mathbf{z}_i$ are not available. Without loss of generality, we assume that $\mathbf{x}_i$ is centered and DR-SC models the centered log-normalized expression vector $\mathbf{x}_i$ with its latent low-dimensional feature $\mathbf{z}_i$ and class label y_i as

$$\mathbf{x}_i = W\mathbf{z}_i + \boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_i \sim N(0, \Lambda), \tag{1}$$

$$\mathbf{z}_i | y_i = k \sim N(\mu_k, \Sigma_k), \tag{2}$$

where $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_p)$ is a diagonal matrix for residual variance, $W \in \mathbb{R}^{p\times q}$ is a loading matrix that transforms the p-dimensional expression vector into q-dimensional embeddings, and $\mu_k \in \mathbb{R}^{q\times 1}$ and $\Sigma_k \in \mathbb{R}^{q\times q}$ are the mean vector and covariance matrix for the kth class, respectively. Eqn. (1) relates the high-dimensional expression vector ($\mathbf{x}_i$) in p genes with a low-dimensional feature ($\mathbf{z}_i$) via a probabilistic PCA model while Eqn. (2) is a Gaussian mixture model (GMM) for this latent feature among all n spots. When spatial coordinates (s_i) are available, we assume each latent class label y_i are interconnected with class labels of their neighborhoods via a Markov random field. To promote spatial smoothness within spot neighborhoods, we assume that the hidden Markov random field $\mathbf{y} = (y_1, \cdots, y_n)^T$ takes the following Potts model [45],

$$\Pr(\mathbf{y}) = C(\beta)^{-1} \exp\{-\frac{1}{2} \sum_i \sum_{i' \in N_i} \beta(1 - \delta(y_i, y_{i'}))\}, \tag{3}$$

where δ is a Dirac function, $C(\beta)$ is a normalization constant that does not have a closed form, N_i is the neighborhood of spot i, and β is the smoothing parameter that controls the similarity

among the neighboring labels, in other words, the degree of spatial smoothness. When this smoothing parameter β goes to zero, the spatial-clustering step in DR-SC, Eqn. (2) and (3), reduces to a latent GMM with no spatial information.

DR-SC unifies both models for dimension reduction and (spatial) clustering (Fig. 1b). By combining the latent GMM in Eqn. (2) and the Potts model in Eqn. (3), DR-SC performs the spatial clustering on low-dimensional embeddings obtained from a probabilistic PCA model in Eqn. (1). Conventionally, embeddings obtained using unsupervised dimension reduction methods, such as PCA, UMAP and tSNE, reflect variations from different sources including batch effects, microenvironments among observed cells/spots, among others, other than cell-type difference. Thus, embeddings from unsupervised dimension reduction analysis may distort the downstream clustering for cell typing [46]. In contrast, DR-SC performs dimension reduction in a self-learning manner, where embeddings, $\mathbf{z}_i$s, are estimated under the supervision of the estimated latent labels for each spot (Fig. 1b). Thus, the obtained embeddings capture information with regard to biological differences, e.g., cell-type or cell-state differences, which in turn improve spatial clustering for cell typing. When no spatial information is available, e.g., scRNA-seq, we could simply apply a latent GMM (2) without considering the Potts model (3). In the later sections, we will show the improved clustering performance in spatial transcriptomics datasets from different platforms. In Additional file 1, we will show DR-SC could improve the clustering performance for single-cell datasets.

Aside from improving the (spatial) clustering performance, the estimated low-dimensional embeddings based on DR-SC could be used in different types of downstream analyses (Figure 1c). First, the estimated embeddings could be used to better visualize the clustering among cells/spots. Second, the performance of trajectory inference could be improved since the reduced dimensional space from DR-SC possesses more relevant information regarding cell clusters. Third, by taking these estimated embeddings as covariates, we could perform hypothesis testing to identify genes with pure spatial variations but no cell-type differences. These genes could be related to cell morphology or tissue types rather than differentially expressed across cell types. In Additional file 1, we will compare the accuracy of downstream lineage inference using the estimated embeddings and cell-type labels from DR-SC with other unsupervised dimension reduction methods in the application of 16 benchmark scRNA-seq datasets. In both later sections and Additional file 1, we will also show both the improved clustering performance and cluster visualization using DR-SC for both non-spatial (CITE-seq) and spatial transcriptomics (10x Visium, Slide-seqV2, and seqFISH) datasets. The basic information (number of spots/cell/genes and platforms) about the used spatial transcriptomics datasets were given in Table S1 (Additional file 2). By applying DR-SC to several spatial transcriptomics datasets, we will further show the utility of using the low-dimensional embeddings obtained from DR-SC to identify genes related to cell morphology and tissue types.

DR-SC improves clustering and estimation for low-dimensional features in simulations

We conducted simulation studies to evaluate the performance of DR-SC in comparison with existing dimension reduction and clustering methods. First, we simulated data with both non-spatial ($\beta = 0$) and spatial ($\beta = 1$) patterns, as well as with both homogeneous and heterogeneous residual variance λ_j (see Methods). Two simulation settings were considered. In Simulation 1, log-normalized and centered gene expressions were generated from Eqn. (1) – (3). In Simulation 2, we first generated a count matrix using Poisson distribution with over-dispersion, which could better mimic the count nature of scRNA-seq and 10x Visium

datasets (see Methods). Then, we performed log-transformation for the raw count matrix using the library size [47]. In all simulations, we set $p = 1,000$ and ran 50 replicates. The details of simulation settings were provided in the Methods Section.

To evaluate the clustering performance, we compared DR-SC with two groups of spatial/non-spatial clustering methods. The first group was tandem analysis using principal components (PCs) from either PCA or WPCA in the dimension-reduction step, while using SpaGCN [24], BayesSpace [48], SC-MEB [26], Giotto [32], Louvain [49], Leiden [50], Gaussian mixture model (GMM), and k-means in the clustering step. Among them, SpaGCN software took the log-normalized expression matrix as the input, used its internally embedded PCA algorithm to obtain PCs, and could only be applied to perform spatial clustering. The second group was joint analysis, including PSC [51] and FKM [38]. By setting the smoothing parameter to zero, BayesSpace, SC-MEB and Giotto could be applied to cluster non-spatial data. On the other hand, to evaluate the estimation accuracy of low-dimensional embeddings, we compared DR-SC with eight dimension reduction methods in all simulation settings, including PCA, weighted PCA (WPCA) [19], FKM [38], tSNE [20], UMAP [21], ZIFA [27], ZINB-WaVE [28], and scVI [29].

First, we compared the clustering performance of each method. For tandem analysis, we considered both PCA and WPCA to obtain low-dimensional embeddings in Simulation 1 (Fig. 2; Additional file 1: Fig. S9). In Simulation 2 (Additional file 1: Fig. S10), besides PCA and WPCA, we also applied ZINB-WaVE to obtain low-dimensional embeddings as the input for different clustering methods in tandem analysis. Since Giotto, k-means, FKM and PSC did not provide a data-driven way to select the number of clusters K, we evaluated their clustering performance using the true cluster number. Fig. 2a and Fig. S9a and S10a (Additional file 1) showed DR-SC achieved the best clustering performance and robust to both homogeneous and heterogeneous residual variances among methods that use the true cluster number. To select the number of clusters K, DR-SC and SC-MEB used a modified Bayesian information criteria (MBIC) [43, 52], GMM used Bayesian information criteria, BayesSpace adopted the average loglikelihood-maximization-based method in early iterations, Leiden and Louvain used a community-modularity-maximzing rule [49] and SpaGCN applied Louvain initialization for this aim [24]. Fig. 2b, Fig. S9b and S10b (Additional file 1) showed the clustering performance of methods that were capable of selecting the number of clusters K while Fig. 2c and Fig. S11b (Additional file 1) showed the barplots of the selected number of clusters for each method. Conventional PCA could not recover the underlying features in the presence of heteroscedastic noise while weighed PCA (WPCA) could be applied to give less informative genes less weights [19]. Thus, when there existed heterogeneous errors, the clustering performance of tandem analysis using conventional PCA was worse than that using WPCA. In all settings, the clustering performance of DR-SC was robust in both homogeneous and heterogeneous cases. Importantly, DR-SC achieved the highest ARI values among all methods here. Moreover, we observed that only DR-SC could correctly choose the number of clusters while BayesSpace tended to overestimate in the non-spatial case and underestimate in the spatial case. This is because BayesSpace fixed the smoothing parameter rather than updated it in a data-driven manner. Thus, the selection of the number of clusters for BayesSpace was sensitive to the choice of the smoothing parameter (Additional file 1: Fig. S11c). In contrast, other methods tended to have a similar pattern across both non-spatial and spatial cases.

Next, we evaluated the performance of DR-SC in estimating the low-dimensional embeddings. Fig. 2d and Fig. S11a (Additional file 1) showed the average canonical correlation between the estimated embeddings $\hat{\mathbf{z}}_i$ and the true latent features $\mathbf{z}_i$. We observed that DR-SC has the highest canonical correlation coefficients, suggesting that the estimated embeddings are

more accurate. In addition, Fig. S9c and Fig. S11a (Additional file 1) showed the Pearson correlation coefficients between the observed expressions $\mathbf{x}_i$ and the estimated cell-type labels $\hat{y}_i$ conditioned on embeddings from DR-SC was much smaller than those from other methods, suggesting that DR-SC captures more relevant information regarding cell types and thus could facilitate the downstream analysis.

In addition, we evaluated the corresponding computational time of each method in all simulation settings. Fig. S9a, S9b and Fig. S10 bottom panel (Additional file 1) showed the boxplots of computational time for each method. Louvain and SpaGCN were the fastest while BayesSpace was the slowest. Moreover, Fig. S9d (Additional file 1) showed DR-SC was computationally efficient and scalable to large sample sizes, where it only took around 30 mins to analyze a data with 1,000 genes and 100,000 spots.

Human dorsolateral prefrontal cortex data

As an emerging spatial transcriptomics technology, 10x Visium assay displays improvements in both resolution and time to run the protocol [53]. Maynard et al. [13] used this technology to generate spatial maps of gene expression matrices for the six-layered human dorsolateral prefrontal cortex (DLPFC) of the adult human brain, and manually annotated Visium spots based on cytoarchitecture. In this dataset, there were 12 tissue sections from three adult donors with a median depth of 291 million reads for each sample, median of 3,844 spots per tissue section and a mean of 1,734 genes per spot. Raw gene expression count matrices were log-transformed and normalized using the library size [47].

In this analysis, we considered both joint and tandem methods for dimension reduction and clustering. In joint methods, we took the log-transformed raw count matrix using the library size as the input while in tandem analysis, we obtained top 15 PCs from either PCA or WPCA as the input for different clustering methods. Since Giotto, k-means, FKM and PSC could not choose the number of clusters K, we fixed the number of clusters using manual annotations to make comparisons with DR-SC. Fig. 3a showed the clustering performance using the ARI values for all 12 tissue sections when K was fixed with values from their manual annotations. Spatial clustering methods, i.e., DR-SC and Giotto, outperformed those without considering spatial information while DR-SC performed much better than Giotto. We further compared the clustering performance for all methods that were capable of selecting the number of cluster as shown in Fig. 3b. We also observed that spatial clustering methods such as DR-SC, SpaGCN, SC-MEB and BayesSpace outperformed the non-spatial ones such as GMM, Leiden and Louvain. Note that there were only minor differences for DR-SC using either the fixed K or the chosen K. We also evaluated the clustering performance using the normalized mutual information (NMI) (Additional file 1: Fig. S12a) and similar patterns could be observed. Fig. 3c, respectively, showed the heatmap of cell types from manual annotations, and the heatmaps of clustering assignments across spatial and non-spatial clustering methods for sample ID 151510. The results for the rest 11 samples were provided in Fig. S13a-S23a (Additional file 1). In addition, Fig. 3d and Fig. S13b-S23b (Additional file 1) showed the tSNE plots for DR-SC and other three dimension reduction methods (PCA, WPCA, UMAP), where tSNE PCs were obtained from the estimated 15-dimensional features of each method with class labels estimated in DR-SC. We observed better separation of tSNE PCs from DR-SC. Moreover, we evaluated the computational efficiency of DR-SC in comparison with other methods in Fig. S12b (Additional file 1), showing that DR-SC was about 10 times faster than FKM, PSC and BayesSpace.

We further performed conditional analysis to investigate the role of spatial variable genes

(SVGs) beyond cell-type differences. Using SPARK [11], we performed spatial variation analysis (SVA) with the embeddings estimated by DR-SC as covariates. The detailed gene list identified at an FDR of 1% was given in Table S2 (Additional file 2). Compared with the gene list identified by SVA without covariates, the number of SVGs dramatically decreased from 1,583 to 113 at an FDR of 1% on average over 12 tissues after adjusting for covariates. Without adjusting for cell-type-relevant covariates, genes identified in SVA could simply reflect the variations primarily due to cell-type differences. Fig. S24a (Additional file 1) showed the Venn diagram between SVGs without adjusting for cell-type-relevant covariates and differentially expressed (DE) genes in cell types, where we observed that the majority of DE genes in cell types were also identified as SVGs without adjusting for covaraites. Fig. S24b&c (Additional file 1) showed the barplots of the proportion of DE genes overlapped with SVGs without/with adjusting for cell-type-relevant covariates, respectively. The overlap proportion reduced substantially by performing conditional spatial variation analysis, suggesting these identified genes may genuinely spatially expressed beyond merely variations between cell types.

Next, we performed functional enrichment analysis for SVGs adjusted for cell-type-relevant covariates. A total of 82 terms from Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Human Protein Atlas (HPA) were enriched with adjusted p-values less than 0.05 in at least three DLPFC tissue sections. Fig. S25 (Additional file 1) showed the top five pathways among all 12 DLPFC tissue sections, where we observed many common terms could be identified after controlling cell-type-relevant covariates. This result suggested that SVGs adjusted for cell-type-relevant covariates shared common spatial patterns in the brain tissue. For example, the same set of top significant HPA terms were identified in 8 out of 12 tissue samples, including process in white matter, process in granular layer and cytoplasm/membrane (Additional file 1: Fig. S25). Nearly all top significant KEGG pathways were identified in all 12 samples, including Huntington disease, Alzheimer disease, Parkinson disease and so forth. Several studies [54, 55] reported a common shared characteristics of abnormal proteins across the brain to cause damage in Alzheimer's disease, Parkinson's disease, Huntington's disease, and other neurodegenerative diseases. Additionally, many common significant GO terms were identified in all 12 samples, such as electron transfer activity, structural molecule activity, structural constituent of cytoskeletion, oxidative phosphorylation, endocytic vesicle lumen, and respiratory chain complex. Details of the top five pathways for all 12 tissue samples were presented in Table S3 (Additional file 2).

Mouse olfactory bulb data

Slide-seq is another spatial transcriptomics technology that simultaneously decouples the imaging from molecular sampling and quantifies expression across the genome with 10-μm spatial resolution [10]. To further improve the magnitude of sensitivity and enable more efficient recovery of gene expression, Slide-seqV2 was recently introduced [56] together with two datasets from mouse olfactory bulb and mouse cortex, respectively. We present the analysis of mouse olfactory bulb dataset in this section and that of mouse cortex dataset in the next one. The olfactory bulb contains 21,041 spots and 37,329 genes with a median of 494 unique molecular identifiers (UMIs) per bead. Raw gene expression count matrices were log-transformed and normalized using the library size.

In the analysis, we considered spatial clustering methods including BayesSpace, SC-MEB, SpaGCN and DR-SC. Among them, all other methods except for DR-SC were based on tandem analysis. Thus, BayesSpace and SC-MEB took top 15 PCs from normalized expression matrix of SVGs as their inputs (see Methods) while SpaGCN package took the normalized expression

matrix as the input and used its internally embedded PCA algorithm to obtain PCs. As a joint method, DR-SC took the normalized expression matrix of SVGs as the input. Fig. 4a showed the spatial heatmap of clustering assignments across four methods while Fig. 4b showed the tSNE plots for these four methods, where tSNE PCs of DR-SC were obtained from its estimated 15-dimensional latent features while tSNE PCs of BayesSpace-O, SC-MEB-O and SpaGCN were based on 15 PCs from PCA. We observed better visualization of cell types using tSNE PCs from DR-SC. We also compared the running time of these methods in Fig. S26a, where DR-SC and BayesSpace, respectively, took 1,217 and 13,193 secs to complete the analysis for all 21,041 spots. To further compare the visualization of different dimension reduction methods, we first applied other eight dimension reduction methods to extract latent features and further obtained two-dimensional tSNE PCs based on each of individual estimated latent ones. In Fig. 4c and Fig. S26b, we visualized the two-dimensional tSNE PCs from different dimension reduction methods with cluster labels estimated in DR-SC, where tSNE PCs from DR-SC were more separable than those from other methods.

Using cluster labels estimated in DR-SC, we performed differential gene expression (DGE) analysis to identify the marker genes for each cluster. Fig. 4d displayed the heatmap of DE genes for each cell type, showing good separation across different cell types. By checking PanglaoDB [57] for the identified marker genes, we were able to identify seven cell types in Fig. 4d, including two major neuron cell types, Purkinje neurons and interneurons consisting of 52% and 10% spots, respectively. The primary output signal of Purkinje cells is the modulated discharge of simple spikes while interneurons could potentially contribute to the modulation of simple spikes [58].

Then, we performed spatial variational analysis using SPARK by controlling the 15-dimensional embeddings estimated by DR-SC. In total, 518 SVGs were identified at an FDR of 1%. The identified gene list was given in Table S4 (Additional file 2). Next, we performed functional enrichment analysis (FEA) for these SVGs. A total of 385 of Gene Ontology (GO) terms were enriched with adjusted p-value less than 0.05. Figure 4e showed the bubble plot for this functional enrichment, where we observed the nervous system development related pathways were enriched in olfactory bulb.

Moreover, we applied Slingshot [35] to perform cell lineage analysis using low-dimensional embeddings and cluster labels estimated by DR-SC. Srivatsan et al. [59] reported that the neuron cells are differentiated after glia cells. To check this result, we focused on studying neuron cells (Purkinje neurons and interneurons) and glia cells (astrocytes and oligodendrocytes) to infer their differentiation trajectory. Fig. S26c showed the inferred trajectory, where we observed that Purkinje neurons differentiated after oligodendrocytes while interneurons was after astrocytes. Additionally, Fig. S26c showed the heatmap of expression levels of top 20 significant genes presenting dynamic expression patterns over pseudotime. From this figure, we observed some genes presented interesting dynamic patterns of expressions, varying from high level to low level and then back to high level, such as *Camk2b* and *Malat1*. Küry et al. [60] reported *Camk2b* was important for learning and synaptic plasticity in mice, while Zhang et al. [61] reported a potential cis-regulatory role of *Malat1* gene transcription in mice.

Mouse E15 neocortex data

The mouse E15 neocortex data from Slide-seqV2 platform contains 33,611 spots and 22,683 genes resolved spatially with their expressions in E15 embryo section. Similarly, we performed spatial clustering by DR-SC in comparison with BayesSpace, SC-MEB and SpaGCN. BayesSpace and SC-MEB took top 15 PCs from normalized expression matrix of SVGs as their inputs (see

Methods) while SpaGCN and DR-SC took the normalized expression matrix as their inputs. Fig. 5a showed the spatial heatmap of clustering assignments across four methods while Fig. 5b showed the tSNE plots for four methods, where tSNE PCs of DR-SC were obtained from its estimated 15-dimensional latent features while tSNE PCs of BayesSpace-O, SC-MEB-O and SpaGCN were based on 15 PCs from PCA. We observed that SC-MEB-O and DR-SC shared similar spatial pattern while a large proportion of spots were assigned to a single cluster (78%) in BayesSpace-O. Moreover, we observed better visualization of clusters using tSNE PCs from DR-SC but not PCA. We also compared the running time of these methods in Fig. S27a, where DR-SC and BayesSpace, respectively, took 1,407 and 18,740 secs to finish the analysis for all 33,611 spots.To further compare the visualization of different dimension reduction methods, we first applied other eight dimension reduction methods to extract latent features and further obtained two-dimensional tSNE PCs based on each of individual estimated latent ones. In Fig 5c and Fig. S27b, we visualized the two-dimensional tSNE PCs from different dimension reduction methods with cluster labels estimated in DR-SC, suggesting DR-SC performed better for visualization.

Based on the cluster labels estimated in DR-SC, we performed DGE analysis to identify the marker genes for each cluster. Fig. 5d displayed the heatmap of DE genes for each cell type, showing good separation across different cell types. By checking PanglaoDB [57] for the identified marker genes, we were able to identify five cell types in Fig. 5d, including two major neuron-related cell types, neurons and neural stem/precursor cells consisting of 40% and 29% spots, respectively.

Next, we applied Slingshot [35] to infer the differentiation lineage of E15 neocortex cells based on the low-dimensional features and cluster labels estimated by DR-SC. Fig. 5e showed the inferred development trajectory of different types of cells and the heatmap of the top 20 significant dynamic expressed genes along the trajectory. We also observed the differentiation of neuron cells (neurons and neural stem/precursor cells) after glia cells (astrocytes, oligodendrocytes and oligodendrocyte progenitor cells). For example, part of neurons and all neural stem/precursor cells differentiated after astrocytes, and the other part of neurons differentiated after oligodendrocytes and oligodendrocyte progenitor cells. From the heatmap, we observed some genes presenting interesting dynamic patterns of expression. For instance, genes *Ttr* had lower expression levels at the initial stage and later their expression levels increased substantially. Buxbaum et al. [62] found that *Ttr* has an effect on the development of the neuropathologic and behavioral phenotypes in mice. In contrast, the expression levels of genes *Nfib*, *Sox11*, *Nnat* and *Map1b* changed from low to high and then to low. Steele-Perkins et al. [63] reported the transcription factor *Nfib* is essential for mouse brain development. Jankowski et al. [64] found transcription factor *Sox11* modulates peripheral nerve regeneration in mice. Sel et al. [65] reported *Nnat* presented the spatial expression pattern during mouse eye development. Meixner et al. [66] found *Map1b* is required for axon guidance and is involved in the development of the central and peripheral nervous system.

Mouse embryo data

We applied DR-SC to analyze a large seqFISH (sequential fluorescence in situ hybridization) dataset of mouse organogenesis [15] which contains 23,194 cells. In this dataset, a panel of 351 genes were resolved spatially with their expressions in multiple 8-12 somite stage mouse embryo sections using the seqFISH platform. Cell labels across the embryo were accurately annotated in [15] based on their nearest neighbors in an existing scRNA-seq atlas (Gastrulation atlas) [67].

In our analysis, we first performed clustering analysis using DR-SC and other existing spatial clustering methods including SpaGCN, BayesSpace, SC-MEB and Giotto. By taking the above manually annotated cell types as reference, we compared the clustering performance of DR-SC and other methods. Since other existing methods were tandem analyses, we obtained top 15 PCs [25] using either PCA or WPCA from all 351 genes as their input except for SpaGCN. DR-SC showed better clustering performance over other clustering methods in terms of the ARI values (Additional file 1: Fig. S28a). The heatmap of cell types from annotations was provided in Fig. S28b (Additional file 1) while heatmaps of cell types inferred by DR-SC, SC-MEB and BayesSpace were provided in Fig. S28c (Additional file 1). We observed cell labels estimated by DR-SC and SC-MEB but not BayesSpace were in agreement with those from manual annotations. BayesSpace incorrectly clustered many cells into the "low quality" cells.

To refine the analysis on the brain regions, we first collected cells manually annotated as "forebrain/midbrain/hindbrain". Then we applied DR-SC to estimate low-dimensional embeddings and infer cell labels for these cells on the brain regions (Fig. 6a). In total, DR-SC identified six clusters. By checking PanglaoDB [57] for the marker genes identified via differential gene expression (DGE) analysis, we were able to identify four cell types (astrocyte, microglia cells, neurons1/2, and ependymal cells1/2) in Fig. 6b and four cortical regions (forebrain, hindbrain1/2/3, midbrain and microglia) in Fig. 6c. Details of cell typing were provided in Table S5 (Additional file 2). Note that neuron cells were from both forebrain and hindbrain regions, and glia (astrocytes, microglia) cells were from both midbrain and microglia regions. A recent study [59] reported that neurons and glia cells could distribute over different brain regions. The tSNE plot for regions and cell types in Fig. 6d showed that DR-SC could separate different clusters well.

To further investigate the development and differentiation of these brain cells, we calculated the pseudotime using Slingshot based on the 15-dimensional embeddings and cluster labels estimated from DR-SC. We identified three lineages that were consistent with the finding in [59]. Fig. 6e showed the inferred lineages and pseudotime, suggesting the dynamic trajectory from glia cells to neurons. Following [59], we used the Allen Brain Reference Atlases (http://atlas.brain-map.org/) as a guide to check how these trajectories were distributed by segmenting the brain. Fig. 6d&e showed that cells and trajectories from each cluster overwhelmingly occupied a different brain region. Combing pseudotime and spatial region information, we observed that cells in early differentiation clustered on the microglia and midbrain regions. Later, cells with differentiated transcriptome emerged in those farther regions, i.e., hinderbrain and forebrain. According to the inferred pseudotime, we identified DE genes along cell pseudotime by using the method in [33]. Fig. 6f showed the heatmap of the expression of the top 20 significant genes, suggesting some interesting dynamic expression patterns over pseudotime. We observed genes *Foxa1*, *Shh* and *Foxa2* had higher expression levels at the initial stage and later their expression levels decreased substantially. A study [68] reported that transcription factors *Foxa1* and *Foxa2* are crucial to maintain key cellular and functional features of dopaminergic neurons in the brain. In contrast, the expression levels of genes *Fgfr2* and *Fgfr3* changed from low to high and then to low. Existing studies [69, 70] reported that genes *Fgfr2* and *Fgfr3* played an important role during early neural development. Moreover, the expression of *Lhx2* and *Nr2f1* varied from low to high level. Peukert et al. [71] found that gene *Lhx2* determined neuronal differentiation and compartition in the caudal forebrain by regulating Wnt signaling. Bertacchi et al. [72] reported that *Nr2f1* regionally controls long-term self-renewal of neural progenitor cells via modulation of cell cycle genes. We observed the expression levels of gene *En1* had a different pattern that changed from high to low then to high along the inferred trajectory. In

humans, the *En1* gene codes for the homeobox protein engrailed (EN) family of transcription factors. A recent study [73] reported that *En1* is a transcriptional dependency in triple negative breast cancer associated with brain metastasis. Carratala-Marco et al. [74] found EN plays an important role in the regionalization of the neural tube and EN distribution regulates the cerebellum and midbrain morphogenesis, as well as retinotectal synaptogenesis.

Discussion

In this paper, we proposed a joint dimension reduction and spatial clustering method (DR-SC) for high-dimensional scRNA-Seq and (spatial) transcriptomics data using a hierarchical model. In contrast to most existing studies that perform dimensional reduction and (spatial) clustering sequentially, DR-SC unifies the low-dimensional feature extraction and (spatial) clustering together, with improved estimation for cell-type-relevant low-dimensional embeddings and enhanced clustering performance for both scRNA-seq and spatial transcriptomic data from different platforms. With simulations studies and analyses of benchmark datasets, we demonstrated that DR-SC can improve clustering performance while effectively estimating the low-dimensional embeddings.

DR-SC relies on a hidden Markov random field with a smoothing parameter to perform spatial clustering. The probabilistic framework of DR-SC allows us to adaptively update the spatial smoothing parameter that promotes the similar cluster assignments for the neighboring tissue locations in a data-driven manner. When the smoothness parameter is set as zero, DR-SC performs clustering for scRNA-seq data without spatial information. We developed an efficient EM algorithm based on iterative conditional mode and expectation-maximization (ICM-EM), making the DR-SC computationally efficient and scalable to large sample sizes.

In-depth analyses using scRNA-seq and spatial transcriptomic data from different platforms showed that the estimated clusters and embeddings from DR-SC could facilitate the downstream analysis. First, using a CITE-seq dataset for CBMC, we demonstrated that the analysis using DR-SC could improve the clustering performance while facilitating the identification of DE genes among different cell types. Second, we applied DR-SC to analyze a 10x Visium dataset for DLPFC, where we demonstrated the improved spatial clustering performance of DR-SC and further carried out conditional spatial variation analysis (SVA) to identify genes with pure spatial variations but not cell-type differences. The majority of genes identified in SVA without adjusting for cell-type-relevant covariates simply reflect cell-type differences. Functional enrichment analysis showed that genes identified in SVA adjusted for covariates were enriched in pathways related to the brain tissue. Third, we applied DR-SC to analyze two Slide-seqV2 datasets and showed it outperformed both the existing dimension reduction methods in visualization and the spatial clustering methods in terms of separation as well as its usefulness in cell trajectory inference. Finally, we applied DR-SC to analyze a seqFISH dataset and showcased its utility to infer cell lineage based on reduced-dimensionality space estimated by DR-SC.

There are several potential extensions for DR-SC. First, in the current study, we only consider a single transcriptional profile. The framework of DR-SC could be naturally extended to perform joint clustering analysis of multiple samples by properly removing their batch effects. Second, fast evolving technology in single-cell omics provides the opportunities and challenges to integrate omics profiles from different modalities of the same individuals. Extension of DR-SC for integration of multiple different omics will achieve higher statistical performance. Third, DR-SC essentially performs unsupervised clustering. With the availability of labels for some cells/spots, it would be interesting to perform semi-supervised clustering for those data.

We will investigate these issues in our future work.

Materials and Methods

An EM algorithm based on the iterative conditional mode

To make inference for the proposed hierarchical model, we designed an efficient EM algorithm based on the iterative conditional mode (ICM) [41, 42], named ICM-EM. It alternates between performing an ICM step (Fig. 1b), which creates a prediction for class labels $\mathbf{y}$ by maximizing its posterior, and an EM step (Fig. 1b), which computes the expectation of the log-likelihood and the model parameters $\hat{\boldsymbol{\theta}}$ by maximizing the expected log-likelihood. First, the complete-data likelihood can be written as

$$P(\mathbf{X}, \mathbf{Z}, \mathbf{y}) = \Pi_{i=1}^{n} P(\mathbf{x}_i|\mathbf{z}_i)P(\mathbf{z}_i|y_i)P(\mathbf{y}). \tag{4}$$

A pseudo-likelihood (PL) technique was used to replace the joint likelihood of class labels $\mathbf{y}$ with a pseudo likelihood, making the joint one separable. This technique was the key to make the computation tractable. Given the current estimates for model parameters, we apply ICM method to obtain a prediction of class labels $\hat{\mathbf{y}}$. Given this $\hat{\mathbf{y}}$, we can write the PL of $\mathbf{y}$ as

$$\tilde{P}(\mathbf{y}; \beta) = \Pi_i P(y_i|\mathbf{y}_{N_i} = \hat{\mathbf{y}}_{N_i}). \tag{5}$$

Plugging (5) into (4) and integrating out $\mathbf{Z}$ and $\mathbf{y}$, we obtain the pseudo log-likelihood for the observed data

$$\ln \tilde{P}(\mathbf{X}; \boldsymbol{\theta}, \hat{\mathbf{y}}) = \sum_i \ln \sum_k P(\mathbf{x}_i|y_i = k; \boldsymbol{\theta})P(y_i = k|\mathbf{y}_{N_i} = \hat{\mathbf{y}}_{N_i}; \boldsymbol{\theta}), \tag{6}$$

where $\boldsymbol{\theta} = (W, \Lambda, \mu_k, \Sigma_k, k = 1, \cdots, K, \beta)$ denotes all parameters in the model. Then we find the lower bound function of the pseudo log-likelihood for the observed data $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}, \hat{\mathbf{y}})$, known as Q-function, by Jensen's inequality. All parameters can be updated by maximizing $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}, \hat{\mathbf{y}})$.

To extract low dimensional representations of $\mathbf{X}$, we use the conditional expectation of $\mathbf{z}_i$ given $(\mathbf{x}_i, \mathbf{y}_{N_i} = \hat{\mathbf{y}}_{N_i})$, that is

$$E(\mathbf{z}_i|\mathbf{x}_i, \mathbf{y}_{N_i} = \hat{\mathbf{y}}_{N_i}) = \sum_k \{R_{ik}\langle \mathbf{z}_i \rangle_k\}, \tag{7}$$

where R_{ik} is the posterior probability of $y_i = k$ defined in (10) of Supplementary Materials (Additional file 1), and $\langle \mathbf{z}_i \rangle_k = \int \mathbf{z}_i P(\mathbf{z}_i|\mathbf{x}_i, y_i = k)d\mathbf{z}_i$. The entire process is summarized in Algorithm 1, where ICM function is defined in Supplementary algorithm 1 (Additional file 1).

Methods for comparison

We conducted comprehensive simulation and real data analysis by comparing DR-SC with existing methods in dimension reduction and clustering performance, respectively.

In detail, we considered the following eight dimension reduction methods to compare the DR performance. (1) PCA implemented in the R package *stats*; (2) weighted PCA (WPCA) [75] implemented in the R package *DR.SC*; (3) FKM [38] implemented in the R package *clustrd*; (4) tSNE; (5) UMAP, where tSNE and UMAP were implemented in the R package *scater*; (6)

Algorithm 1 The proposed ICM-EM algorithm for DR-SC

Input: $\mathbf{X}$, $\mathcal{S} = \{s_i\}_{i=1}^n$, q, K, grid points of β, *beta_grid*, maximum iterations of EM *maxIter*, relative tolerance of pseudo loglikelihood *epsLogLike*.

Output: $\hat{\mathbf{y}}, \hat{\mathbf{Z}}$ and $\hat{\boldsymbol{\theta}}$

```
Initialize ŷ^(0) and θ^(0)
for each t ∈ 1, ··· , maxIter do
    Update ŷ^(t) based on function ICM(X, ŷ^(t−1), θ^(t−1), S) in Supplementary algorithm 1;
    Update θ^(t) by maximizing Q-function Q(θ; θ^(t−1), ŷ^(t));
    Evaluate the pseudo observational log-likelihood, LogLike(t).
    if |LogLike(t) − LogLike(t − 1)|/|LogLike(t − 1)| < epsLogLike then
        break;
    end if
end for
Evaluate Ẑ based on Equation (7) by replacing R_ik and ⟨z_i⟩_k with R_ik^(t) and ⟨z_i⟩_k^(t), respectively.
return ŷ, Ẑ and θ̂
```

ZIFA is implemented in the Python module *ZIFA*; (7) ZINB-WaVE is implemented in the R package *zinbwave*; (8) scVI is implemented in the Python module *scvi*. Since the last three methods ZIFA, ZINB-WaVE and scVI can be applied to only raw count data, we compared DR-SC with them in Simulation 2 and real datasets.

We considered the following ten clustering methods for comparing the clustering performance. (1) BayesSpace [48] implemented in the R package *BayesSpace*; (2) Giotto [32] implemented in the R package *Giotto*; (3) SC-MEB [26] implemented in the R package *SC.MEB*; (4) SpaGCN [24] implemented in the Python module *SpaGCN*; (5) Louvain [49] implemented in the R package *igraph*; (6) Leiden [50] implemented in the R package *leiden*; (7) GMM implemented in the R package *mclust*; (8) k-means implemented in the R package *stats*; (9) FKM [38] implemented in the R package *clustrd*; (10) PSC [51] implemented in the R package *orclus*. In tandem analysis, BayesSpace, Giotto, SC-MEB and SpaGCN were recently developed to perform spatial clustering while Louvain, Ledien, GMM, and k-means were conventional non-spatial clustering algorithms. On the other hand, FKM and PSC were joint dimension reduction and clustering analysis.

Evaluation metrics

We evaluated the performance of DR-SC from four aspects, including feature extraction, clustering performance, selection of the number of clusters, and computational efficiency. Here, we briefly presented the evaluation metrics for feature extraction and clustering performance and refer details for other two aspects to Supplementary Materials (Additional file 1).

In simulations, we considered two metrics to assess the performance for feature extraction including both the canonical correlations between the estimated features and the underlying true ones and the conditional correlation between gene expression $\mathbf{x}_i$ and cell type label y_i given the estimated latent features. Canonical correlation measures the similarity of two sets of random variables. Thus, larger values in canonical correlation coefficients suggest better estimation for $\mathbf{z}_i$. Here, we calculated the mean of canonical correlation coefficients as follows

$$\mathrm{CCor}(\mathbf{z}_i, \hat{\mathbf{z}}_i) = \frac{1}{q} \sum_{l=1}^{q} r_l(\mathbf{z}_i, \hat{\mathbf{z}}_i),$$

where $r_l(\mathbf{z}_i, \widehat{\mathbf{z}}_i)$ is the l-th canonical correlation coefficient between $\mathbf{z}_i$ and $\widehat{\mathbf{z}}_i$. On the other hand, $\hat{\mathbf{z}}_i$ capturing all information regarding cell types was preferred, which was encapsulated as the mean of conditional correlation coefficients between gene expression $\mathbf{x}_i$ and cell type label y_i given the estimated latent features defined as

$$\text{ConCor} = \frac{1}{p}\sum_{j=1}^{p} corr(y_i, resid_{ij}),$$

where $resid_{ij}$ is the residual of x_{ij} regressing on $\widehat{\mathbf{z}}_i$ and $corr(y_i, resid_{ij})$ is the Pearson correlation coefficient between y_i and $resid_{ij}$. Optimally, we prefer to obtain the estimated features $\widehat{\mathbf{z}}_i$ that contain all information regarding cell types, in other words, $\mathbf{y}_i \perp \mathbf{x}_i | \widehat{\mathbf{z}}_i$.

To compare the clustering performance, we evaluated both adjusted Rand index (ARI) [76] and normalized mutual information (NMI) [77]. ARI [76] is the corrected version of Rand index (RI) [78] to avoid some drawback of RI [76] and defined as

$$\text{ARI} = \frac{RI - E(RI)}{\max(RI) - E(RI)},$$

where $E(RI)$ and $\max(RI)$ is the expected value and maximum value of RI, respectively. Suppose there are n cells/spots in scRNA-seq/spatial transcriptomics dataset. And let $U = (u_1, \cdots, u_i, \cdots, u_K)$ and $V = (v_1, \cdots, v_j, \cdots, v_L)$ denote two clustering labels for n cells/spots from two different methods, where K and L are corresponding to the numbers of clusters, respectively, from these two methods. Let n_{ij} be the number of cells/spots belonging to both classes u_i and v_j, and a_i and b_j be the number of cells/spots in classes u_i and v_j, respectively; then the specific formula of ARI is given by

$$\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}.$$

ARI is to measure the similarity of two different partitions and lies between -1 and 1. A larger value of ARI means a higher similarity between two partitions. When the two partitions are equal up to a permutation, the ARI takes a value of 1.

Normalized mutual information (NMI) is a corrected version of mutual information (MI) to make the value of MI range from zero to 1. MI originates from probability theory and information theory, and measures the mutual dependence between the two random variables. More specifically, it quantifies the "amount of information" in units such as shannons (bits) obtained about one random variable by observing the other random variable. Let x and y be two discrete random variables, i.e. the random variables taking value of class label on two different partitions, then MI of them can be defined as

$$\text{MI}(x, y) = \sum_x \sum_y P(x, y) \ln \frac{P(x, y)}{P(x)P(y)} = H(x) + H(y) - H(x, y), \tag{8}$$

where $P(x, y)$ is the joint distribution of (x, y), $P(x)$ and $P(y)$ are the marginal distribution of x and y, respectively, and $H(x)$, $H(y)$ and $H(x, y)$ are the marginal entropy of x, y and the joint entropy of (x, y), respectively. Intuitively, mutual information measures the information that x and y share. If x and y do not have shared information and are independent mutually, then $MI(x, y) = 0$. At the other extreme, if $y = x$, then $MI(x, y) = H(x)$, where $H(x)$ is the marginal entropy of x. This indicates MI does not take values in the interval of zero and one.

So some normalized versions are proposed and we used one version of them defined as

$$\mathrm{NMI}(x, y) = \frac{MI(x, y)}{\max(H(x), H(y))}.$$

From the above formula, we know when the two partitions are equal up to a permutation, the NMI takes a value of 1.

Simulations

Simulation 1. Log-normalized gene expression data. In this simulation, we generated the non-spatial/spatial log-normalized gene expressions. In detail, we generate the class label y_i for each $i = 1, \cdots, n$ in a rectangular 70×70 lattice from a K-state (K=7) Potts model with smoothing parameter $\beta = 0$ or 1 using function `sampler.mrf` in R package *GiRaF*. Then we generate latent low-dimensional features $\mathbf{z}_i$ from the conditional Gaussian such that $\mathbf{z}_i | y_i = k \sim (\mu_k, \Sigma_k)$, where $\mathbf{z}_i \in R^q$ with $q = 10$ and structures for μ_k and Σ_k are shown in Table S6 (Additional file 2). Next, we generate $\widetilde{W} = (\widetilde{w}_{ij}, i \leq p, j \leq q)$ with each $\widetilde{w}_{ij} \overset{i.i.d.}{\sim} N(0, 1)$, perform a QR decomposition on $\widetilde{W}$ such that $\widetilde{W} = \widetilde{Q}\widetilde{R}$, and assign $W = \widetilde{Q}$ that is a column orthogonal matrix. Finally, we generate a high-dimensional expression matrix using $\mathbf{x}_i = W\mathbf{z}_i + \boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \Lambda)$, where $\Lambda = \mathrm{diag}(\lambda_j), j = 1, \ldots, p$. In the case of homoscedasticity, $\lambda_j = 9, \forall j$, while in the case of heteroscedasticity, $\lambda_j = 2 + 4|a_j|, a_j \overset{i.i.d.}{\sim} N(0, 9)$.
Simulation 2. Raw gene expression data. In the second simulation, we generated the non-spatial/spatial raw gene expressions. The way to generate the class label y_i, loading matrix w and latent features $\mathbf{z}_i$ is the same as that in simulation 1 except that μ_k has a different value, see Table S6 (Additional file 2). The difference is to generate log-normalized gene expressions $\mathbf{x}_i$ by using $\mathbf{x}_i = W\mathbf{z}_i + \tau + \boldsymbol{\varepsilon}_i, \tau_j \sim N(0, 1), \boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \Lambda)$ and raw gene expressions $\tilde{\mathbf{x}}_i$ by using $\tilde{\mathbf{x}}_{ij} \sim Poisson(x_{ij})$, where τ_j is the j-th element of τ, $\Lambda = \mathrm{diag}(\lambda_j), j = 1, \cdots, p$. To ensure a proper signal, we set $\lambda_j = 1, \forall j$, in the case of homoscedasticity and $\lambda_j = 0.1 + |a_j|, a_j \overset{i.i.d.}{\sim} N(0, 1)$ in the case of heteroscedasticity. In this simulation, we only observe raw gene expression $\tilde{x}_{ij}$ of gene j and cell i for non-spatial setting, observe raw gene expression $\tilde{x}_{ij}$ of gene j and spot i and spatial coordinates s_i for spot i.

Data resources and analyses

Human dorsolateral prefrontal cortex datasets

We downloaded spatial transcriptomics on 10x Visium platform for human dorsolateral prefrontal cortex (DLPFC) at `https://github.com/LieberInstitute/spatialLIBD`. This dataset collects 12 human postmortem DLPFC tissue sections from three independent neurotypical adult donors. The raw data for each sample included 33538 genes. We first selected the genes with spatial variation using SPARK [11] without adjusting for any covariates. In detail, we selected the spatial variable genes (SVGs) either with adjusted p-values less than 0.05 or top 2,000 SVGs (Additional file 2: Table S7). Then we performed log-normalization using the library size. The detailed information for 12 samples was given in Table S7 (Additional file 2). By taking manual annotations based on cytoarchitecture as the benchmark, we would be able to evaluate the clustering performance of DR-SC and other methods. In tandem analysis, we first obtained top 15 PCs from either PCA or weighted PCA (WPCA) and then applied other clustering methods using top 15 PCs [25]. We further performed the spatial variation analysis to identify SVGs adjusted for cell-type-relevant covariates using SPARK in comparision

with SVGs without adjusting for these covariates. We then performed DEG analysis using the function *FindAllMarkers* in the R package *Seurat* to identify DE genes based on cell type labels estimated using DR-SC. Finally, we performed functional enrichment analysis using g:profiler [79] `https://biit.cs.ut.ee/gprofiler/gost` based on the SVGs with adjustment.

Mouse olfactory bulb data

We downloaded mouse olfactory bulb data from `https://singlecell.broadinstitute.org/single_cell/data/public/SCP815/sensitive-spatial-genome-wide-expression-profiling-at-cellular-resolution#study-summary`. We first selected top 2,000 genes with spatial variation using SPARK [11] without adjusting for any covariates. Then we performed log-normalization for these SVGs using the library size and obtained top 15 PCs based on PCA. Since BayesSpace and SC-MEB are both based on tandem analysis, the obtained top PCs from PCA were used as their inputs. SpaGCN is also based on tandem analysis but it used its internally embedded PCA algorithm. As joint method, DR-SC took the 2,000 SVGs as inputs. By applying DR-SC in this dataset, we clustered all spots into 12 clusters. Using the estimated class labels for 12 clusters from DR-SC, we performed DGE analysis using the function *FindAllMarkers* in the R package *Seurat* to identify the marker genes for each cluster. Next, We performed cell typing using PanglaoDB database [57] for each class to obtain seven cell types in Fig. 4d. In the following, we performed spatial variation analysis using SPARK by adjusting the 15 features extracted from DR-SC. Finally, we performed trajectory inference by Slingshot method based on the extracted features and cell class estimated by DR-SC and detected the the DE genes along the inferred cell pseudotime by using the function *testPseudotime* in the R package *TSCAN*.

Mouse E15 neocortex data

We first obtained mouse E15 neocortex data from `https://singlecell.broadinstitute.org/single_cell/data/public/SCP815/sensitive-spatial-genome-wide-expression-profiling-at-cellular-resolution#study-summary`. We then selected top 2,000 genes with spatial variation using SPARK [11] without adjusting for any covariates. Then we performed log-normalization for these SVGs using the library size and obtained top 15 PCs based on PCA as the input of BayesSpace and SC-MEB. As joint method, DR-SC took the 2,000 SVGs as inputs. By applying DR-SC in this dataset, we clustered all spots into 15 clusters. Next, we performed DGE analysis based on the estimated class labels for these 15 clusters using the function *FindAllMarkers* in the R package *Seurat* to identify the marker genes for each cluster. Next, We conducted cell typing using PanglaoDB database [57] for each class to obtain five cell types in Fig. 5d. Finally, we performed trajectory inference with Slingshot method and detected the the DE genes along the inferred cell pseudotime by using the function *testPseudotime* in the R package *TSCAN*.

Mouse embryo datasets

We downloaded mouse embryo dataset [15] at `https://content.cruk.cam.ac.uk/jmlab/SpatialMouseAtlas2020/` that were measured on sequential fluorescence in situ hybridization (seqFISH) platform. This dataset contains 23,194 cells, 351 genes and two-dimensional spatial coordinates. Cell types were annotated based on their nearest neighbors in an existing scRNA-seq atlas (Gastrulation atlas) [15]. By taking these manual annotations as the benchmark, we compared the clustering performance of DR-SC and other spatial clustering methods. We

performed log-normalization with the library size on gene-expression matrix. In tandem analysis, we first obtained top 15 PCs [25] from either PCA or weighted PCA (WPCA) and then applied other spatial clustering methods using top 15 PCs. We further restricted our analysis to the cells manually annotated as "orebrain/midbrain/hindbrain" and performed the downstream trajectory analysis. By applying DR-SC, we obtained six subclusters in the brain region. Then we performed DGE analysis using the function *FindAllMarkers* in the R package *Seurat* to identify DE genes between the estimated clusters and further mapped six clusters to either four cortical regions or four cell types using PanglaoDB database [57]. To visualize the clustering results, we applied tSNE to reduce the 15-dimensional embeddings from different methods to a 2-dimensional representation. Finally, we applied Slingshot to conduct trajectory inference based on the features and clusters from DR-SC and detected the the DE genes along the inferred cell pseudotime by using the function *testPseudotime* in the R package *TSCAN*.

Benchmark datasets in trajectory inference

We downloaded 16 benchmark datasets with linear trajectory information from website `https://zenodo.org/record/1443566#.XNV25Y5KhaR` [80]. These datasets consist of single-cell gene-expression measurement in the form of raw read counts. The detailed information for these datasets was given in Table S8 (Additional file 2), including species, the number of cells, the number of genes, platform and so on. We first pre-processed raw count data using *Seruat*, including the selection of top 2,000 highly variable genes and log-normalization using the library size [47] for methods taking the normalized expressions except for ZINB-WaVE and scVI. After normalization, we estimated the low-dimensional embeddings and class labels using both joint and tandem methods. In joint analysis, we considered the proposed DR-SC and FKM while in tandem analysis, we performed dimension reduction using other methods followed by clustering analysis using the Gaussian mixture model (GMM). The number of clusters was chosen using modified BIC and BIC by default for DR-SC and GMM, respectively. Since FKM cannot select the number of clusters automatically, the one selected for DR-SC was set for FKM. To further perform the lineage development analysis, we applied Slingshot implemented in R package *slingshot* with the default values for parameters using the estimated embeddings and class labels as its input.

Cord blood mononuclear cells datasets

We obtained cord blood mononuclear cells datasets (CBMC) from NCBI `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE100866` by the access number GSE100866. This dataset contains 8,167 cells measured using CITE-Seq technology [81] for the cord blood mononuclear cells from two species (Human and mouse). In addition to genome-wide expression measurements of 20,511 genes in the form of read counts, this dataset also measures the protein levels of 13 cell-surface markers. We first perform pre-processing for top 2,000 highly variable genes and log-normalization using the library size [47]. Following the guidelines of Stoeckius et al. [81], we dropped three cell-surface markers because of low quality and performed clustering analysis for the rest 10 markers using the *FindClusters* function in the R package *Seurat*. By taking the class labels from nine clusters estimated using these 10 surface markers as the benchmark, we evaluated the clustering performance of DR-SC and other methods using the ARI values. DR-SC and FKM simultaneously estimated the embeddings and class labels while other methods are tandem analyses. Using the estimated class labels for eleven clusters from DR-SC, we performed differential gene expression (DGE) analysis using the R package *BPSC* [37] for human cells. Next, We performed cell typing using PanglaoDB database [57]

for each class identified by DR-SC to obtain nine cell types in Fig. S4a. For each cell type, we further performed functional enrichment analysis by selecting the significant genes with adjusted p-values less than 0.05 and log fold change greater than 0.5.

Availability of data and materials

All codes and analysis results in this paper are publicly available at GitHub [82]. The source code is released under the GNU general public license. The 16 benchmark datasets with linear trajectory informationare are available at `https://zenodo.org/record/1443566#.XNV25Y5KhaR`. The cord blood mononuclear cells datasets aer available at `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE100866` by the access number GSE100866. The dorsolateral prefrontal cortex datasets for human on 10x Visium platform are accessible at `https://github.com/LieberInstitute/spatialLIBD`. The mouse olfactory bulb data and mouse E15 neocortex data on Slide-seqV2 platform are available at `https://singlecell.broadinstitute.org/single_cell/data/public/SCP815`. The mouse embryo dataset on seqFISH platform is accessible at `https://content.cruk.cam.ac.uk/jmlab/SpatialMouseAtlas2020/`.

Ethics approval and consent to participate

No ethical approval was required for this study. All utilized public data sets were generated by other organizations that obtained ethical approval.

Competing interests

The authors have no competing interests.

Consent for publication

Not applicable.

Authors' contributions

J.L., X.S. and X.Z. initiated and designed the study, W.L. and X.L. implemented the model and performed simulation studies and benchmarking evaluation, J.L., X.S. and X.Z. wrote the manuscript, and all authors edited and revised the manuscript.

References

[1] Simone Picelli, Omid R Faridani, **r**Asa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length rna-seq from single cells using smart-seq2. *Nature protocols*, 9(1):171–181, 2014.

[2] Helena L Crowell, Charlotte Soneson, Pierre-Luc Germain, Daniela Calini, Ludovic Collin, Catarina Raposo, Dheeraj Malhotra, and Mark D Robinson. Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nature communications*, 11(1):1–12, 2020.

[3] Simone Picelli, **r**Asa K Björklund, Omid R Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods*, 10(11):1096–1098, 2013.

[4] Kuanwei Sheng, Wenjian Cao, Yichi Niu, Qing Deng, and Chenghang Zong. Effective detection of variation in single-cell transcriptomes using matq-seq. *Nature methods*, 14(3):267–270, 2017.

[5] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[6] Saiful Islam, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson. Highly multiplexed and strand-specific single-cell rna 5' end sequencing. *Nature protocols*, 7(5):813–828, 2012.

[7] Sheel Shah, Eric Lubeck, Wen Zhou, and Long Cai. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron*, 92(2):342–357, 2016.

[8] Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233), 2015.

[9] Patrik L St**r**ahl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

[10] Samuel G Rodriques, Robert R Stickels, Aleksandrina Goeva, Carly A Martin, Evan Murray, Charles R Vanderburg, Joshua Welch, Linlin M Chen, Fei Chen, and Evan Z Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, 2019.

[11] Shiquan Sun, Jiaqiang Zhu, and Xiang Zhou. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature methods*, 17(2):193–200, 2020.

[12] Jiaqiang Zhu, Shiquan Sun, and Xiang Zhou. Spark-x: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biology*, 22(1):1–25, 2021.

[13] Kristen R Maynard, Leonardo Collado-Torres, Lukas M Weber, Cedric Uytingco, Brianna K Barry, Stephen R Williams, Joseph L Catallini, Matthew N Tran, Zachary Besich, Madhavi Tippani, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience*, 24(3):425–436, 2021.

[14] Shahar Alon, Daniel R Goodwin, Anubhav Sinha, Asmamaw T Wassie, Fei Chen, Evan R Daugharthy, Yosuke Bando, Atsushi Kajita, Andrew G Xue, Karl Marrett, et al. Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. *Science*, 371(6528), 2021.

[15] Tim Lohoff, Shila Ghazanfar, Alsu Missarova, Noushin Koulena, Nico Pierson, Jonathan A Griffiths, Evan S Bardot, Chee-Huat Linus Eng, Richard CV Tyser, Ricard Argelaguet, et al. Highly multiplexed spatially resolved gene expression profiling of mouse organogenesis. *BioRxiv*, 2020.

[16] Shiquan Sun, Jiaqiang Zhu, Ying Ma, and Xiang Zhou. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell rna-seq analysis. *Genome biology*, 20(1):1–21, 2019.

[17] Tae Hyun Kim, Xiang Zhou, and Mengjie Chen. Demystifying "drop-outs" in single-cell umi data. *Genome biology*, 21(1):1–19, 2020.

[18] Shixiong Zhang, Xiangtao Li, Qiuzhen Lin, and Ka-Chun Wong. Review of single-cell rna-seq data clustering for cell type identification and characterization. *arXiv preprint arXiv:2001.01006*, 2020.

[19] David Hong, Jeffrey A Fessler, and Laura Balzano. Optimally weighted pca for high-dimensional heteroscedastic data. *arXiv preprint arXiv:1810.12862*, 2018.

[20] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[21] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[22] Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282, 2019.

[23] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.

[24] Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J Irwin, Edward B Lee, Russell T Shinohara, and Mingyao Li. Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature methods*, pages 1–10, 2021.

[25] Edward Zhao, Matthew R Stone, Xing Ren, Jamie Guenthoer, Kimberly S Smythe, Thomas Pulliam, Stephen R Williams, Cedric R Uytingco, Sarah EB Taylor, Paul Nghiem, et al. Spatial transcriptomics at subspot resolution with bayesspace. *Nature Biotechnology*, pages 1–10, 2021.

[26] Yi Yang, Xingjie Shi, Qiuzhong Zhou, Mai Chan Lau, Lei Sun, Joe Yeong, and Jin Liu. Sc-meb: spatial clustering with hidden markov random field using empirical bayes. *bioRxiv*, 2021.

[27] Emma Pierson and Christopher Yau. Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology*, 16(1):1–10, 2015.

[28] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell rna-seq data. *Nature communications*, 9(1):1–17, 2018.

[29] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

[30] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

[31] Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631, 2002.

[32] Ruben Dries, Qian Zhu, Rui Dong, Chee-Huat Linus Eng, Huipeng Li, Kan Liu, Yuntian Fu, Tianxiao Zhao, Arpan Sarkar, Feng Bao, et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome biology*, 22(1):1–31, 2021.

[33] Zhicheng Ji and Hongkai Ji. Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. *Nucleic acids research*, 44(13):e117–e117, 2016.

[34] Xiaojie Qiu, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A Pliner, and Cole Trapnell. Reversed graph embedding resolves complex single-cell trajectories. *Nature methods*, 14(10):979–982, 2017.

[35] Kelly Street, Davide Risso, Russell B Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*, 19(1):1–16, 2018.

[36] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.

[37] Trung Nghia Vu, Quin F Wills, Krishna R Kalari, Nifang Niu, Liewei Wang, Mattias Rantalainen, and Yudi Pawitan. Beta-poisson model for single-cell rna-seq data analyses. *Bioinformatics*, 32(14):2128–2135, 2016.

[38] Angelos Markos, Alfonso Iodice DEnza, and Michel van de Velden. Beyond tandem analysis: Joint dimension reduction and clustering in r. *Journal of Statistical Software (Online)*, 91(10), 2019.

[39] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

[40] Yao-Hung Hubert Tsai, Martin Q Ma, Muqiao Yang, Han Zhao, Louis-Philippe Morency, and Ruslan Salakhutdinov. Self-supervised representation learning with relative predictive coding. *arXiv preprint arXiv:2103.11275*, 2021.

[41] Julian Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(3):179–195, 1975.

[42] W Qian and DM Titterington. Estimation of parameters in hidden markov models. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 337(1647):407–428, 1991.

[43] Hansheng Wang, Bo Li, and Chenlei Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683, 2009.

[44] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck III, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 2021.

[45] Fran**c**cois Graner and James A Glazier. Simulation of biological cell sorting using a two-dimensional extended potts model. *Physical review letters*, 69(13):2013, 1992.

[46] Tara Chari, Joeyta Banerjee, and Lior Pachter. The specious art of single-cell genomics. *bioRxiv*, 2021.

[47] Aaron TL Lun, Karsten Bach, and John C Marioni. Pooling across cells to normalize single-cell rna sequencing data with many zero counts. *Genome biology*, 17(1):1–14, 2016.

[48] Edward Zhao, Matthew R Stone, Xing Ren, Thomas Pulliam, Paul Nghiem, Jason H Bielas, and Raphael Gottardo. Bayesspace enables the robust characterization of spatial gene expression architecture in tissue sections at increased resolution. *bioRxiv*, 2020.

[49] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[50] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

[51] Charu C Aggarwal and Philip S Yu. Finding generalized projected clusters in high dimensional spaces. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 70–81, 2000.

[52] Shujie Ma and Jian Huang. A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, 112(517):410–423, 2017.

[53] Michaela Asp, Joseph Bergenstr**r**ahle, and Joakim Lundeberg. Spatially resolved transcriptomesnext generation tools for tissue exploration. *BioEssays*, 42(10):1900221, 2020.

[54] William P Flavin, Luc Bousset, Zachary C Green, Yaping Chu, Stratos Skarpathiotis, Michael J Chaney, Jeffrey H Kordower, Ronald Melki, and Edward M Campbell. Endocytic vesicle rupture is a conserved mechanism of cellular invasion by amyloid proteins. *Acta neuropathologica*, 134(4):629–653, 2017.

[55] Johannes Weickenmeier, Mathias Jucker, Alain Goriely, and Ellen Kuhl. A physics-based model explains the prion-like features of neurodegeneration in alzheimers disease, parkinsons disease, and amyotrophic lateral sclerosis. *Journal of the Mechanics and Physics of Solids*, 124:264–281, 2019.

[56] Robert R Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L Marshall, Daniela J Di Bella, Paola Arlotta, Evan Z Macosko, and Fei Chen. Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqv2. *Nature biotechnology*, 39(3):313–319, 2021.

[57] Oscar Franzén, Li-Ming Gan, and Johan LM Björkegren. Panglaodb: a web server for exploration of mouse and human single-cell rna sequencing data. *Database*, 2019, 2019.

[58] Neal H Barmack and Vadim Yakhnitsa. Functions of interneurons in mouse cerebellum. *Journal of Neuroscience*, 28(5):1140–1152, 2008.

[59] Sanjay R Srivatsan, Mary C Regier, Eliza Barkan, Jennifer M Franks, Jonathan S Packer, Parker Grosjean, Madeleine Duran, Sarah Saxton, Jon J Ladd, Malte Spielmann, et al. Embryo-scale, single-cell spatial transcriptomics. *Science*, 373(6550):111–117, 2021.

[60] Sébastien Küry, Geeske M van Woerden, Thomas Besnard, Martina Proietti Onori, Xénia Latypova, Meghan C Towne, Megan T Cho, Trine E Prescott, Melissa A Ploeg, Stephan Sanders, et al. De novo mutations in protein kinase genes camk2a and camk2b cause intellectual disability. *The American Journal of Human Genetics*, 101(5):768–788, 2017.

[61] Bin Zhang, Gayatri Arun, Yuntao S Mao, Zsolt Lazar, Gene Hung, Gourab Bhattacharjee, Xiaokun Xiao, Carmen J Booth, Jie Wu, Chaolin Zhang, et al. The lncrna malat1 is dispensable for mouse development but its transcription plays a cis-regulatory role in the adult. *Cell reports*, 2(1):111–123, 2012.

[62] Joel N Buxbaum, Zhengyi Ye, Natàlia Reixach, Linsey Friske, Coree Levy, Pritam Das, Todd Golde, Eliezer Masliah, Amanda R Roberts, and Tamas Bartfai. Transthyretin protects alzheimer's mice from the behavioral and biochemical effects of aβ toxicity. *Proceedings of the National Academy of Sciences*, 105(7):2681–2686, 2008.

[63] George Steele-Perkins, Céline Plachez, Kenneth G Butz, Guanhu Yang, Cindy J Bachurski, Stephen L Kinsman, E David Litwack, Linda J Richards, and Richard M Gronostajski. The transcription factor gene nfib is essential for both lung maturation and brain development. *Molecular and cellular biology*, 25(2):685–698, 2005.

[64] Michael P Jankowski, Sabrina L McIlwrath, Xiaotang Jing, Pamela K Cornuet, Kathleen M Salerno, H Richard Koerber, and Kathryn M Albers. Sox11 transcription factor modulates peripheral nerve regeneration in adult mice. *Brain research*, 1256:43–54, 2009.

[65] Saadettin Sel, Eva Patzel, Lucia Poggi, Delia Kaiser, Thomas Kalinski, Martin Schicht, Friedrich Paulsen, and Norbert Nass. Temporal and spatial expression pattern of nnat during mouse eye development. *Gene Expression Patterns*, 23:7–12, 2017.

[66] Arabella Meixner, Silke Haverkamp, Heinz Wässle, Susanne Führer, Johann Thalhammer, Nina Kropf, Reginald E Bittner, Hans Lassmann, Gerhard Wiche, and Friedrich Propst. Map1b is required for axon guidance and is involved in the development of the central and peripheral nervous system. *The Journal of cell biology*, 151(6):1169–1178, 2000.

[67] Blanca Pijuan-Sala, Jonathan A Griffiths, Carolina Guibentif, Tom W Hiscock, Wajid Jawaid, Fernando J Calero-Nieto, Carla Mulas, Ximena Ibarra-Soria, Richard CV Tyser, Debbie Lee Lian Ho, et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, 566(7745):490–495, 2019.

[68] Alessandro Pristerà, Wei Lin, Anna-Kristin Kaufmann, Katherine R Brimblecombe, Sarah Threlfell, Paul D Dodson, Peter J Magill, Cathy Fernandes, Stephanie J Cragg, and Siew-Lan Ang. Transcription factors foxa1 and foxa2 maintain dopaminergic neuronal properties and control feeding behavior in adult mice. *Proceedings of the National Academy of Sciences*, 112(35):E4929–E4938, 2015.

[69] Jennifer Walshe and Ivor Mason. Expression of fgfr1, fgfr2 and fgfr3 during early neural development in the chick embryo. *Mechanisms of development*, 90(1):103–110, 2000.

[70] Jonna Saarimäki-Vire, Paula Peltopuro, Laura Lahti, Thorsten Naserke, Alexandra A Blak, Daniela M Vogt Weisenhorn, Kai Yu, David M Ornitz, Wolfgang Wurst, and Juha Partanen. Fibroblast growth factor receptors cooperate to regulate neural progenitor properties in the developing midbrain and hindbrain. *Journal of Neuroscience*, 27(32):8581–8592, 2007.

[71] Daniela Peukert, Sabrina Weber, Andrew Lumsden, and Steffen Scholpp. Lhx2 and lhx9 determine neuronal differentiation and compartition in the caudal forebrain by regulating wnt signaling. *PLoS biology*, 9(12):e1001218, 2011.

[72] Michele Bertacchi, Anna Lisa Romano, Agnès Loubat, Frederic Tran Mau-Them, Marjolaine Willems, Laurence Faivre, Philippe Khau van Kien, Laurence Perrin, Fran**c**coise Devillard, Arthur Sorlin, et al. Nr2f1 regulates regional progenitor dynamics in the mouse neocortex and cortical gyrification in bbsoas patients. *The EMBO journal*, 39(13):e104163, 2020.

[73] Guillermo Peluffo, Ashim Subedee, Nicholas W Harper, Natalie Kingston, Bojana Jovanović, Felipe Flores, Laura E Stevens, Francisco Beca, Anne Trinh, Chandra Sekhar Reddy Chilamakuri, et al. En1 is a transcriptional dependency in triple-negative breast cancer associated with brain metastasis. *Cancer research*, 79(16):4173–4183, 2019.

[74] Francisco Carratala-Marco, Patricia Andreo-Lillo, Marta Martinez-Morga, Teresa Escamez-Martínez, Arancha Botella-López, Carlos Bueno, and Salvador Martinez. Clinical phenotypes associated to engrailed 2 gene alterations in a series of neuropediatric patients. *Frontiers in neuroanatomy*, 12:61, 2018.

[75] Jushan Bai and Yuan Liao. Inferences in panel data with interactive effects using large covariance matrices. *Journal of Econometrics*, 200(1):59–78, 2017.

[76] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[77] Thomas M Cover and Joy A Thomas. *Elements of information theory 2nd edition (wiley series in telecommunications and signal processing)*. Wiley-Interscience, 2006.

[78] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[79] Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research*, 47(W1):W191–W198, 2019.

[80] Robrecht Cannoodt, W Saelens, H Todorov, and Y Saeys. Single-cell-omics datasets containing a trajectory. *Zenodo (Oct. 2018). DOI*, 10, 2018.

[81] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Large-scale simultaneous measurement of epitopes and transcriptomes in single cells. *Nature methods*, 14(9):865, 2017.

[82] Wei Liu, Xu Liao, Xiang Zhou, Xingjie Shi, and Jin Liu. Statistical analysis of joint dimension reduction and (spatial) clustering for both single-cell rna-seq and spatial transcriptomics data. *Github*, 2021.

Figures

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

Additional Files

Additional file 1 — Supplementary-Materials.pdf

It includes all the supplementary figures and the details of DR-SC.

Additional file 2 — Supplementary-Tables.xlsx

It includes all the supplementary tables.

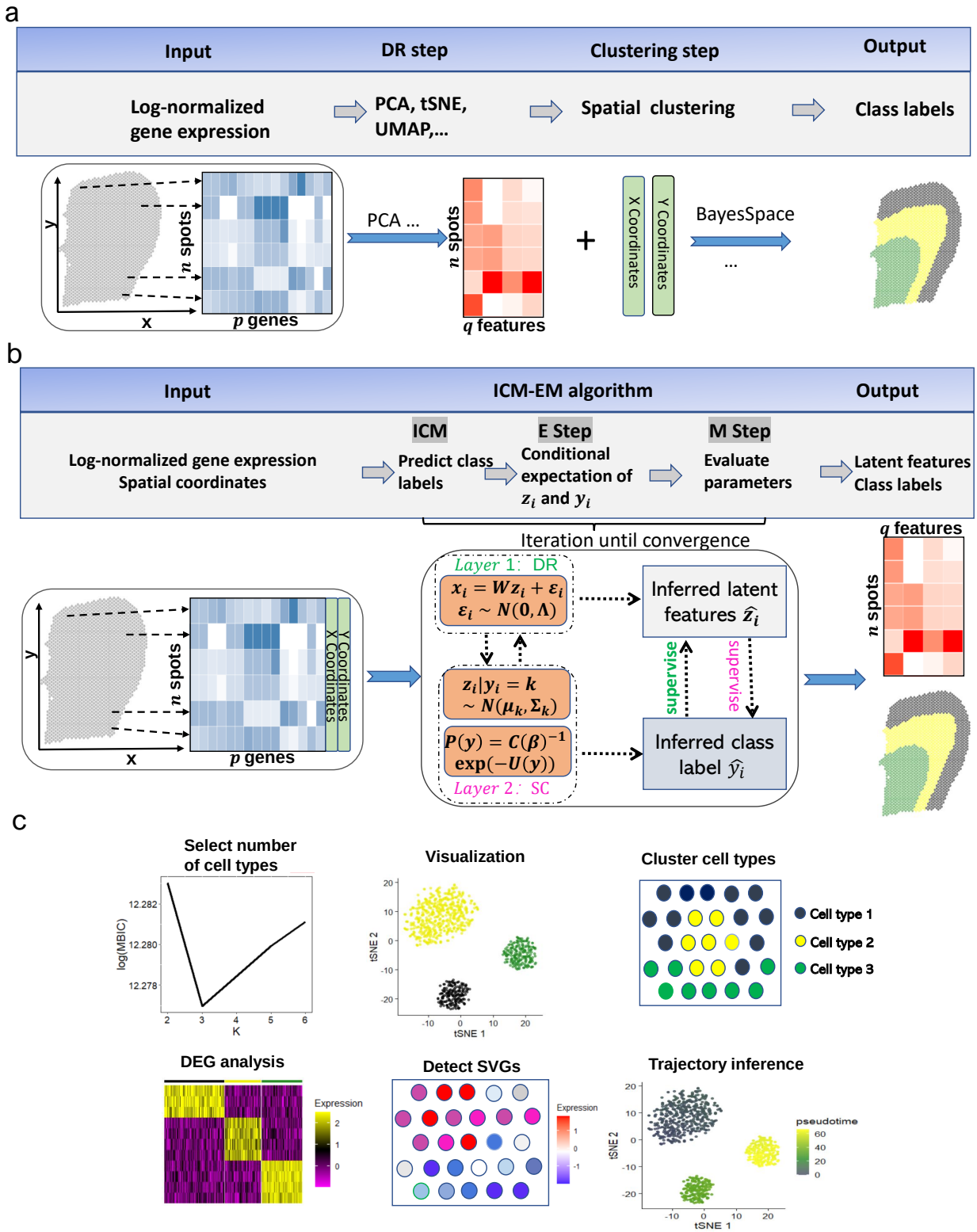

Figure 1: Workflows for both tandem analysis (a) and DR-SC (b) and potential applications of DR-SC in the downstream analysis (c). a&b. Compared with tandem analysis, DR-SC iteratively performs dimension reduction and (spatial) clustering with improved estimation for both clustering and low-dimensional embeddings. c. DR-SC could be used to perform clustering for cell types with the number of clusters selected in a data-driven manner. The estimated cell types could be used to perform differential gene expression analysis. The estimated low-dimensional embeddings from DR-SC could be used for visualization, trajectory inference, and detection of genes with spatially variations by controlling cell-type-relevant covariates.

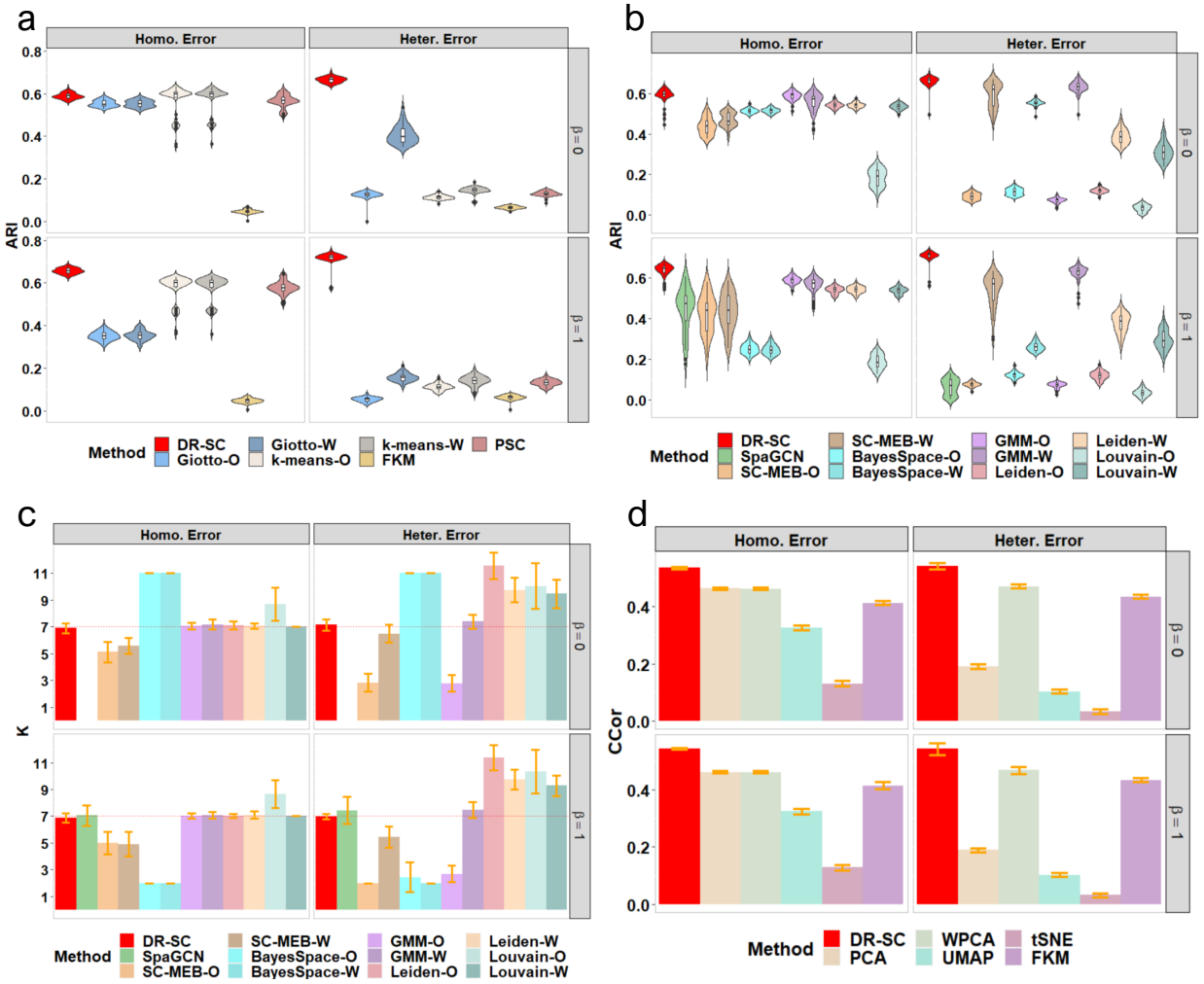

Figure 2: Comparisons in Simulation 1 that log-normalized gene expressions were simulated. For methods of tandem analysis, we consider both PCA and WPCA to obtain PCs and name them as method-O or method-W. a. comparison of the clustering performance for seven methods given the true number of clusters. b. comparison of the clustering performance for 12 methods that can choose number of clusters. c. comparison of the performance of selecting number of clusters for 12 methods that can choose the number of clusters. d. comparison of the performance of dimension reduction for six methods.

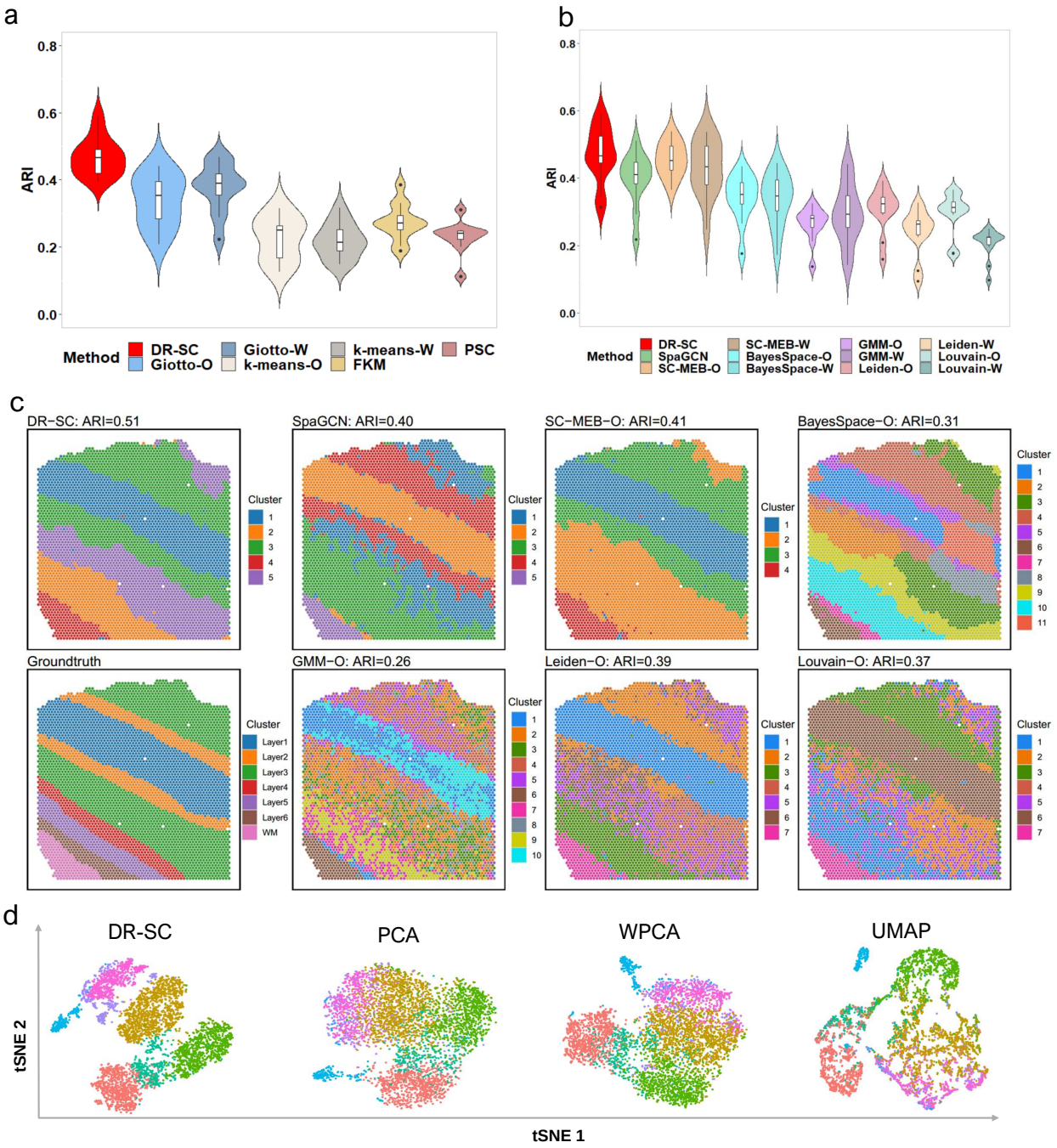

Figure 3: Analysis of human dorsolateral prefrontal cortex data. a. Boxplot of ARI values across 12 samples for DR-SC and other methods that cannot choose the number of clusters. The number of clusters was fixed by their manual annotations in the analysis. b. Boxplot of ARI values across 12 samples for DR-SC and other methods that are capable of choosing the number of clusters. c. Spatial heatmaps for cluster assignment of sample ID 151510 using DR-SC and other spatial and non-spatial clustering methods. The left bottom corner denotes cell assignment from manual annotation, the upper panel corresponds to the cell assignment from spatial clustering methods, and the rest of lower panel corresponds to the cell assignment from non-spatial clustering methods. d. visualization of the cluster labels for sample ID 151510 from DR-SC given the annotated number of clusters based on two-dimensional tSNE embeddings from four different DR methods including DR-SC, PCA, WPCA and UMAP.

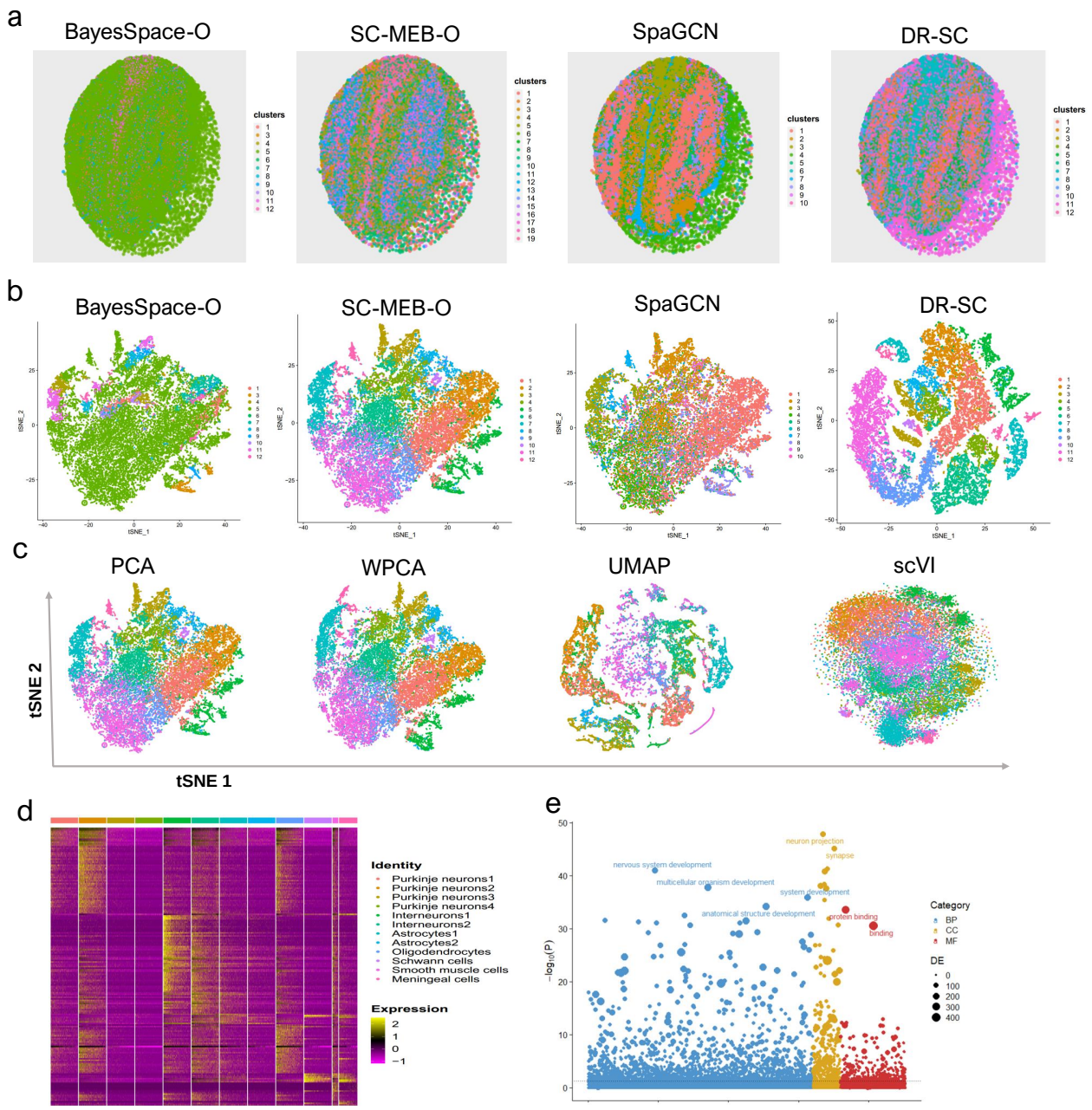

Figure 4: Analysis of mouse olfactory bulb data. a. spatial heatmap for clusters from four spatial clustering methods. b. tSNE plots for these four methods, where tSNE PCs of DR-SC were obtained based on its extracted 15-dimensional features while tSNE PCs of BayesSpace-O, SC-MEB-O and SpaGCN were based on 15 PCs from PCA. c. visualization of the cluster labels from DR-SC based on two-dimensional tSNE embeddings from four different DR methods including PCA, WPCA, UMAP and scVI. d. heatmap of DE genes for each cell type identified by DR-SC. e. Bubble plot of $-\log10(p-\text{values})$ for pathway enrichment analysis on 518 SVGs with an adjusted p-value less than 0.05. The dashed line represents a p-value cutoff of 0.05. Gene sets are colored by categories: GO biological process (BP, blue), and GO cellular component (CC, yellow), GO molecular function (MF, brown).

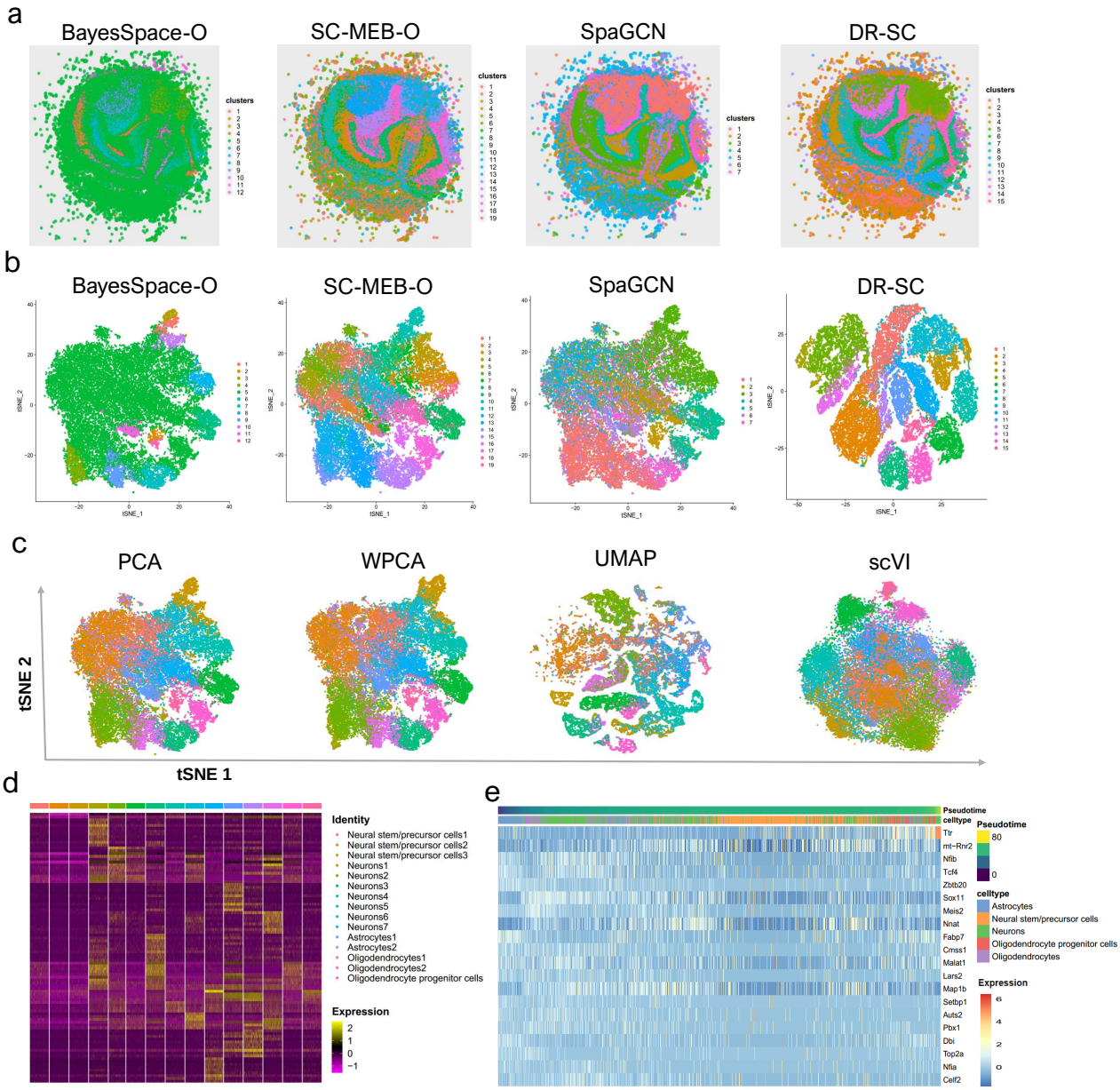

Figure 5: Analysis of mouse E15 neocortex data. a. spatial heatmap for clusters from four spatial clustering methods including BayesSapce-O, SC-MEB-O, SpaGCN and DR-SC. b. tSNE plots for these four methods, where tSNE PCs of DR-SC were obtained based on its extracted 15-dimensional features while tSNE PCs of BayesSpace-O, SC-MEB-O and SpaGCN were based on 15 PCs from PCA. c. visualization of the cluster labels estimated by DR-SC based on two-dimensional tSNE embeddings from four different DR methods including PCA, WPCA, UMAP and scVI. d. heatmap of DE genes for each cell type identified by DR-SC. e. Heatmap of gene expression levels for top 20 genes with significant changes with respect to the Slingshot pseudotime. Each column represents a spot that is mapped to this path and is ordered by its pseudotime value. Each row denotes a top significant gene.

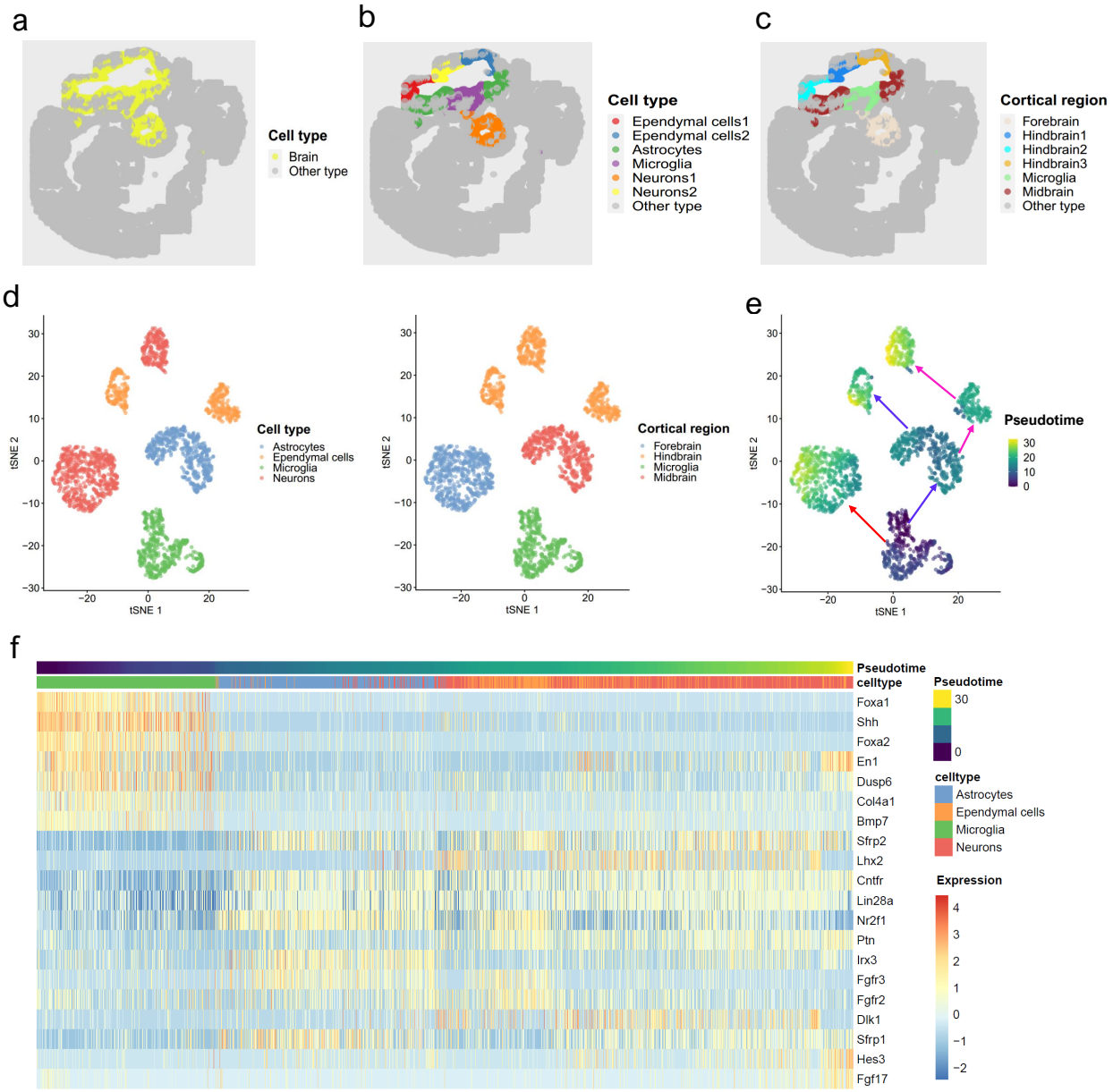

Figure 6: Analysis of mouse embryo data. a. Spatial heatmap of brain area and other area. b. Spatial heatmap of cell types based on clusters identified by DR-SC. c. Spatial heatmap of the cortical region identified by DR-SC. d. tSNE plot of cell types and corresponding cortical regions, where the tSNE projection is evaluated based on the estimated low-dimensional embeddings using DR-SC. Note that the cortical regions and cell types are well separated. e. tSNE plot of inferred pseudotime by using Slingshot based on the estimated low-dimensional embeddings and cluster labels for the cortical region from DR-SC. f. Heatmap of gene expression levels for top 20 genes with significant changes with respect to the Slingshot pseudotime. Each column represents a spot that is mapped to this path and is ordered by its pseudotime value. Each row denotes a top significant gene.