Novel tensor decomposition-based approach for cell type deconvolution in Visium datasets when reference scRNA-seqs include multiple minor cell types

Y-H Taguchi[1*] and Turki Turki[2]

[1*]Department of Physics, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8551, Japan.
[2]Department of Computer Science, King Abdulaziz University, Jeddah, 21589, Saudi Arabia.

*Corresponding author(s). E-mail(s): tag@granular.com;
Contributing authors: tturki@kau.edu.sa;

Abstract

We have applied tensor decomposition (TD)-based unsupervised feature extraction (FE) to integrate multiple Visium datasets, as a platform for spatial gene expression profiling (spatial transcriptomics). As a result, TD-based unsupervised FE successfully obtains singular value vectors consistent with spatial distribution, i.e., singular value vectors with similar values are assigned to neighboring spots. Further, TD-based unsupervised FE successfully infers the cell type fractions within individual Visium spots (i.e., successful deconvolution) by referencing single-cell RNA-seq experiments that include multiple minor cell types, for which other conventional methods—RCTD, SPOTlight, SpaCET, and cell2location—fail. Therefore, TD-based unsupervised FE can be used to perform deconvolution even when other conventional methods fail as it includes multiple minor cell types in the reference profiles, although it cannot be in usual cases. TD-based unsupervised FE is thus expected to be applied to a wide range of deconvolution applications.

Keywords: Visium, Spatial transcriptomics, tensor decomposition, deconvolution

1 Introduction

Spatial transcriptomics, involving spatial measurement of gene expression is currently a popular topic [1]. In addition to the development of measurement technology, analytical technology has also been co-developed. Spatial transcriptomics, particularly the 10x Genomics Visium platform [2], has fundamentally transformed our ability to study tissue architecture by capturing both transcriptomic information and its spatial context. However, each Visium "spot" often encapsulates gene expression from multiple cells [2], complicating the direct interpretation of cell-type composition [3]. A host of computational deconvolution methods has emerged to address this issue, aiming to disentangle mixed spot-level signals into individual cellular components using single-cell RNA sequencing (scRNA-seq) references or other probabilistic and machine learning paradigms [3].

A key example is Robust Cell Type Decomposition (RCTD), which employs a probabilistic framework to robustly assign cell-type identities to spots, modeling both biological and technical variability [4]. SPOTlight adopts a seeded non-negative matrix factorization (NMF) regression strategy using scRNA-seq data to deconvolute spot-level gene expression and provide interpretable spatial maps of cell-type distributions [5]. SpaCET combines deconvolution with spatial autocorrelation analysis, illuminating local cell neighborhoods and potential interactions within tissue sections [6]. Meanwhile, cell2location uses a Bayesian approach to jointly model cell abundance and gene expression, inferring the spatial organization of fine-grained cell types across Visium spots [7].

Beyond these four applications, several methods have become integral to the spatial deconvolution toolkit. Stereoscope employs variational inference to probabilistically reconcile scRNA-seq references with spatial transcriptomics data, thus estimating the presence and abundance of different cell types at each spot [8]. Tangram leverages deep learning to align high-dimensional single-cell RNA-seq profiles with Visium spots, inferring how various cell types and subtypes map onto histological structures [9]. Additionally, integrative pipelines within widely used bioinformatics frameworks (e.g., Seurat [10–12], Giotto [13]) facilitate spot deconvolution in tandem with downstream analyses, such as trajectory inference and cell-cell interaction studies. Together, these methods provide a dynamic and continuously expanding arsenal for uncovering cellular heterogeneity and spatial organization, enabling unprecedented insight into tissue biology across both healthy and disease contexts.

On this occasion, we notice that some of these tools fail to deconvolute Visium when the reference single-cell gene expression includes multiple minor cell types (see the later part of this study). Two kinds of single-cell gene expression are used as a reference. One is the reference for general use, i.e, not specific to the targeted Visium data and another is specific to the targeted Visium data. The latter is particularly important, for example, in precision medicine intended for personal therapy. Thus, although deconvolution is expected to be accurate, especially when the reference single-cell gene expression is associated with the targeted Visium dataset, this is not the case when reference single-cell gene expression includes multiple minor cell types. To address these problems, we developed a strategy where TD based unsupervised FE [14] is applied to the Visium dataset. In this study, after demonstrating that four

of conventional methods, RCTD, SPOTlight, SpaCET, and cell2location fail when reference single-cell gene expression includes multiple minor cell types, we demonstrate the success of our proposed strategy.

2 Results

2.1 Failure of existing methods

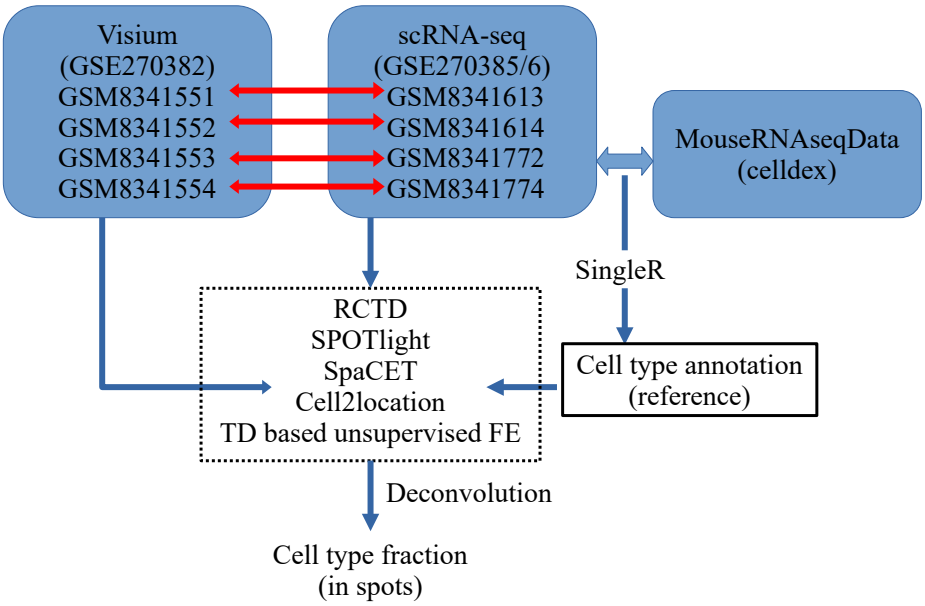

Fig. 1 Flow chart of the analyses used in this study. Single-cell RNA-seq profiles are annotated using SingleR with the reference of MouseRNAseqData from celldex. Annotated cell types and single-cell RNA-seq are used to deconvolute Visium datasets with RCTD, SPOTlight, SpaCET, cell2location, and TD-based unsupervised FE. Horizontal red arrows show the correspondence between Visium datasets and reference single-cell RNA-seqs.

First, we would like to demonstrate how poorly deconvolution is performed by some of conventional methods when reference single-cell gene expression includes multiple minor cell types (Table 1). As observed, neurons, micoroglia, and oligodendrocytes form the majority cell, consistent with the conclusion of the original paper [15]. Thus, the primary test of the following deconvolution methods is whether the majority cell types inferred by several methods include these three cell types. Below, deconvolution is performed with four conventional methods, RCTD [4], SPOTlight [5], SpaCET [6] and cell2location [7], which were briefly introduced in the Introduction as well as with TD-based unsupervised FE. Figure 1 shows the outline of this analysis.

No.	k	1	2	3	4
1	Astrocytes	15	8	50	21
2	B cells	4	2	0	1
3	Epithelial cells	1	0	1	0
4	Fibroblasts	8	4	7	1
5	Macrophages	3	4	1	0
6	Microglia	3273	2950	1414	1697
7	Neurons	4638	3956	3987	4590
8	NK cells	2	1	1	0
9	Oligodendrocytes	6476	5365	3829	4274
10	T cells	30	16	39	11
11	Granulocytes	0	4	17	0
12	Monocytes	0	1	3	0
13	Dendritic cells	0	0	1	0
14	Endothelial cells	0	0	1	0

Table 1 Fraction of cell-type annotation in reference single-cell gene expression associated with four Visium datasets ($1 \leq k \leq 4$).

2.1.1 RCTD

To determine whether conventional methods can perform deconvolution successfully even when the reference single-cell gene expression profile includes minor cell types as shown in Table 1, we first tested RCTD [4], one of the SOTA that aims for deconvolution (Fig. 2). The primary problem is that RCTD detects many astrocytes that were reported to be absent in the original paper [15]. Thus, it is obvious that RCTD is unsuccessful. The reason of this failure is that RCTD cannot consider cell populations in the reference single-cell gene expression. The model used in RCTD is as follows. Suppose that there are M spots where the expression of N genes is measured. Then x_{ijk} is modeled by RCTD as

$$x_{ijk}|\lambda_{ijk} \sim \text{Poisson}(X_{jk}\lambda_{ijk}) \tag{1}$$

$$\log(\lambda_{ijk}) = \alpha_{jk} + \log\left(\sum_{s=1}^{S_k} \beta_{jsk}\mu_{sik} + \gamma_{ik} + \epsilon_{ijk}\right) \tag{2}$$

where X_{jk} is the total transcript count in the jth spot of the kth Visium dataset, S_k is the number of cell types present in the kth Visium dataset, α_{jk} is a fixed spot-specific effect, μ_{sik} is the mean gene expression profile for cell type s and gene i, β_{jsk} is the proportion of the contribution of cell type s to spot j, γ_{ik} is a gene-specific platform random effect, ϵ_{ijk} is a random effect to account for other sources of variation, such as spatial effects. As RCTD does not consider cell population in the reference, it cannot infer cell populations in individual spots of Visium data based on the cell populations in the reference. This is why RCTD wrongly counts too many (not existing) astrocyte populations in individual spots even though astrocytes are actually missing.

2.1.2 SPOTlight

Next we employed SPOTlight [5] (Fig. 3). As seen in Fig. 3, for $k = 2, 3$, the cell proportion is too heterogeneous to be true. This is possibly because of the marker

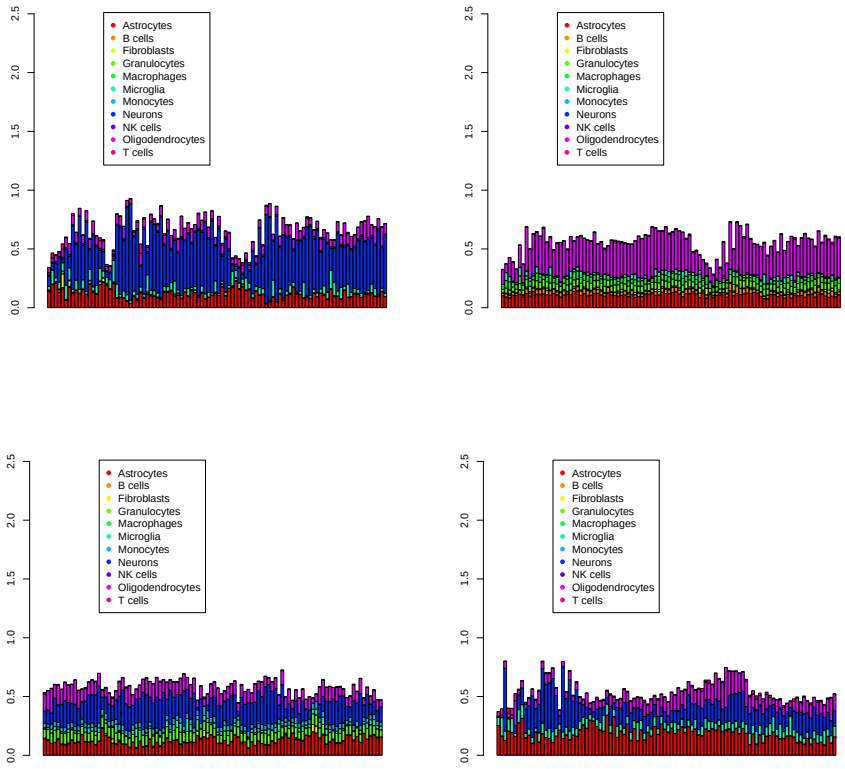

Fig. 2 The first 100 cell populations for four Visium datasets ($1 \leq k \leq 4$) provided by RCTD.

gene identification step in SPOTlight that selects marker genes prior to deconvolution. SPOTlignt selects marker genes that are distinct among all cell types as much as possible; this results in an unintended emphasis toward minor cell types and unrealistic heterogenous cell proportions. It is thus obvious that SPOTlight is also unsuccessful.

2.1.3 SpaCET

Although SPOTlight and RCTD are SOTA, because they are a bit old fashioned, we employ SpaCET [6] as a new alternative to RCTD and SPOTlight. Although SpaCET has been designed to process tumors, it can process gene expression other than that of tumors when reference-single-cell gene expression profiles associated with Visium data are available (Fig. 4). As none of the three major cell types in four Visium datasets are neurons, micoroglia, and oligodendrocytes and cell fractions are too distinct between individual Visium datasets, SpaCET is hardly successful. The reason why SpaCET fails is because it considers only correlation coefficients of gene expression between

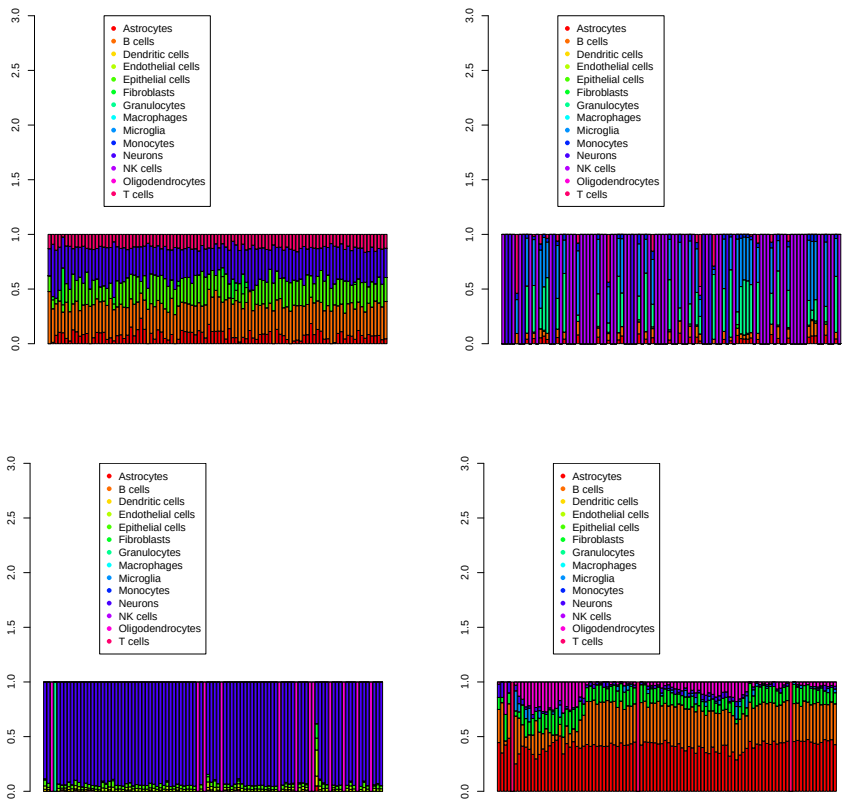

Fig. 3 The first 100 cell populations for four Visium datasets ($1 \leq k \leq 4$) provided by SPOTlight.

individual single cells in the reference and the spots in Visium that cannot consider the cell type fraction in reference single cell profiles as in the case of RCTD.

2.1.4 cell2location

Finally, we tested the more advanced, Bayesian-based method, cell2location [7] (Fig. 5). Although the Bayesian method is expected to be more accurate and robust, it is very time consuming as it must optimize probability itself (and cell2location is greatly time consuming). As none of three majority cell types in four Visium datasets are neurons, microglia, and oligodendrocytes, cell2location is unlikely to be successful despite a very long running time (it takes a few days whereas other three finish less than at most a few tens of minutes).

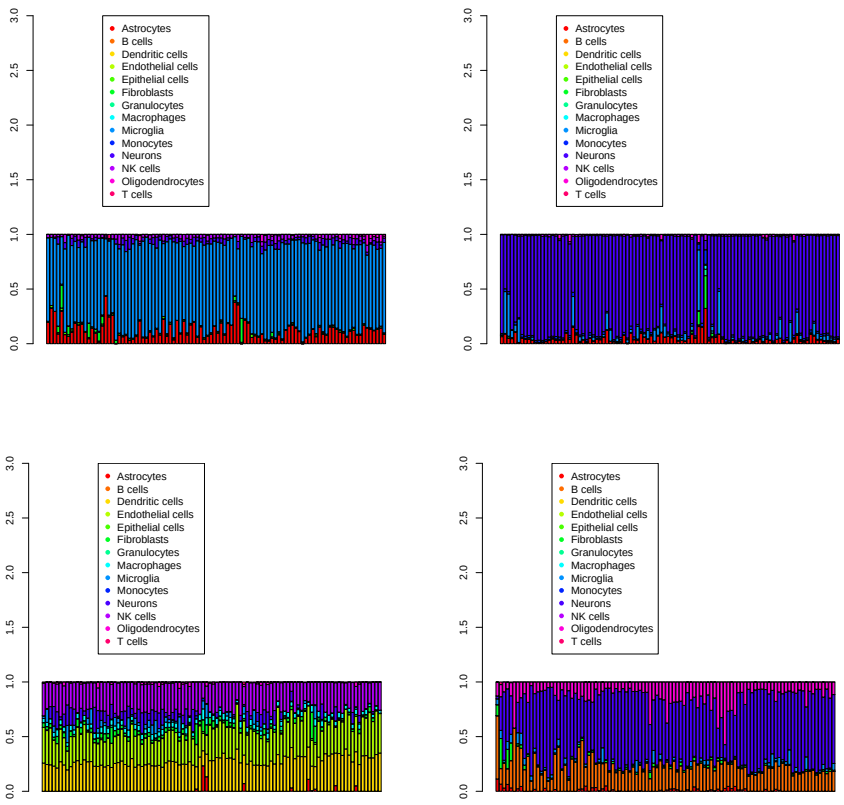

Fig. 4 The first 100 cell populations for four Visium datasets ($1 \leq k \leq 4$) provided by SpaCET.

2.2 TD based unsupervised FE applied to Visium

To determine whether TD-based unsupervised FE can outperform the above four methods, we applied TD- based unsupervised FE to Visium data (To see how we can get the result in this section, see the Discussion section). Figure 6 shows $c_{j's}$ (cell type profiles computed using TD-based unsupervised FE. See below for more details) for $1 \leq k \leq 4$. It is clear that only three cell types, "Microglia" "Neurons" and "Oligodendrocytes" are dominant, consistent with the original study [15]. Thus our inference of cell types is reasonable and TD-based unsupervised FE is the only method to successfully process Visium datasets even when the associated reference profiles include many minor cell types.

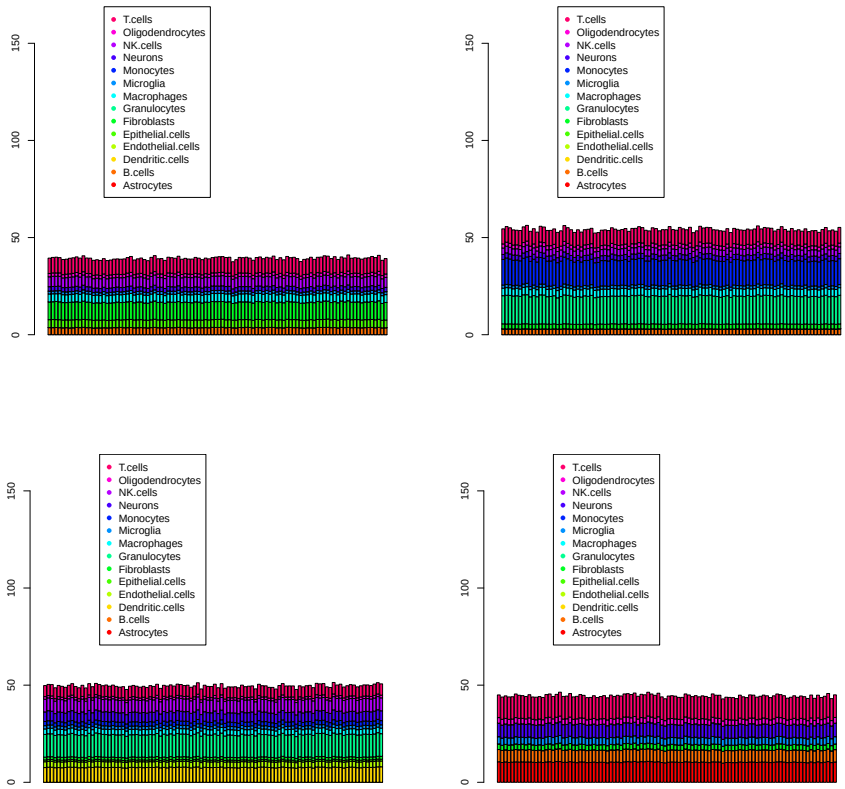

Fig. 5 The first 100 cell population for four Visium datasets ($1 \leq k \leq 4$) provided by cell2location.

3 Discussion

To see how we can reach the results in Fig. 6, we explain how we apply TD based unsupervised FE to the Visium datasets, step by step (Fig. 7). First, we apply TD-based unsupervised FE, obtain $v_{\ell_1 j'}$, and plot $v_{2j'}$ on the spatial coordinate (Fig. 8). Here the j'th spots in $1 \leq j' - 4992(k-1) \leq 4992$ belong to the kth Visium datasets. As nearby spots share the similar amount of v_{2j}, i.e., consistent with the spatial coordinate, we conclude that $\ell_2 = 2$ is the informative and variable component. Then using u_{2i} corresponding to v_{2j}, we attributed P-values to individual genes (is) and selected 277 genes to identify genes whose expression was expected to be consistent with the spatial coordinate. We can expect these 277 genes are consistent with the spatial coordinate as well as biologically important.

Next, we would like to see if TD-based unsupervised FE applied to Visium can be used for biological interpretation, which is unclear because individual spots in Visium include multiple cells. Without annotations of individual spots, interpreting outcomes

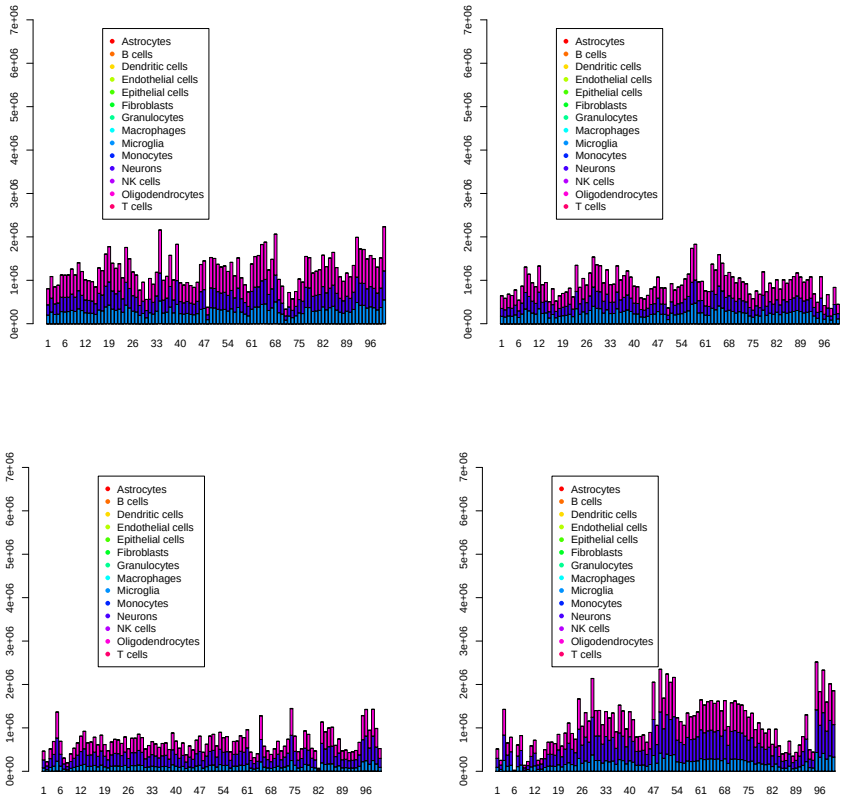

Fig. 6 The first 100 $c_{j's}$ for four Visium datasets ($1 \leq k \leq 4$) by-TD based unsupervised FE.

is difficult. One possible annotation to be attributed to individual spots is the cell fraction, i.e., deconvolution. Fortunately, as we have associated the single-cell gene expression profile, spots can be annotated based on single-cell RNA-seq. To do this, we need to annotate individual cells in single-cell RNA-seq (see Methods).

Next we decided to employ 277 genes as biomarker genes to relate the gene expression of individual spots to that of single-cell RNA-seq. We computed SVD, $u^{\mathrm{sc}}_{\ell_1 ik}$, of single-cell RNA-seq using only 277 genes. Linear regression

$$x_{ij'} = x_{i[j'-4992(k-1)]k} = a_{j'0} + \sum_{\ell_1=1}^{5} a_{j'\ell_1} u^{\mathrm{sc}}_{\ell_1 ik} \tag{3}$$

was performed to evaluate if 277 gene expression profiles in the Visium dataset can be represented by SVD computed from single-cell RNA-seq where $a_{\ell_1 j'}, 1 \leq \ell_1 \leq 5$ are regression coefficients. Figure 9 shows the spatial distribution of $v^{\mathrm{sel}}_{1j'}$ and the

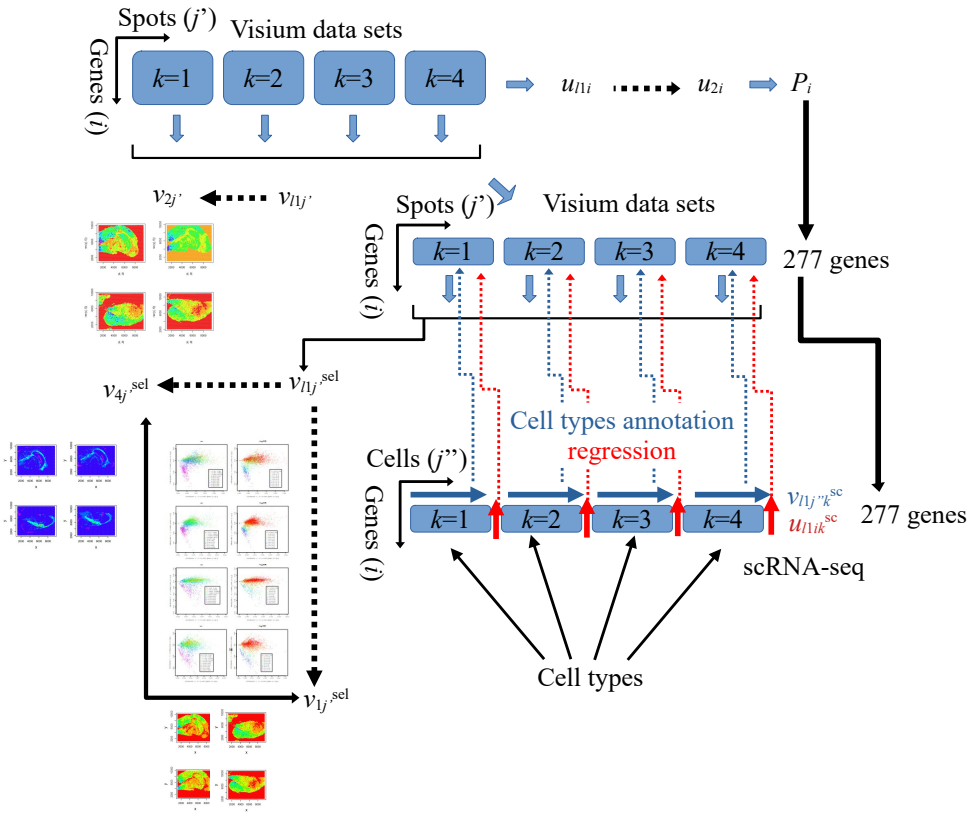

Fig. 7 First, we integrate four Visium datasets and get $v_{\ell_1 j'}$ to see if it is consistent with the spatial coordinate. Then, as we found that $v_{2j'}$ is consistent with the spatial coordinate, the corresponding u_{2i} is used to select 277 genes with adjusted P-values less than 0.01. Using 277 genes, $v^{\text{sel}}_{\ell_1 j'}$ is computed and $v^{\text{sel}}_{1j'}$ is still consistent with the spatial coordinate. Regression analysis is then performed wherein $x_{ij'}$ is fitted with $u^{\text{sc}}_{\ell_1 ik}$ and the spot-associated significant P-values are selected. P-values are especially related to $v^{\text{sel}}_{4j'}$. Finally, cell types of spots associated with significant adjusted P-values are inferred based on $v^{\text{sc}}_{\ell_1 j''k}$

Spearman correlation coefficient between $x_{ij'}$ and $a_{j'0} + \sum_{\ell_1=1}^{5} a_{j'\ell_1} u^{\text{sc}}_{\ell_1 ik}$, i.e., the left and right hand sides of eq. (3). The reason why not $v^{\text{sel}}_{2j'}$ but $v^{\text{sel}}_{1j'}$ is because $v^{\text{sel}}_{1j'}$ does not exhibit consistency with the spatial coordinate whereas $v^{\text{sel}}_{2j'}$ does. Not only is $v^{\text{sel}}_{1j'}$ still consistent with the spatial coordinate, $x_{ij'}$ is well computed from $v^{\text{sc}}_{\ell_1 jk}$. Table 2 lists the number of spots associated significant correlation between $x_{ij'}$ and $a_{j'0} + \sum_{\ell_1=1}^{5} a_{j'\ell_1} u^{\text{sc}}_{\ell_1 ik}$. It is obvious that the gene expression of most spots has a significant correlation with SVD computed from single RNA-seq.

Next, we would like to know the condition wherein spots are associated with significant correlation. For this purpose, we compute the Spearman correlation coefficient

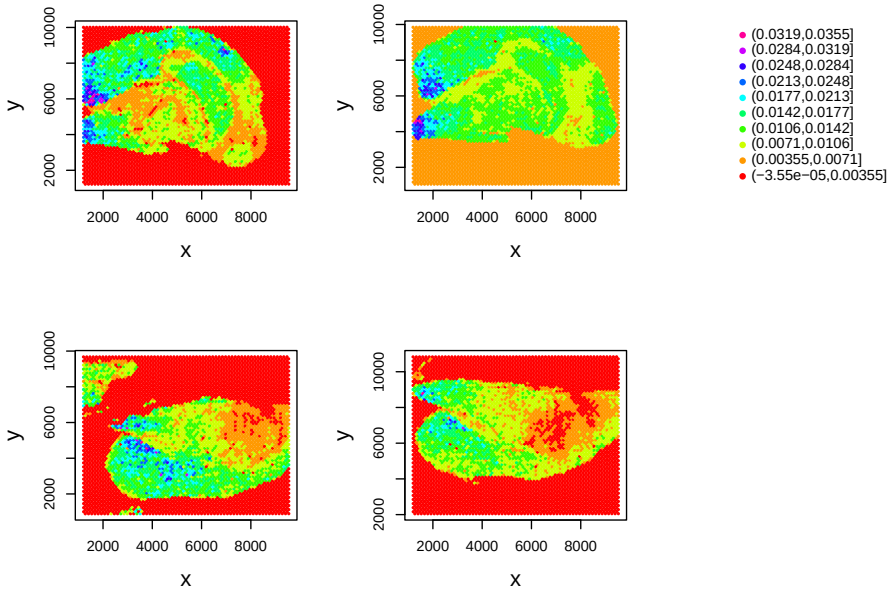

Fig. 8 $v_{2j'}$ attributed to the location of the j'th spot. Top left: $k = 1, 1 \leq j' \leq 4992$. Top right: $k = 2, 4993 \leq j' \leq 9984$. Bottom left: $k = 3, 9985 \leq j' \leq 14976$. Bottom right: $k = 4, 14977 \leq j' \leq 19968$.

Table 2 the number of spots associated significant correlation

	Adjusted P-values	
k	> 0.05	≤ 0.05
1	971	2076
2	1271	1769
3	1261	1349
4	1230	1351

between $v^{\text{sel}}_{\ell_1 j'}$ and negative signed logarithmic P-values of linear regression. We find that $v^{\text{sel}}_{4j'}$ has the largest correlation coefficient (Fig. 10). Figure 11 shows the scatter plot between $v^{\text{sel}}_{1j'}$ and $v^{\text{sel}}_{4j'}$ where colors correspond to the absolute Spearman correlation coefficient between $x_{ij'}$ and $a_{j'0} + \sum_{\ell_1=1}^{5} a_{j'\ell_1} u^{\text{sc}}_{\ell_1 ik}$ or negative-signed logarithmic P-values associated with the Spearman correlation coefficient. It is obvious that the negatively larger $v^{\text{sel}}_{4j'}$ is associated with more significant correlation. To determine whether $v^{\text{sel}}_{4j'}$ is still coincident with the spatial coordinate, we display $v^{\text{sel}}_{4j'}$ on spots (Fig. 12). It is thus obvious that $v^{\text{sel}}_{4j'}$ is still coincident with spatial coordinates.

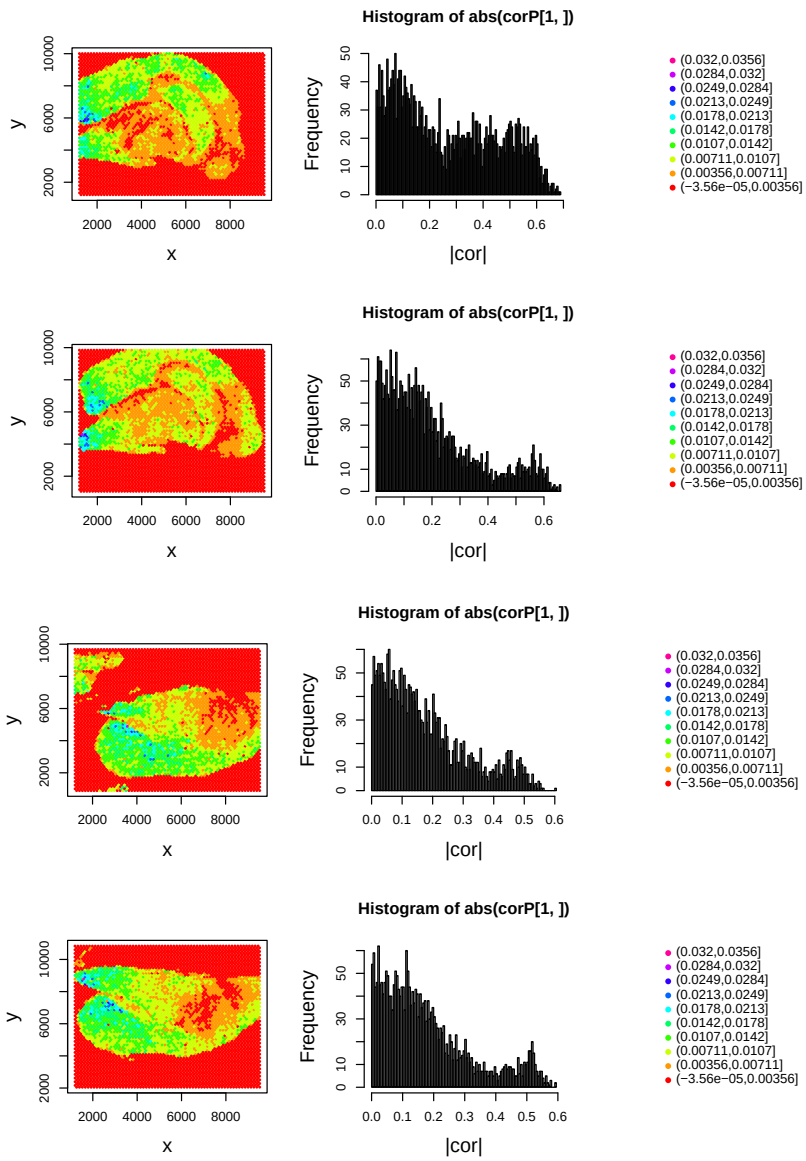

Fig. 9 Left column: $v_{1j'}^{\text{sel}}$ attributed to the location of the j'th spot. Right column: Histogram of the Spearman correlation coefficient between $x_{ij'}$ and $a_{j'0} + \sum_{\ell_1=1}^{5} a_{j'\ell_1} u_{\ell_1 ik}^{\text{sc}}$. The first row: $k = 1, 1 \leq j' \leq 4992$. The second row: $k = 2, 4993 \leq j' \leq 9984$. The third row: $k = 3, 9985 \leq j' \leq 14976$. The fourth row: $k = 4, 14977 \leq j' \leq 19968$.

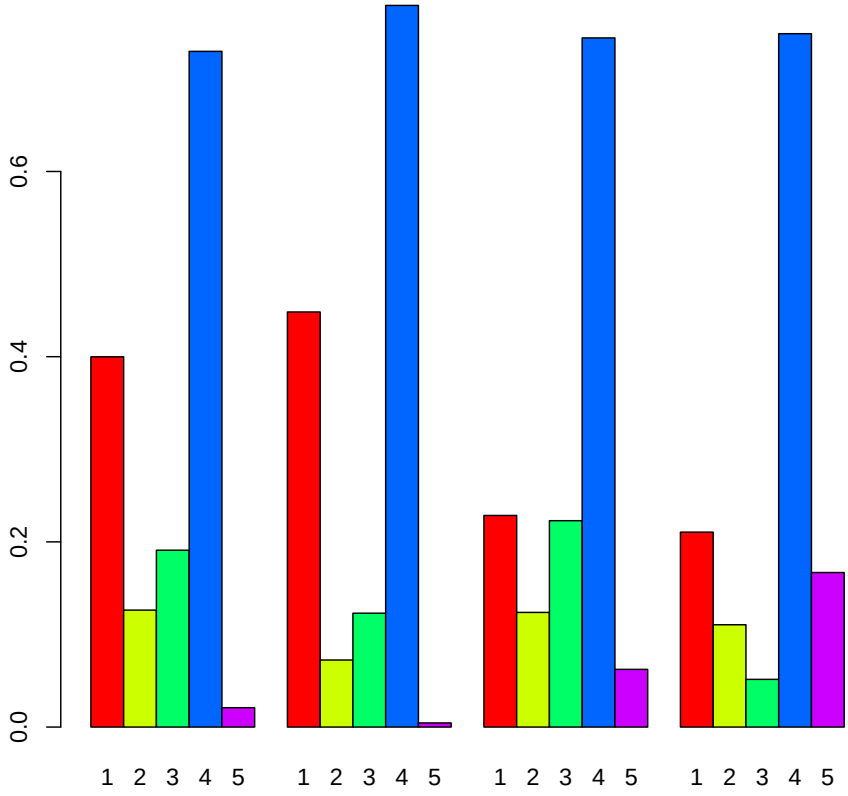

Fig. 10 Absolute Spearman correlation coefficient between $v^{\mathrm{sel}}_{\ell_1 j}$ and negative-signed logarithmic P-values of linear regression, eq. (3). Numbers below horizontal axis correspond to ℓ_1. From the left to right, $k = 1, 2, 3, 4$.

Then the fraction of the sth cell type, $c_{j's}$, of the j'th spot is computed as

$$c_{j's} = a_{j'0} + \sum_{\ell_1=1}^{5} \sum_{j''=1}^{M_k} a_{j'\ell_1} v^{\mathrm{sc}}_{\ell_1 j'' k} \delta_{j''s} \tag{4}$$

where $\delta_{j''s}$ takes 1 only when the j''th cell in single-cell RNA-seq belongs to the sth cell type otherwise 0. M_k is the total number of single cells in the kth single-cell RNA-seq. $a_{j'\ell_1}$s are those computed in eq. (3). $c_{j's}$ was computed only for spots associated with adjusted P-values less than 0.05 in Table 2. This is the result shown in Fig. 6.

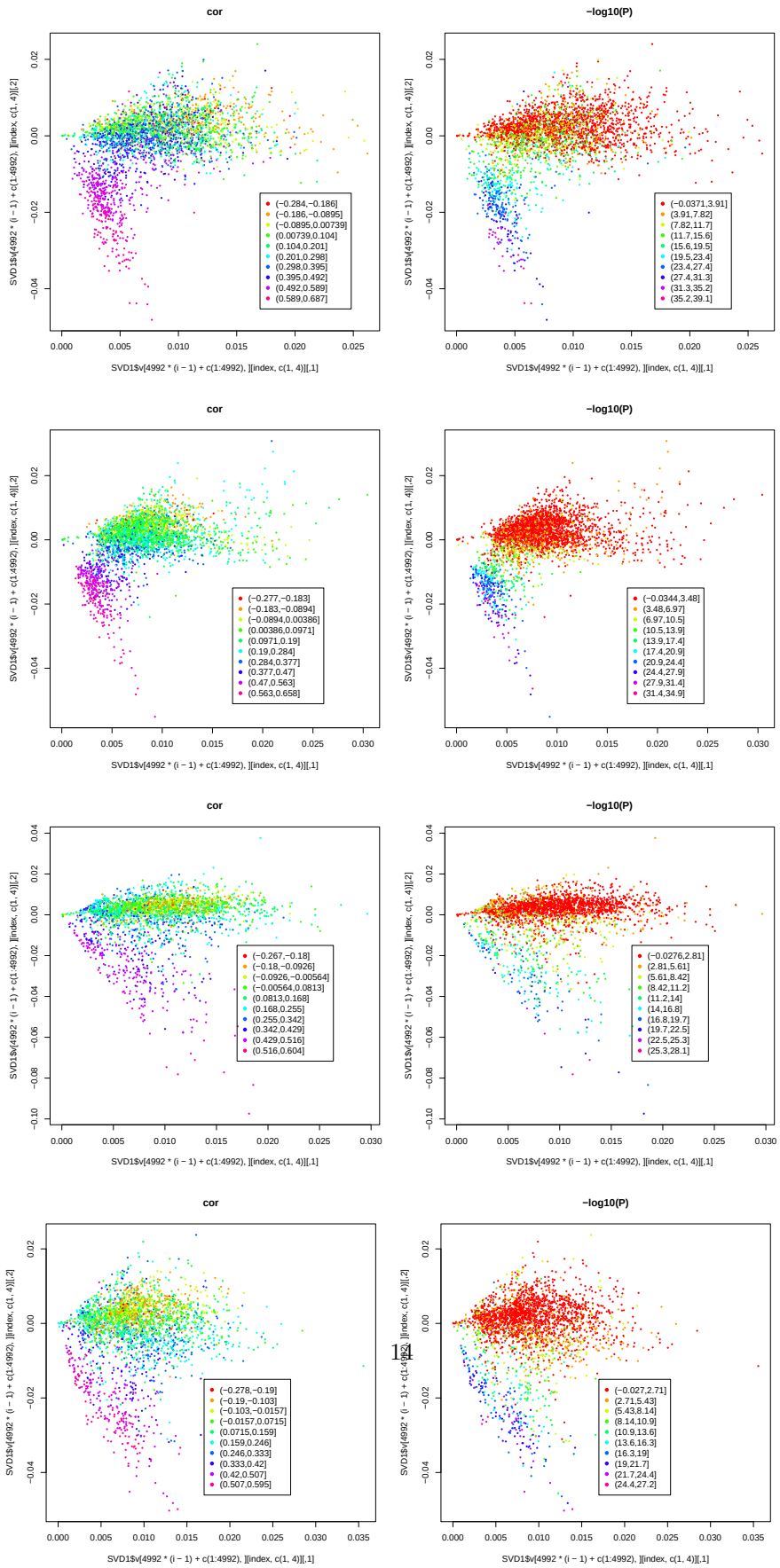

Fig. 11 Scatter plot between $v_{1j'}^{\mathrm{sel}}$ (horizontal axis) and $v_{4j'}^{\mathrm{sel}}$ (vertical axis) where colors in the left column correspond to the absolute Spearman correlation coefficient between $x_{ij'}$ and $a_{j'0} + \sum_{\ell_1=1}^{5} a_{j'\ell_1} u_{\ell_1 ik}^{\mathrm{sc}}$ and those in the right column are negative-signed logarithmic P-values associated with the Spearman correlation coefficient. The first row: $k = 1, 1 \leq j' \leq 4992$. The second row: $k = 2, 4993 \leq j' \leq 9984$. The third row: $k = 3, 9985 \leq j' \leq 14976$. The fourth row: $k = 4, 14977 \leq j' \leq 19968$.

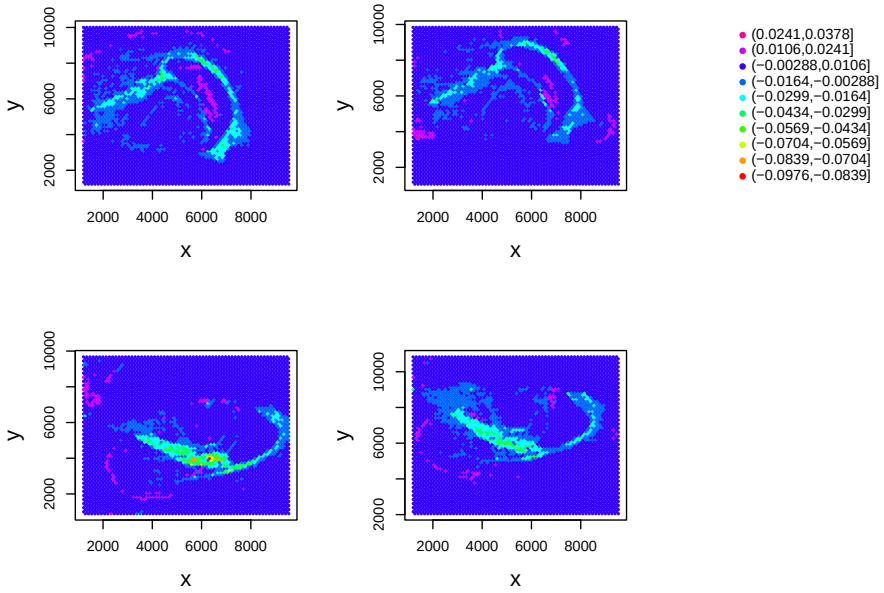

Fig. 12 $v_{4j'}^{\mathrm{sel}}$ attributed to the location of the j'th spot. Top left: $k = 1, 1 \leq j' \leq 4992$. Top right: $k = 2, 4993 \leq j' \leq 9984$. Bottom left: $k = 3, 9985 \leq j' \leq 14976$. Bottom right: $k = 4, 14977 \leq j' \leq 19968$.

Although one might wonder why we did not employ more established benchmark to test the proposed method, we have already found that TD cannot compete with the other SOTA when the established benchmark [16] is employed. Thus, we have decided to insist that TD is superior to SOTA only when multiple minor cell types in the reference single cell profiles. Although one might also wonder why we did not check all the methods cited in the introduction for the comparison, since cell2location and RCTD are the top ranked methods [16] it is the waste of time to check other inferior methods.

4 Methods

4.1 Visium

The vVisium dataset [15] analyzed in this study was retrieved from GEO with GEO ID, GSE270382, which is composed of four Visium datasets, two of which are Tau-dKI, uninjected and other two of which are Tau-dKI, injected, respectively. The downloaded matrix files are loaded into R by `read.csv` command and are formatted as a sparse matrix format using `spMatrix` function in Matrix package.

4.2 Single-cell RNA-seq and annotated cell fraction reference

Single-cell RNA-seq [15] was retrieved from GEO with GEO ID, GSE270385 and GSE270386. Two datasets were taken from GSE270385 and two additional ones were taken from GSE270386 with GEO IDs, GSM8341772 and GSM8341774. The .csv files were loaded into R using the `read.csv` command. We also inferred the cell types of individual cells in single-cell RNA-seq using the `SingleR` function in the SingleR package [17] in R. Annotation as performed using the MouseRNAseqData() dataset in celldex package [18]. The generated cell type fraction is shown in Table 1.

4.3 Detailed setups and parameters of conventional methods

For all cases, individual Visium datasets were analyzed separately in contrast to TD-based unsupervised FE, which integrates four Visium datasets.

4.3.1 RCTD

Function `create.RCTD` was performed by max_cores = 2, CELL_MIN_INSTANCE=1 to ensure that RCTD does not ignore the minor cell types. Then the function `create.RUN` was performed in the doublet_mode = 'multi' to let RCTD to attribute more than two cell types to individual spots.

4.3.2 SPOTlight

SPOTlight was performed with the default settings (for more details, see the source code provided in the Github link)

4.3.3 SpaCET

SpaCET was also performed with the default settings (for more details, see the source code provided in the Github link)

4.3.4 cell2location

Model learning was performed using the `train` function with max_epochs=300, batch_size=2500, train_size=1, lr=0.002, accelerator="cpu". `Cell2location` to initialize Visium data cell type annotation was performed with N_cells_per_location=30, detection_alpha=200. Then the `train` function was performed with max_epochs=30000, batch_size=adata_vis.n_obs, train_size=1 (for more details, see the source code provided in the Github link).

4.4 TD based unsupervised FE

Suppose we have tensor $x_{ijk} \in \mathbb{R}^{31053\times4992\times4}$ which represents the ith gene expression at the jth spot of the kth Visium dataset. Higher order singular value decomposition (HOSVD) is applied to x_{ijk} and we get

$$x_{ijk} = \sum_{\ell_1=1}^{31053} \sum_{\ell_2=1}^{4992} \sum_{\ell_3=1}^{4} G(\ell_1\ell_2\ell_3) u_{\ell_1 i} u_{\ell_2 j} u_{\ell_3 k} \tag{5}$$

where $G \in \mathbb{R}^{31053\times4992\times4}$ is a core tensor that represents the contribution of $u_{\ell_1 i}u_{\ell_2 j}u_{\ell_3 k}$ toward x_{ijk}, $u_{\ell_1 i} \in \mathbb{R}^{31053\times31053}$, $u_{\ell_2 j} \in \mathbb{R}^{4992\times4992}$, $u_{\ell_3 k} \in \mathbb{R}^{4\times4}$ are singular value matrices and orthogonal matrices.

In this study, HOSVD was performed by applying the `irlba` function in the irlba package to unfolded matrix $x_{ij'}$ where $j' = j + 4992(k-1)$ as

$$x_{ijk} = x_{ij'} \in \mathbb{R}^{31053\times19968},\ (1 \le k \le 4) \tag{6}$$

and we got

$$x_{ij'} = \sum_{\ell_1} \lambda_{\ell_1} u_{\ell_1 i} v_{\ell_1 j'} \tag{7}$$

The reason why HOSVD can be replaced with SVD applied to the unfolded matrix is because HOSVD is nothing but SVD applied to the unfolded matrix.

Spots specifically contributing to the ℓth singular value vector, $u_{\ell i}$, can be selected by attributing P-values to the ith gene by assuming that the null hypothesis $u_{\ell i}$ obeys Gaussian distribution,

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{\ell_1 i}}{\sigma_{\ell_1}}\right)^2\right] \tag{8}$$

where $P_{\chi^2}[> x]$ is the cumulative χ^2 distribution where the argument is larger than x and σ_{ℓ_1} is the standard deviation. The obtained P-values are corrected by the BH criterion and the ith genes associated with adjusted P-values less than 0.01 are selected.

In the above computation, σ_{ℓ_1} is optimized such that $u_{\ell_1 i}$ obeys Gaussian distribution as much as possible. The procedure is as follows. First, we need to compute P_i assuming some σ_{ℓ_1}. Then P_is are corrected with the BH criterion, and ith genes associated with adjusted P_i less than 0.01 are excluded. Next, we compute the histogram of P_i for the non-excluded genes and the standard deviation of the histogram, σ_h, as

$$\langle h_s \rangle = \frac{1}{S}\sum_{s=1}^{S} h_s \tag{9}$$

$$\sigma_h = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(h_s - \langle h_s \rangle\right)^2} \tag{10}$$

where h_s is the histogram of P_i at the sth bin and S is the total number of bins. We tune σ_{ℓ_1} so as to minimize σ_h.

The reason why we need to optimize σ_s is as follows. As σ_{ℓ_1} should be the standard deviation of $u_{\ell_1 i}$ supposed to obey the null hypothesis, Gaussian. Thus, σ_{ℓ_1} should be computed using only non-excluded genes. If we do not exclude $u_{\ell_1 i}$s associated with smaller adjusted P-values, because $u_{\ell_1 i}$s associated with smaller adjusted P-values have larger absolute $u_{\ell_1 i}$, σ_{ℓ_1} is over estimated and P_is are also overestimated. This results in a fewer number of selected genes. To avoid this situation, we need to compute σ_{ℓ_1} using only non-excluded $u_{\ell_1 i}$s.

5 Conclusion

In this paper, we demonstrated how TD-based unsupervised FE successfully performs deconvolution even when the reference single cell profile includes multiple minor cell types, for which four conventional methods fail. We also demonstrate that the identification of singular value vectors consistent with spatial cell type distribution may be the reason of the successful deconvolution.

Declarations

- Funding
  This study was supported by a KAKENHI Grant (Number 24K15168).

- Conflict of interest/Competing interests
  We declare that the authors have no conflicts of interest.
- Ethics approval and consent to participate
  Not applicable
- Consent for publication
  Not applicable
- Data availability
  All data analyzed in this study are from GEO and are publicly available.
- Materials availability
  Not applicable
- Code availability
  https://github.com/tagtag/TDbasedUFE_deconvolution
- Author contribution
  Y. H. T. planned the study and performed the analyses. Y.-H.T. and T.T. evaluated the results and wrote and reviewed the manuscript. All authors have read and agreed to the published version of the manuscript. The conceptualization, data curation, analysis were performed by Y. H. T.

References

[1] Williams, C.G., Lee, H.J., Asatsuma, T., Vento-Tormo, R., Haque, A.: An introduction to spatial transcriptomics for biomedical research. Genome Medicine **14**(1), 68 (2022) https://doi.org/10.1186/s13073-022-01075-1

[2] Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg, Pontén, F., Costea, P.I., Sahlén, P., Mulder, J., Bergmann, O., Lundeberg, J., Frisén, J.: Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science **353**(6294), 78–82 (2016) https://doi.org/10.1126/science.aaf2403 https://www.science.org/doi/pdf/10.1126/science.aaf2403

[3] Li, B., Dong, X., Zhang, J., Sui, T., Colozo, J.M., Ko, A.T., Qiu, W., Li, X., Chang, A., Ni, S., Li, S., Gherardini, P., Marton, R., Hashemi, B., Chen, L.,

Collado-Torres, L., Li, L., Fei, Y., Zhu, J., Seykora, J.T., Kim, B.J., Litt, B., Fullerton, J.M., Storm, P.B., Dahlquist, J.Y., He, X., Li, M.: Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. Nature Methods **19**(5), 662–670 (2022) https://doi.org/10.1038/s41592-022-01420-x

[4] Cable, D.M., Murray, E., Zou, L.S., Goeva, A., Macosko, E.Z., Chen, F., Irizarry, R.A.: Robust decomposition of cell type mixtures in spatial transcriptomics. Nature Biotechnology **40**(4), 517–526 (2022) https://doi.org/10.1038/s41587-021-00830-w

[5] Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I., Heyn, H.: Spotlight: seeded nmf regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. Nucleic Acids Research **49**(9), 50–50 (2021) https://doi.org/10.1093/nar/gkab043 https://academic.oup.com/nar/article-pdf/49/9/e50/60906660/gkab043.pdf

[6] Ru, B., Huang, J., Zhang, Y., Aldape, K., Jiang, P.: Estimation of cell lineages in tumors from spatial transcriptomics data. Nature Communications **14**(1), 568 (2023) https://doi.org/10.1038/s41467-023-36062-6

[7] Kleshchevnikov, V., Shmatko, A., Dann, E., Aivazidis, A., King, H.W., Li, T., Elmentaite, R., Lomakin, A., Kedlian, V., Gayoso, A., Jain, M.S., Park, J.S., Ramona, L., Tuck, E., Arutyunyan, A., Vento-Tormo, R., Gerstung, M., James, L., Stegle, O., Bayraktar, O.A.: Cell2location maps fine-grained cell types in spatial transcriptomics. Nature Biotechnology **40**(5), 661–671 (2022) https://doi.org/10.1038/s41587-021-01139-4

[8] Andersson, A., Bergenstråhle, J., Asp, M., Bergenstråhle, L., Jurek, A., Navarro, J.F., Lundeberg, J.: Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. Communications Biology **3**(1), 565 (2020) https://doi.org/10.1038/s42003-020-01247-y

[9] Biancalani, T., Scalia, G., Buffoni, L., Avasthi, R., Lu, Z., Sanger, A., Tokcan, N., Vanderburg, C.R., Segerstolpe, .A., Zhang, M., Avraham-Davidi, I., Vickovic, S., Nitzan, M., Ma, S., Subramanian, A., Lipinski, M., Buenrostro, J., Brown, N.B., Fanelli, D., Zhuang, X., Macosko, E.Z., Regev, A.: Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. Nature Methods **18**(11), 1352–1362 (2021) https://doi.org/10.1038/s41592-021-01264-7

[10] Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., Regev, A.: Spatial reconstruction of single-cell gene expression data. Nature Biotechnology **33**(5), 495–502 (2015) https://doi.org/10.1038/nbt.3192

[11] Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W.M., Hao, Y., Stoeckius, M., Smibert, P., Satija, R.: Comprehensive integration of single-cell data. Cell **177**(7), 1888–190221 (2019) https://doi.org/10.1016/j.

cell.2019.05.031

[12] Hao, Y., Hao, S., Andersen-Nissen, E., Mauck III, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., Granja, J.M., Hoffman, P., Stoeckius, M., Papalexi, E., Srivastava, A., Mimitou, E.P., Satija, R.: Integrated analysis of multimodal single-cell data. Cell **184**(13), 3573–358729 (2021) https://doi.org/10.1016/j.cell.2021.04.048

[13] Dries, R., Zhu, Q., Eng, C.-H.L., Li, R., Liu, K., Fu, Y., Jiang, Z., Ricker, A., Barillas, D., Sampaio, P., *et al.*: Giotto: a toolbox for integrative analysis and visualization of spatial expression data. Genome Biology **22**(1), 78 (2021) https://doi.org/10.1186/s13059-021-02286-2

[14] Taguchi, Y.-h.: Unsupervised Feature Extraction Applied to Bioinformatics: A PCA Based and TD Based Approach, 2nd edn. Unsupervised and Semi-Supervised Learning. Springer, Switzland (2024)

[15] Nagata, K., Hashimoto, S., Joho, D., Fujioka, R., Matsuba, Y., Sekiguchi, M., Mihira, N., Motooka, D., Liu, Y.-C., Okuzaki, D., Kikuchi, M., Murayama, S., Saido, T.C., Kiyama, H., Sasaguri, H.: Tau accumulation induces microglial state alterations in alzheimer's disease model mice. eNeuro **11**(12) (2024) https://doi.org/10.1523/ENEURO.0260-24.2024 https://www.eneuro.org/content/11/12/ENEURO.0260-24.2024.full.pdf

[16] Sang-aram, C., Browaeys, R., Seurinck, R., Saeys, Y.: Spotless, a reproducible pipeline for benchmarking cell type deconvolution in spatial transcriptomics. eLife **12**, 88431 (2024) https://doi.org/10.7554/eLife.88431

[17] Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., Butte, A.J., Bhattacharya, M.: Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nature Immunology **20**, 163–172 (2019) https://doi.org/10.1038/s41590-018-0276-y

[18] Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., Butte, A.J., Bhattacharya, M.: Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. **20**, 163–172 (2019) https://doi.org/10.1038/s41590-018-0276-y