OXFORD

Briefings in Bioinformatics, 2022, **23**(3), 1–17
https://doi.org/10.1093/bib/bbac116
Problem Solving Protocol

SPCS: a spatial and pattern combined smoothing method for spatial transcriptomic expression

Yusong Liu, Tongxin Wang, Ben Duggan, Michael Sharpnack, Kun Huang, Jie Zhang, Xiufen Ye and Travis S. Johnson

Corresponding authors: Travis S. Johnson, Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN 46202, USA. Tel.: +1-317-278-5451; Fax: 317-274-2678. E-mail: johnstrs@iu.edu; Xiufen Ye, College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin, Heilongjiang 150001, China. Tel.: +86-451-8251-9410; Fax: +86-451-8251-9410. E-mail: yexiufen@hrbeu.edu.cn; Jie Zhang, Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN 46202, USA. Tel.: +1-317-274-2839; Fax: 317-321-2003. E-mail: jizhan@iu.edu

Abstract

High-dimensional, localized ribonucleic acid (RNA) sequencing is now possible owing to recent developments in spatial transcriptomics (ST). ST is based on highly multiplexed sequence analysis and uses barcodes to match the sequenced reads to their respective tissue locations. ST expression data suffer from high noise and dropout events; however, smoothing techniques have the promise to improve the data interpretability prior to performing downstream analyses. Single-cell RNA sequencing (scRNA-seq) data similarly suffer from these limitations, and smoothing methods developed for scRNA-seq can only utilize associations in transcriptome space (also known as one-factor smoothing methods). Since they do not account for spatial relationships, these one-factor smoothing methods cannot take full advantage of ST data. In this study, we present a novel two-factor smoothing technique, spatial and pattern combined smoothing (SPCS), that employs the *k*-nearest neighbor (kNN) technique to utilize information from transcriptome and spatial relationships. By performing SPCS on multiple ST slides from pancreatic ductal adenocarcinoma (PDAC), dorsolateral prefrontal cortex (DLPFC) and simulated high-grade serous ovarian cancer (HGSOC) datasets, smoothed ST slides have better separability, partition accuracy and biological interpretability than the ones smoothed by preexisting one-factor methods. Source code of SPCS is provided in Github (https://github.com/Usos/SPCS).

Keywords: spatial transcriptomics, imputation, two-factor expression smoothing, *k*-nearest neighbors, tissue region partition, pancreatic ductal adenocarcinoma, dorsolateral prefrontal cortex, high-grade serous ovarian cancer

Introduction

Mammalian tissue is highly heterogeneous with phenotypes that depend on their spatial distribution [1, 2]. Until recently, studies of tissue heterogeneity have either sacrificed spatial relationships (e.g. scRNA-seq) or produced low-dimensional measurements [e.g. immunohistochemistry (IHC)] [3–7]. Novel ST techniques allow whole transcriptome profiles to be measured while preserving spatial relationships [8, 9]. These techniques have already been profoundly useful in understanding tumor [10–12] and non-tumor tissue [13–16] heterogeneity. However, improvements in ST library preparation [17], sequencing techniques and bioinformatic analysis pipelines [18] are still necessary and ongoing in contrast to more established scRNA-seq standard practice protocols [19, 20].

The most widely utilized ST technologies are based on highly multiplexed sequence barcoding, which suffers from expression noise and dropout events [21, 22]. Barcoding-based scRNA-seq data suffer from similar

Yusong Liu is a PhD student at the College of Intelligent Systems Science and Engineering, Harbin Engineering University. His research focus is on machine learning and network analysis in bioinformatics.
Dr Tongxin Wang is a recently graduated PhD student from the Department of Computer Science, Indiana University, who is currently employed as a research scientist at Facebook. His research focus is on deep learning, transfer learning and adversarial networks.
Ben Duggan is a medical student at the Indiana University School of Medicine. His research interests include bioinformatics, machine learning and using computing to improve patient care.
Dr Michael Sharpnack is a pathology resident at the Department of Pathology at the University of California San Francisco. His research is primarily concerned with novel immunotherapies, antigen presentation and bioinformatics methods development as it applies to lung cancer.
Dr Kun Huang is a professor and chair of the Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indiana University School of Medicine Precision Health Initiative Chair for Genomic Data Science, Director of Data Science and Informatics for the Precision Health Initiative, Associate Director of Data Science for the Indiana University Simon Comprehensive Cancer Center and an Investigator at the Regenstrief Institute. His main research interests are in medical image analysis, multi-omics and machine learning.
Dr Jie Zhang is an assistant professor of Medical and Molecular Genetics, who is also a member of the Center for Computational Biology and Bioinformatics. Her research interests are applied translational bioinformatics and systems biology for cancer and neurological disease.
Dr Xiufen Ye is a professor of College of Intelligent Systems Science and Engineering, Harbin Engineering University, IEEE Senior Member. Her main research interest includes image processing, pattern recognition and artificial intelligence.
Dr Travis S. Johnson is an assistant research professor at the Department of Biostatistics and Health Data Science, Indiana University School of Medicine. His research interests include applied machine learning for use with high-dimensional omic data as it applies to cancer and dementia research.
Received: October 27, 2021. **Revised:** February 24, 2022. **Accepted:** March 9, 2022

limitations, while plate- and *in vitro* transcription-based techniques, such as Smart-seq2 [23] and CEL-seq2 [24], respectively, provide more representative expression profiles per cell at the cost of fewer cells measured per experiment. As a result, a multitude of techniques have been developed to impute the missing expression values and smooth the noise that comes directly from the barcode-based non-spatial scRNA-seq. SAVER [25] and MAGIC [26] use sets of correlated genes and relative cell similarity in transcriptome space to impute the dropout events and eliminate other types of expression errors via machine learning techniques. These methods are termed as 'one-factor methods', given that they only incorporate expression values. The smoothened expression values give more accurate representations of the true underlying RNA abundances than the raw read counts. ST data have the advantage of providing spatial relationships that can be used in addition to transcriptomic similarity for smoothing based on the assumption that nearby cells will have more similar expression profiles than distant cells.

Here, we present a novel two-factor smoothing method, termed spatial and pattern combined smoothing, i.e. SPCS, specifically designed for ST data, which utilizes both the associations of spatial locations in transcriptome space (expression pattern knowledge) and in Euclidean space (spatial knowledge). By performing SPCS on multiple ST slides from PDAC, DLPFC and HGSOC datasets, smoothed ST slides have better separability, partition accuracy and biological interpretability than the ones smoothed by preexisting one-factor methods.

Methods

Datasets

The datasets that we use in this study include two real-world ST datasets, PDAC [10] and DLPFC [14], and a simulating dataset generated from HGSOC single-cell datasets [27]. For the two real-world datasets, PDAC includes 10 ST slides sourced from the traditional ST platform, while DLPFC is a Visium platform dataset with 12 slides. All the data in these datasets consist of two different matrices containing gene expressions and spatial coordinates. One matrix consists of the gene expression values for each spatial barcode hybridized from its corresponding spot on the ST slide. The other matrix contains the spatial locations in 2D space for each spot's spatial barcode. Using these two matrices, we can generate a 2D representation of each gene's expression value throughout the biopsied tissue section. Because these datasets are sourced from different ST platforms, i.e. traditional ST platform and new developed Visium platform, we can explore the influence of smoothing methods more comprehensively. A detailed statistical summary of the data we used is provided in Supplementary Table S1 available online at https://academic.oup.com/bib.

To better explore the ability of smoothing methods to deal with outlier spots, we designed a simulation experiment based on those used in the BayesSpace study [28]. Simulated ST data are based on HGSOC single-cell datasets and an immunofluorescence stained image of an ovarian cancer biopsy. In the original single-cell analysis of the HGSOC dataset, all the cells were divided into 15 clusters by the DBSCAN clustering method and annotated [27]. Considering the limited number of cells, we only used some of the slides. Ground-truth cluster labels were derived from single-cell level annotation of tumor and stroma compartments within the image. To make the simulated data reflect biology, we separated the slide into four clusters: intratumor (including dendritic and fibroblast cells), stroma (corresponding to macrophage cells) and two tumor clusters (associated with two different malignant cell clusters). Detailed information about ground-truth clusters is provided in Supplementary Table S2 available online at https://academic.oup.com/bib. To test the ability of different smoothing methods to find outlier spots, we randomly mixed 5% of other cell types as perturbation for each spatial cluster. We generated 10 sets of simulating data in the simulation analysis.

SPCS of spatial transcriptomic expression

For each spot on an ST slide, there exists not only the gene expression but also its spatial positions. This means we can improve the quality of the expression values within each specific spot using the relative similarity to the other spots based on both expression pattern and spatial location on the ST slide. To achieve this goal, we propose a kNN-based method, SPCS, to perform smoothing and padding. We display the procedure of the SPCS method in Figure 1.

In our method, we obtained the smoothed expression of each spot by integrating the contribution-weighted expression of its pattern and spatial neighbors. Let $\mathbf{X}_i$ be a vector of gene expression values for spot i, smoothed expression $\mathbf{X}'_i$ can be calculated by the following:

$$\mathbf{X}'_i = (1-\alpha)\,\mathbf{X}_i + \alpha\left(\beta\frac{\sum_{j\in\mathbf{N_s}(i)} c^{s}_{ji}\mathbf{X}_j}{\sum_{k\in\mathbf{N_s}(i)} c^{s}_{ki}} + (1-\beta)\frac{\sum_{j\in\mathbf{N_p}(i)} c^{p}_{ji}\mathbf{X}_j}{\sum_{k\in\mathbf{N_p}(i)} c^{p}_{ki}}\right). \quad (1)$$

In Equation (1), there are two ratio parameters α and β. α is used to regularize the ratio of original expression to the corrected expression, which avoids the expression of the object spot becoming over-smoothed. β is used to adjust the ratio of spatial-based to pattern-based smoothing for different applications. $\mathbf{N_s}(i)$ and $\mathbf{N_p}(i)$ are the spatial and pattern neighborhoods of the object spot i, respectively. The size of the neighborhoods can be determined by parameters τ_s and τ_p, which also should be specified in advance. c^{s}_{ji} and c^{p}_{ji} represent the spatial and pattern contributions of a corresponding neighbor

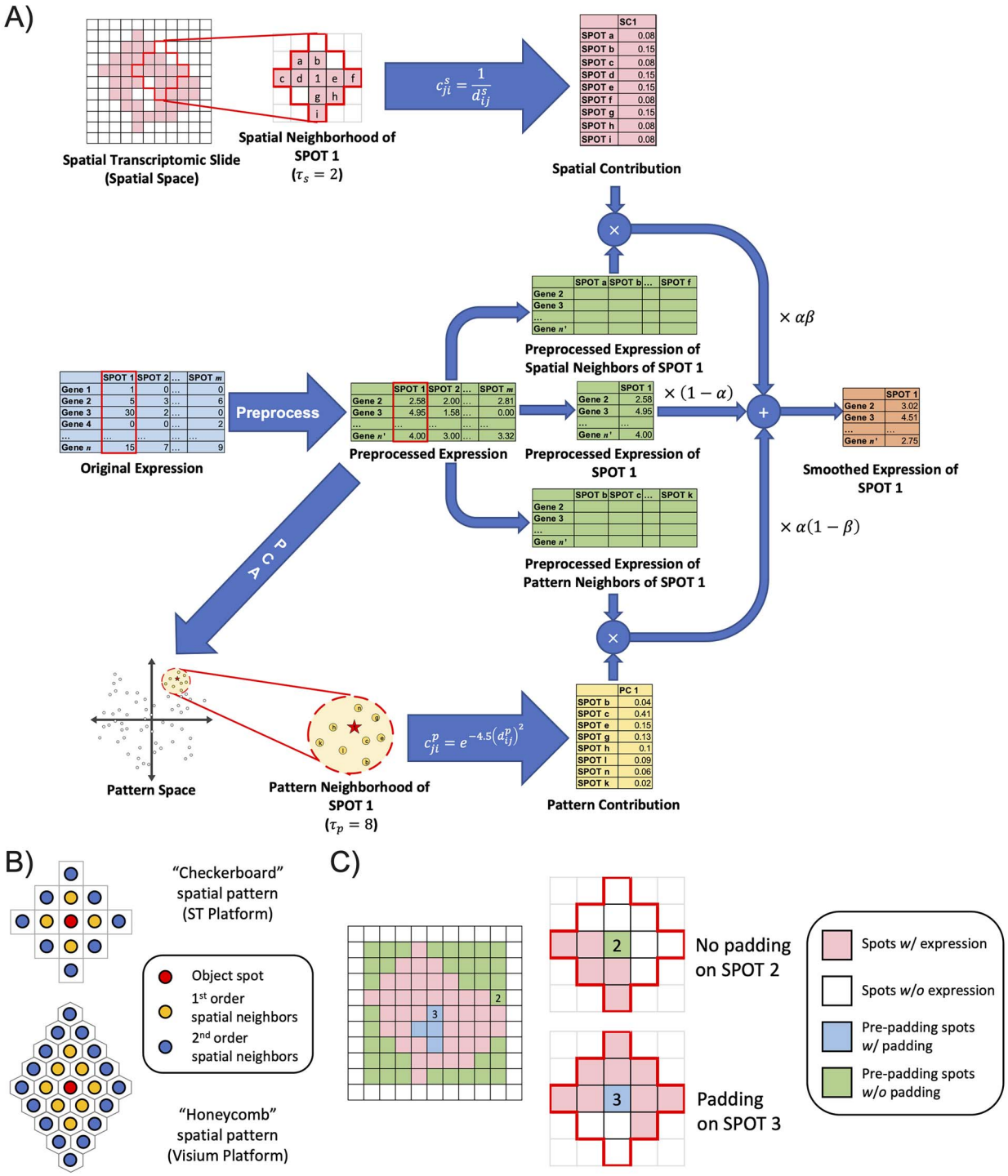

Figure 1. Workflow of our proposed SPCS smoothing method. (**A**) An example of smoothing. In this sample, SPOT 1 is the object spot that is going to be smoothed. For ST slides, preprocessing steps normalize the expressions and filter out genes with zero expression in most spots. Spatial neighborhood is second-order neighborhood of SPOT 1 with nine spots. Spots, where all gene expression is 0, are treated as non-tissue regions and are excluded from spatial neighborhood. Pattern neighborhood consists of the top eight spots with the most similar expression pattern to SPOT1. Both spatial and pattern contributions are normalized, which means the total contribution of all neighbors is =1. Smoothing is performed by integrating parameter- and contribution-weighted expression of the object spot itself and both its spatial and pattern neighbors. (**B**) Shape of second-order spatial neighborhood for ST and Visium platforms. Spatial neighborhood is determined by Manhattan distance. Spots with the same Manhattan distance to object spot belong to the same order of spatial neighborhood. Due to the different geometries of spots, the shape of spatial neighborhood could be different in different platforms. (**C**) Padding strategy of SPCS. A missing spot will not be padded unless there are >50% of non-missing spots inside its spatial neighborhood.

spot j to the object spot i. For traditional ST platform-based PDAC dataset, we set $\alpha = 0.6, \beta = 0.4, \tau_s = 2$ and $\tau_p = 16$; and for the Visium platform-based DLPFC and simulating datasets, we make $\tau_s = 4$ due to the larger number of spots. We will display the influence of these parameters on data separability and discuss the selection of them in the Discussion section. In addition, SPCS will also fill the missing spots using multiple non-missing spatial neighbors. The source code of our proposed SPCS method is provided in our Github repository (https://github.com/Usos/SPCS). In the next section, we introduce the detailed mathematical definitions of neighborhood, neighborhood contribution and our missing spot padding strategy.

Pattern neighborhood

ST data are a type of transcriptomic data that measure gene expression patterns similar to scRNA-seq. Like some scRNA-seq data, where cells can be localized to tissue location of origin, spots that have a similar expression pattern are more likely to belong to the same region in a tissue. Therefore, smoothing the expression of a given spot using other spots with a similar gene expression can improve data quality and is similar to one-factor smoothing methods [25, 26] designed for scRNA-seq. A group of the most similar spots based on the expression 'pattern' of the spot can be defined as that spot's 'pattern neighborhood'. Here, we provide the explicit definition of the pattern neighborhood used by the SPCS method.

Definition 1 (Pattern neighborhood): $\mathbf{S_p}$ is the gene expression pattern space of an ST slide; $i, j, k \in \mathbf{S_p}$ are different spots; $d_{ij}^{\mathbf{p}}$ and $d_{ik}^{\mathbf{p}}$ are pattern distances between spots i, j and i, k, respectively. $\mathbf{N_p}(i)$ is the τ_p pattern neighborhood of spot i if $\mid \mathbf{N_p}(i) \mid = \tau_p, \forall j \in \mathbf{N_p}(i), \forall k \in \mathbf{S_p} - (\mathbf{N_p}(i) \cup \{i\})$ s.t. $d_{ij}^{\mathbf{p}} < d_{ik}^{\mathbf{p}}$.

For gene expression data, the overall shapes of gene expression patterns are of greater interest than the individual magnitudes of each feature [29]. Hence, we used the Pearson correlation distance to measure the pattern distance between different spots. Let ρ_{ij} represents Pearson correlation coefficient (PCC) of coordinate of spots i and j in pattern space, Pearson correlation distance d_{ij} of spots i and j is given as follows [30]:

$$d_{ij} = 1 - \rho_{ij}. \tag{2}$$

In ST data, some genes are expressed at identical or near-identical levels that lack the variance to establish an accurate pattern neighborhood. Therefore, we used principal component analysis (PCA) [31] to transform the expression of spots into a 10D principal component space before smoothing. These uncorrelated components with the largest variance from our PCA are considered as the gene expression pattern space.

Spatial neighborhood

In contrast to scRNA-seq data, ST data provide the spatial position for each spot in the slide. Regions in proximity on histopathology slides are more likely to be the same tissue type. Aside from the pattern associations between spots, we can also use spatial associations to smooth the expression as a second factor. We define the group of spots that are spatially near a given spot as the 'spatial neighborhood' of that spot, which is defined explicitly below.

Definition 2 (Spatial neighborhood): $\mathbf{S}$ represents the set of spatial location indices of an ST slide, and $i, j \in \mathbf{S}$, $d_{ij}^{\mathbf{s}}$ is the spatial distance between spots i and j. $\mathbf{N_s}(i)$ is the τ_s spatial neighborhood of spot i, if $\forall j \in \mathbf{N_s}(i)$ s.t. $d_{ij}^{\mathbf{s}} \leq \tau_s$.

ST spots are spatially distributed in a checkerboard or honeycomb pattern. Due to the geometric patterns inherent to ST spot layout, Manhattan distance is a suitable metric to measure the spatial distance between spots inside an ST slide. Thus, we chose Manhattan distance as the spatial distance to define our spatial neighborhood. Due to the difference in the spatial pattern of spots, the shape of spatial neighborhood could be different in different platforms. Figure 1**B** illustrates the shape of second-order neighborhood of both traditional ST platform (checkerboard spatial pattern) and Visium platform (honeycomb spatial pattern).

Contribution of neighbors on smoothing

Different neighbors in the spatial or pattern neighborhood will have different impacts on the smoothing for a given spot, which we refer to as 'contribution'. Since the definitions of spatial and pattern distance are different, we model the corresponding contributions in different ways. The contributions of both spatial and pattern neighbors are still comparable since the range of both is $[0, 1]$. For spots outside the neighborhood (both spatial and pattern) of object spot, we assigned their corresponding contribution to 0, which means they have no contribution to smoothing of the object spot.

In pattern space, to better capture global gene expression patterns, we used PCC distance described in Equation (2) as the distance metric, whose range is $[0, 2]$. If the expression of two spots has a negative correlation, the distance based on Equation (2) will become >1. However, smoothing with negative correlation spots is not performed since they are dissimilar. Therefore, we set the contribution to 0 if the pattern distance between the object spot and one of its neighbors is >1. We used an exponential transformation to achieve this goal. For object spot i and its pattern neighbor j, pattern contribution $c_{ji}^{\mathbf{p}}$ can be defined as follows:

$$c_{ji}^{\mathbf{p}} = \begin{cases} \exp\left(-\left(\frac{d_{ij}^{\mathbf{p}}}{\sigma}\right)^2\right) & d_{ij}^{\mathbf{p}} < 1 \\ 0 & \text{otherwise} \end{cases}. \tag{3}$$

The exponential function in Equation (3) limits the range of c_{ji}^{p} to [0, 1] and ensures that c_{ji}^{p} decreases as d_{ij}^{p} increases. σ in this equation is a tuning parameter that controls how the pattern contribution decays with pattern distance. When $d_{ij}^{p} > 3\sigma/\sqrt{2}$, c_{ji}^{p} will quickly decay to 0 [32, 33]. Hence, setting σ to $\sqrt{2}/3$ will remove the effect of negative correlation neighbors and we can simplify Equation (3) to

$$c_{ji}^{p} = \exp\left(-4.5\left(d_{ij}^{p}\right)^{2}\right). \quad (4)$$

For spatial neighbors, we used Manhattan distance, an integer >0, as spatial distance to measure their distance to the object spot. Since the inverse proportional function has a similar decay nature as exponential function, we used the inverse of spatial distance as the spatial contribution, i.e.

$$c_{ji}^{s} = \frac{1}{d_{ij}^{s}}. \quad (5)$$

In this case, i is the spot being smoothed and j is the spatial neighbor of spot i.

Padding of missing spots

For ST slides, there are two types of missing values. The first type of missing value is missing genes, which means the spot itself is located in the tissue region, but some specific gene expression is missing. This is the main form of dropout events and also frequently happens in single-cell data, which most of the smoothing methods can handle. The second type is missing spots, i.e. expression of all genes in the spot is 0. These kinds of absent data are unique in ST data and cannot be padded without spatial position information, which often indicates the spot has been removed due to a quality problem. Hence, it is necessary to judge whether the missing spot needs to be padded. As shown in Figure 1**C**, for each blank spot in a slide, SPCS will only pad the ones whose predetermined spatial neighborhood has >50% non-blank spots. This criterion ensures that the boundary of the tissue will not be erroneously expanded. Since there is no expression on missing spots at all, we estimate the expression of missing spots by their spatial neighbors only. Let i become a missing spot, and its expression X_i can be estimated by the following:

$$\mathbf{X}_i = \frac{\sum_{j\in\mathbf{N_S}(i)} c_{ji}^{\mathbf{s}}\mathbf{X}_j}{\sum_{k\in\mathbf{N_S}(i)} c_{ki}^{\mathbf{s}}}. \quad (6)$$

In Equation (6), $\mathbf{N_s}(i)$ is τ_s-order spatial neighborhood of spot i, c_{ji}^{s} and c_{ki}^{s} are spatial contributions of spot j and k to spot i. Spot padding is performed after non-missing spot smoothing. To evaluate the adjusted Rand index (ARI) score for padding spots, we assign ground-truth labels for them. The ground-truth labels of the padded spots are determined by spatial contribution-weighted major voting of their spatial neighbors and manually inspected by a pathology resident.

Performance evaluation

To evaluate the effectiveness of our proposed SPCS method, we first performed SPCS and other one-factor smoothing methods (SAVER and MAGIC) on PDAC, DLPFC, and simulated ST datasets. SAVER and MAGIC are two representative one-factor smoothing methods using different techniques (i.e. statistical model-based and kNN-based). Then, we partitioned both smoothed and the original unsmoothed real-world slides using the *K*-medoids clustering method [34] and judged how well the clusters were separated after smoothing by internal evaluation of the unsupervised clustering. In addition, we also explored how the parameters in SPCS will influence the smoothing. Next, we performed multiple unsupervised clustering methods, including *K*-medoids, Louvain [35], mclust [36], BayesSpace [28] and spaGCN [37], on all the smoothed and unsmoothed slides to find out how well the clusters match the histopathological labels from the corresponding image as an external evaluation. Simulating data were also used here to detect outliers for each smoothing method. As a gene filter, we only kept genes with <70% zero expressed spots in our analysis. For normalization, we performed logarithmic count per million normalization before smoothing using SPCS, SAVER and MAGIC. The distance metric used during clustering was Pearson correlation distance as described in Equation (2). For each dataset, the number of clusters is predetermined by ground truth, which is provided in Supplementary Tables S1 and S2 available online at https://academic.oup.com/bib. PCA was performed prior to clustering, and the eigenvectors with top 20 eigenvalues were selected to reduce the dimensions of expression matrices and to enhance the clustering results.

Internal evaluation

For the internal evaluation, silhouette score was used as the evaluation indicator [38]. Silhouette score, a metric whose range is [−1, 1], estimates the average distance between clusters. A greater silhouette score indicates better cluster separations. For an ST spot i, let $a(i)$ be the average distance between i and all the other ST spots within the same cluster and $b(i)$ be the smallest average distance between i and all the ST spots with any other clusters. The silhouette coefficient $S(i)$ of ST spot i can be expressed by the following:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}. \quad (7)$$

The silhouette score of an entire ST slide is the average silhouette coefficient of all spots in it.

In this work, Pearson correlation-based measurement in Equation (2) was used to calculate the dissimilarity

between ST spots from the imputed ST data. Traditionally, Euclidean distance is used, but here, we adopted the dissimilarity metric defined in Equation (2) since our clustering method is based on PCC distance. Padded spots are excluded in this analysis. To make the average silhouette coefficient of clustering under different smoothing methods comparable, we used the same distance matrix, which was based on dimensionality-reduced original unsmoothed ST expression. Hence, in our experiments, compared with unsmoothed slides, an increase or lack of change in silhouette score on smoothed slides represents the smoothing method has enhanced the original data distribution, while a decrease indicates the smoothing method has corrupted the original data distribution.

External evaluation

For the external evaluation, we obtained the histopathological labels of the ST spots. The correspondence between smoothed slides and histopathological labels was evaluated at both clustering and gene expression levels. At the clustering level, the imputed ST data are clustered using multiple clustering algorithms. Technical details of the clustering methods are shown in Supplementary Table S4 available online at https://academic.oup.com/bib. Next, the concordance was evaluated between the clusters from the imputed ST data and the labels from the corresponding histopathological images. Then, the distribution of marker gene expression was compared to the locations of histopathological labels.

The ARI was used to evaluate the similarity between the clustering results from the imputed ST data and the histopathological labels [39]. For clusters from the imputed ST data $\{\mathbf{I}_1, \mathbf{I}_2, \mathbf{I}_3, \mathbf{I}_4\}$ and the histopathological categories of ST spots $\{\mathbf{H}_1, \mathbf{H}_2, \mathbf{H}_3, \mathbf{H}_4\}$, we denoted n_{ij} as the number of ST spots that are in both $\mathbf{I}_i$ and $\mathbf{H}_j$. The ARI is defined as follows:

$$\mathrm{ARI} = \frac{\Sigma_{ij}\binom{n_{ij}}{2} - \frac{\Sigma_i\binom{n_i}{2}\Sigma_j\binom{n_j}{2}}{\binom{n}{2}}}{\frac{1}{2}\left[\Sigma_i\binom{n_i}{2} + \Sigma_j\binom{n_j}{2}\right] - \frac{\Sigma_i\binom{n_i}{2}\Sigma_j\binom{n_j}{2}}{\binom{n}{2}}}, \quad (8)$$

where n_i is the number of ST spots in $\mathbf{I}_i$ and n_j is the number of ST spots in $\mathbf{H}_j$. A higher ARI value indicates that the imputed ST clusters and the histopathological labels are more similar. In this analysis, padded spots that are unable to be clustered with the ground truth will be treated as an error cluster.

In the marker gene evaluation with the PDAC dataset, two marker genes, *PRSS1* and *TM4SF1*, were used to compare their expression distribution spatially. Both genes are protein-coding genes. *PRSS1* encodes a trypsinogen, which is often highly expressed in normal pancreatic tissues, while *TM4SF1* is a common proto-oncogene and is highly expressed in pancreatic cancer among other malignancies [40, 41]. High expression of *TM4SF1* in the cancerous regions of the PDAC dataset has been detected in previous research [10]. For the DLPFC dataset, three other marker genes, *MOBP*, *PCP4* and *SNAP25*, were used in the analysis. Previous research has reported that these genes can delineate different cortical layers [14]. The expression distribution of these representative genes can stratify the histopathological regions and can be used to measure the accuracy of that partition.

Biological analysis

To examine how our algorithm aids in informing the biology of an ST sample, we identified differentially expressed genes (DEGs) and performed gene ontology enrichment analysis (GOEA). DEGs were identified in the PDAC slides by comparing two groups of spots with *TM4SF1*-high (neoplastic tissue) and low expression (non-neoplastic tissue). To define the two groups, we first linearly transformed the expression values of *TM4SF1* between [0, 1] by dividing each value by the maximum expression value. The spots on each slide were split into two groups, with one having a transformed expression value <0.7 (*TM4SF1* under-expressed) and the other greater (*TM4SF1* over-expressed). The cutoff 0.7 was chosen as it can reflect the boundary of the cancer region accurately.

A Kruskal–Wallis test was performed between the over- and under-expressed groups, and the *P*-values were adjusted using the Benjamini–Hochberg method to account for the multiple comparisons [42]. Only genes with an adjusted *P*-value < 0.05 were included. In addition, we used logarithmic foldchange (logFC) to determine the up-regulated and down-regulated events for genes. Genes with a logFC > 1 were considered as up-regulated and as down-regulated if the logFC were <-1.

GOEA was performed on the DEGs using 'g:Profiler' [43]. The significance of enriched terms was tested by cumulative hypergeometric test, and *P*-values were corrected by g:SCS method [44]. Only terms with an adjusted *P*-value < 0.05 were reported. All data sources offered by g:Profiler were used, including gene ontology (GO) [45], Reactome [46], KEGG [47], WikiPathways [48], TRANSFAC [49], CORUM [50], Human Protein Atlas [51] and the Human Phenotype Ontology [52]. Heatmaps for each sample were then generated to compare the terms found by each smoothing algorithm.

The DLPFC slides were processed slightly differently than the PDAC slides. To find the DEGs, one layer of the cortex was compared with the other six layers. For example, the expression values for all the sports in Layer 1 (L1) were compared to the expression values for the spots in layers L2, L3, L4, L5, L6 and WM (white matter). A Kruskal–Wallis test was performed between one layer of

the cortex and all other layers, *P*-values corrected using the Benjamini–Hochberg method and logFC calculated. This resulted in seven sets of comparisons. The same cutoffs of *P*-value < 0.05 and logFC > 1 were used. The significant DEGs were then compared to a set of DLPFC layer marker genes published by Zeng *et al*. [53]. Enrichment analysis was performed using g:Profiler in the same way as the PDAC slides were processed.

Results

In this study, we applied our two-factor smoothing method SPCS and two state-of-the-art one-factor smoothing methods (MAGIC and SAVER) to smooth ST slides. We first evaluated the computational cost of SPCS and provided the results in Supplementary Table S3. To compare the performance of different smoothing methods, we evaluated the quality of generated clusters after performing unsupervised clustering on both smoothed and unsmoothed expressions. The generated clusters and the distribution of marker gene expression were compared with pre-labeled histopathological partitions to check if the smoothed expression more accurately reflected the pathology features of the corresponding images. In addition, DEGs were identified in different regions and GOEA was performed to reveal the biological meaning from the smoothed data.

Internal evaluation

The silhouette score indicates the average distance between clusters in a slide. Since we used the same unsmoothed distance matrix in the different smoothing methods, the silhouette score reflects how well the smoothing methods kept the original data distribution. For different smoothing methods, a greater silhouette score than an unsmoothed slide usually represents better separability of the smoothed expressions compared to the original data distribution. In contrast, a decreased silhouette score indicates the smoothing method has changed the original data distribution making it less separable. After clustering the spots into the corresponding clusters, we calculated the silhouette score of 10 PDAC slides and 8 DLPFC slides with 7 clusters for each smoothing method, as shown in Figure 2**A** and Supplementary Figure S1 available online at https://academic.oup.com/bib. In most of the slides from these two datasets, SPCS and MAGIC got a similar or even greater silhouette score to the unsmoothed slide, while SAVER got a significantly lower score. In addition, SPCS had the greatest average silhouette score over the other smoothing methods even slightly higher than the unsmoothed slide. This result indicates that kNN-based smoothing generally can keep the characteristics of original data distribution resulting in better data separability.

Figure 2**B** and **C** illustrates how the parameters influenced the data separability of SPCS smoothed PDACA1 slide. For the two ratio parameters α and β, while $\beta \leq 0.4$, silhouette score was not affected much by changing α, but the result became more sensitive to α once β is large. This result indicates that pattern-based smoothing can help keep the original data distribution. In addition, the results reveal that the data distribution of spatial neighbors is different from the pattern neighbors. For the neighborhood size parameters, the results show that τ_s and τ_p have no significant influence on data separability. However, while $\tau_s = 2$ and $\tau_p = 4$, silhouette score is significantly lower than other parameter combinations, which indicates that neighborhood size parameters also should be well tuned according to the dataset to avoid unexpected results.

External evaluation

In the external evaluation, different smoothing methods were evaluated based on the consistency of unsupervised clusters with histopathological labels. The result of PDACA1 alone is shown in Figure 3 since the histopathological labels were not available for other PDAC slides. This slide can be well clustered by *K*-medoids clustering even without smoothing. When examining the results in detail, Figure 3**B** shows that most ST spots in the cancer cells and desmoplasia region were separated from other clusters. Most ST spots in the duct epithelium region were also well separated, with slight mixing of the interstitium and the normal pancreatic tissue regions. It is worth mentioning that by including spatial position information, SPCS can pad missing spots, which other one-factor smoothing methods cannot do. Figure 3**C** and **D** shows the influence of different smoothing methods on marker gene expression. Compared with other smoothing methods, SPCS generated a better marker gene contrast between distinctive histopathological areas due to two reasons. First, marker gene expressions showed fewer dropouts and better-reflected expression patterns across different tissue regions with SAVER and SPCS (MAGIC does not impute the missing expressions). Second, and in contrast to SAVER, SPCS imputed the missing values and kept the spatial distribution of non-missing values stable.

Figure 4 illustrates external evaluation results from 12 slides in DLPFC dataset. Figure 4**A** shows the influence of smoothing methods on different clustering methods. To make the evaluation comprehensive, we choose three commonly used general clustering methods (i.e. *K*-medoids, Louvain and mclust) and two state-of-the-art clustering methods developed specifically for ST slides (i.e. BayesSpace and SpaGCN). The results revealed that smoothing achieved a higher ARI score using various clustering methods and that SPCS outperforms other one-factor smoothing methods for every clustering method. When combining SPCS with Louvain or mclust clustering methods on the DLPFC dataset, we got an ARI score near BayesSpace, indicating that the combination of SPCS with general clustering methods improves the performance of spatial clustering on ST slides.

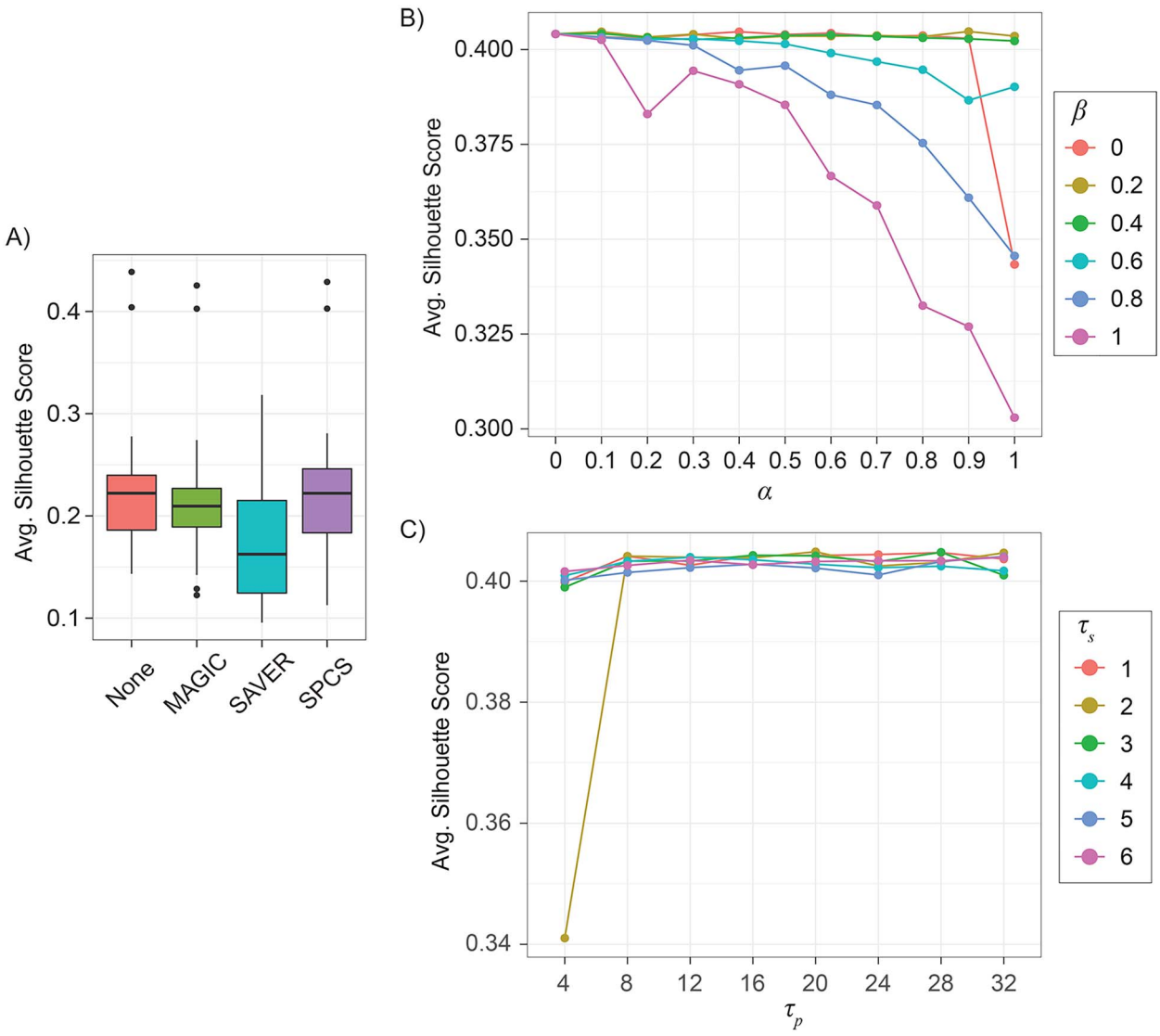

Figure 2. Influence of smoothing on data separability and data distribution. (**A**) Box plot of average silhouette score of 10 PDAC samples and 8 seven-layer DLPFC samples without smoothing and with smoothing by different methods (MAGIC, SAVER and SPCS). (**B**) Influence of parameters α and β on average silhouette score for SPCS smoothed PDACA1 slide when $\tau_p = 16, \tau_s = 2$. (**C**) Influence of parameters τ_p and τ_s on average silhouette score for SPCS smoothed PDACA1 slide when $\alpha = 0.6, \beta = 0.4$.

In addition, combining SPCS and BayesSpace got the highest ARI score among all the combinations. Due to the difference in preprocessing steps, the results of SpaGCN clustering were moderately lower than the original publication [37], which indicates that hyperparameter settings of SpaGCN are sensitive to preprocessing steps and we hope to optimize this in the future. Figure 4**B** and Supplementary Figure S2**A**, available online at https://academic.oup.com/bib, illustrate clustering ground-truth and BayesSpace clustering results combined with different smoothing methods on slides 151675 and 151673. Compared with other one-factor smoothing or without smoothing, SPCS can achieve more clear and accurate boundaries between different regions, which contribute to a higher ARI score. In addition, smaller spatial neighborhood (τ_s) for SPCS can help to capture long narrow regions in the slide but may cause over clustering in thicker regions with a similar length and width. Similar to the PDAC dataset, SPCS also helps to obtain a better marker gene contrast between distinctive cortical layers in DLPFC dataset as shown in Supplementary Figure S2**B**–**D** available online at https://academic.oup.com/bib.

We also performed simulation analysis to test the clustering accuracy for different smoothing methods. Louvain and BayesSpace, the two best general and ST dedicated clustering methods in the DLPFC experiment, were used in this analysis. The clustering ground truth is shown in Figure 5**A**. For the 10 simulated slides, Figure 5**B** illustrates the distribution of ARI score of each combination of smoothing and clustering method. In general, since the original single-cell dataset was well clustered, every combination of smoothing and clustering methods got high ARI scores in this experiment. With Louvain clustering, there was only a small difference in ARI scores on different slides. For different smoothing methods,

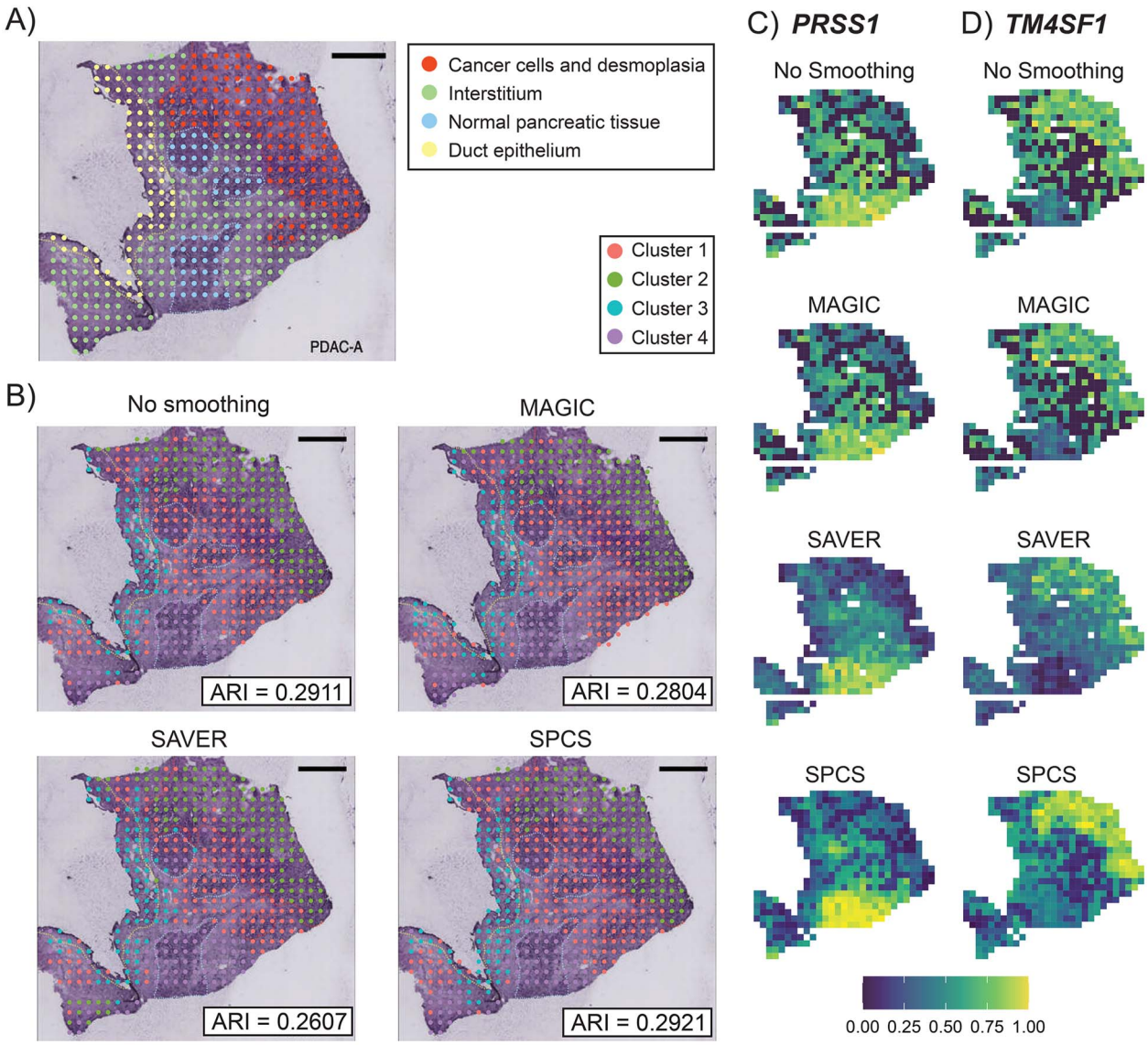

Figure 3. Influence of smoothing on clustering accuracy in PDACA1. (**A**) Original ST slide of PDACA1 and its histopathological partitions. (**B**) Results of *K*-medoids clustering on expression smoothed by different methods. ARI score, which reflects the correlation between clusters and histopathological labels, is marked at bottom-right corner of each figure. Clusters are ordered by their size. Heatmap of smoothed expression of two marker genes, (**C**) *PRSS1* and (**D**) *TM4SF1*, are also shown. For demonstration purposes, expressions of genes are linearly transformed into the range of 0–1 as normalization.

SPCS gets a slightly higher average ARI score, which indicates that spatial information helps general clustering methods to cluster ST slides. However, Louvain separates the spots into five clusters instead of four, which means Louvain tends to cluster the spots according to cell types. For the BayesSpace experiments, the ARI scores were lower than Louvain on average and were greater in variance. To better review the higher variance of BayesSpace, we illustrate BayesSpace clustering results of two simulating slides in Figure 5**C** and **D**. Obviously, BayesSpace tends to merge small clusters into nearby larger ones, which leads to oversmoothing in the simulated dataset, and smoothing with SAVER and SPCS will aggravate this problem. Overall, all methods performed well on the simulated data (ARI > 0.85), but it is worth noting that SPCS has a higher potential ARI as evaluated by the 75th percentile. This is surprising considering the other smoothing methods are specifically designed for scRNA-seq data from which these simulated ST slides are generated.

Biological analysis

The biological interpretability of the smoothed results was compared between different smoothing methods. Comparing between *TM4SF1* over- and under-expressed regions for the PDAC slides, the significant DEG numbers are displayed in Figure 6**A**. There were no DEGs found in any smoothed or unsmoothed slides of PDACB2 and PDACG because they lacked *TM4SF1*, while a higher number of DEGs were detected by SPCS in six out of the rest eight ST slides. Correspondingly, the number of enriched GO terms was also more from SPCS than the other methods, as shown in Figure 6**B**, which are further examined below.

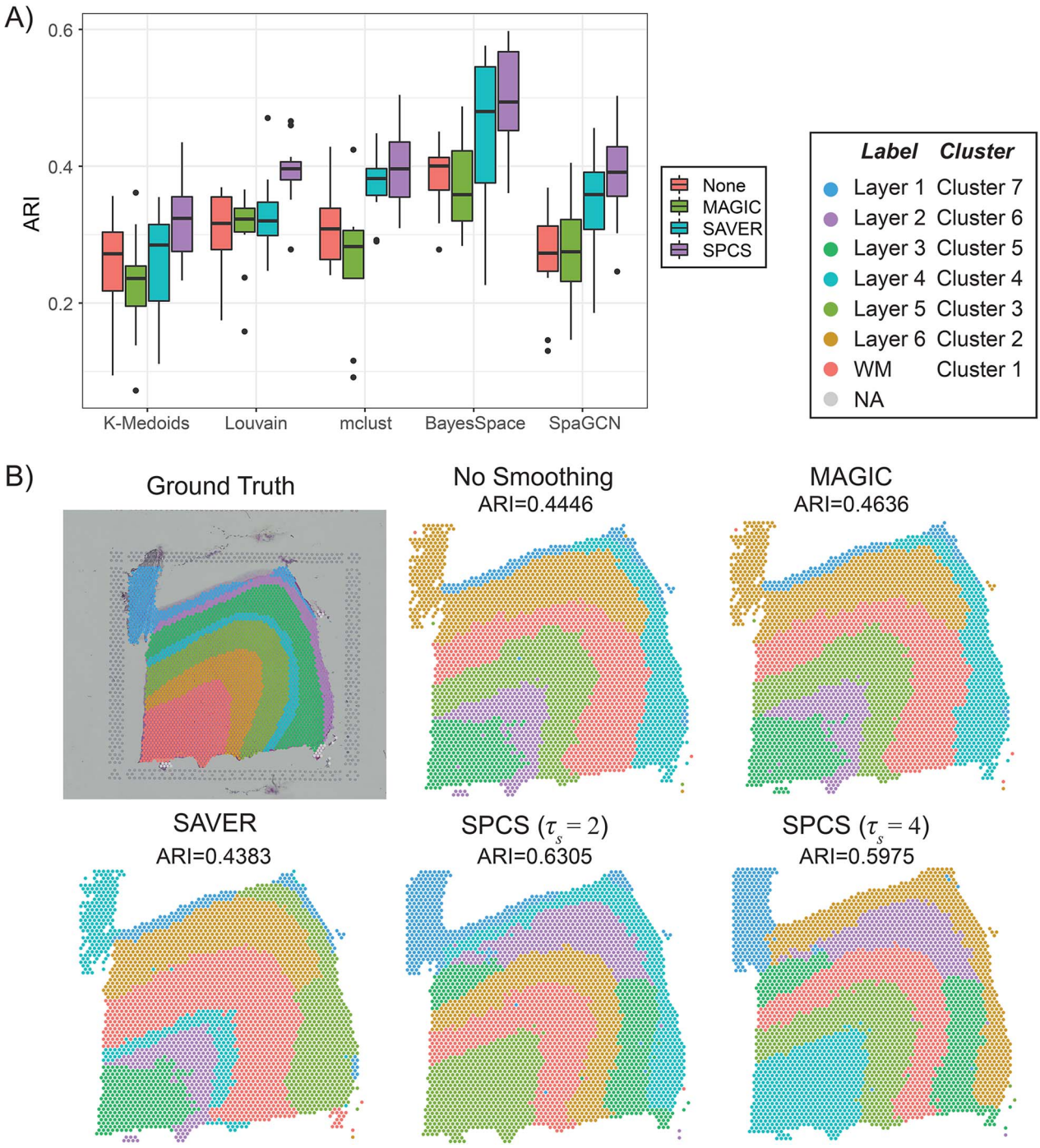

Figure 4. Influence of smoothing on clustering accuracy in DLPFC dataset. (**A**) Box plot of ARI score on various combinations of smoothing (No smoothing, MAGIC, SAVER and SPCS) and clustering methods (K-medoids, Louvain, mclust, BayesSpace and SpaGCN) in 12 DLPFC samples. (**B**) Ground-truth label and BayesSpace clustering results of smoothed and unsmoothed sample 151675. Clusters are ordered by their size.

The top 10 most significant GO terms found in slide PDACA1 for each algorithm are shown in Figure 6**C** for the up-regulated DEGs and in Figure 6**D** for the down-regulated DEGs. More than 10 terms are shown in the heatmap because most GO terms were not shared between smoothing methods. Without smoothing the slides, the up-regulated DEGs identified are related to interleukin-1. Instead, the terms found by SPCS smoothing are related primarily to cell adhesion, extracellular matrix (ECM) organization and MET-activated PTK2 signaling. The GO terms from all smoothed and unsmoothed up-regulated slides can be seen in Supplementary Figure S3 available online at https://academic.oup.com/bib. g:Profiler could not find any enriched terms for slides PDACB2, PDACD and PDACG. Up-regulated enriched terms for all SPCS slides except PDACB1 appear similar to PDACA1. The results for the other methods had a small number of terms which

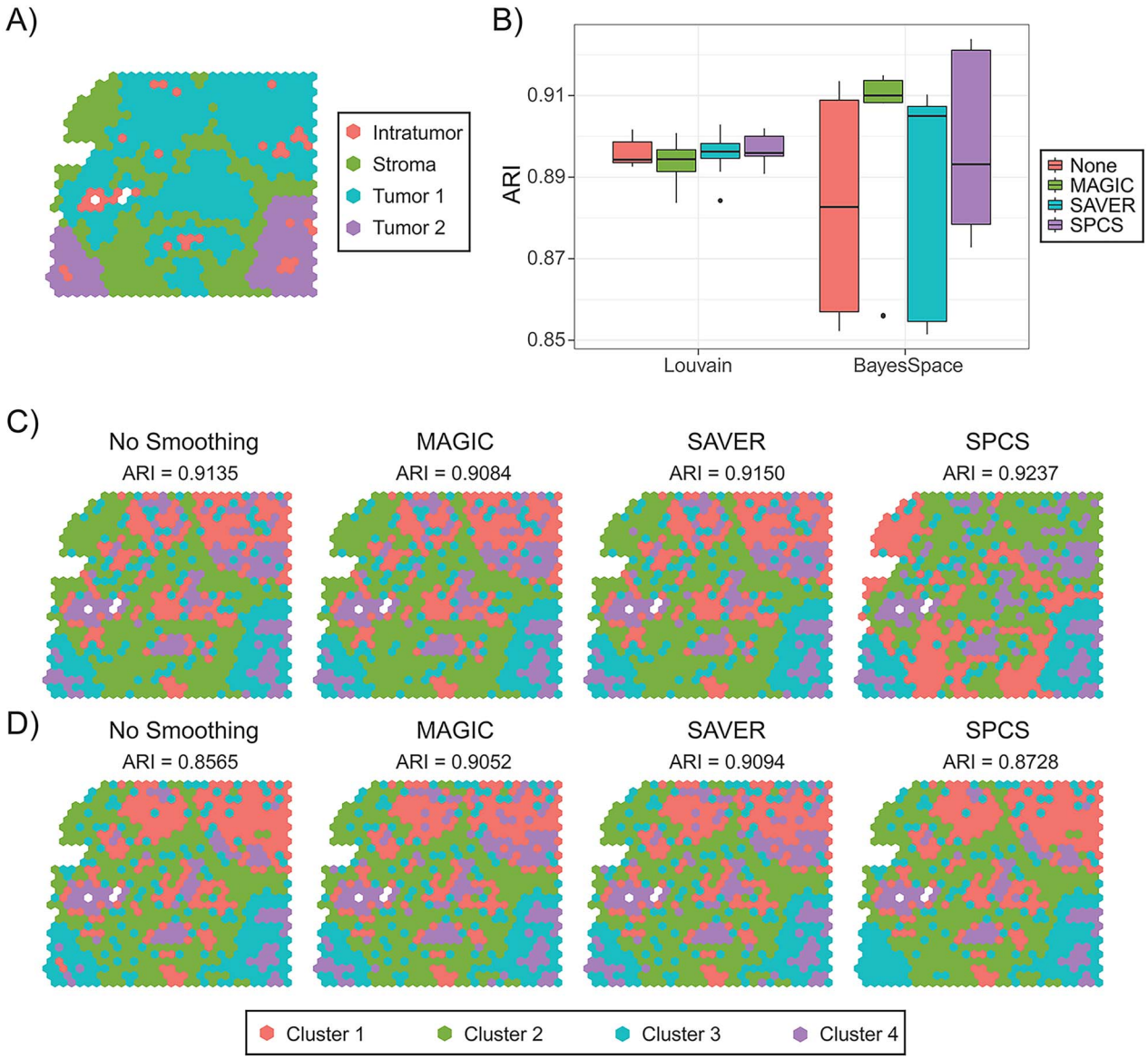

Figure 5. Influence of smoothing on clustering accuracy in HGSOC simulating dataset. (**A**) Ground-truth labels of a simulated ST slide. (**B**) Box plot of ARI score on Louvain and BayesSpace clustered 10 smoothed simulated ST slides. (**C** and **D**) BayesSpace clustering results of two of the simulated ST slides are illustrated. Clusters are ordered by their size.

do not seem related to PDAC, except for slide PDACF that contained cell adhesion and ECM terms for the no smoothing and MAGIC.

The enriched GO terms for the down-regulated DEGs (Figure 6**D**) tell a different story. Without smoothing, GO terms related to digestion as well as infection and autoimmune-related pathways and complement cascade components (C3, C5) were identified. GO terms related to digestion were seen with SAVER as well. After performing smoothing, we found many more GO terms related to the ECM, more complement binding, as well as apoptosis regulation, but the infection and autoimmune-related pathways were absent. Applying MAGIC and SPCS also helped to find other complement cascade components such as C1q complex binding. The other slides shown are in Supplementary Figure S3, available online at https://academic.oup.com/bib, and have similar results to the terms described above.

Figure 6**E** contains terms enriched in the SPCS smoothed data from up-regulated DEGs that have previously been reported in PDAC [54–57]. These terms, such as cell adhesion, cadherin binding, PI3K-Akt signaling pathway and focal adhesion, are all significantly enriched in DEGs from SPCS-smoothed data (albeit not among the top 10 enriched terms), but were absent from unsmoothed data, reflecting the enhancement of biological interpretability by performing our novel SPCS smoothing method.

The results of the biological analysis of the DLPFC slides are shown in Supplementary Figure S4 available online at https://academic.oup.com/bib. The significant DEGs are shown in Supplementary Figure S4**A**, available online at https://academic.oup.com/bib, and show

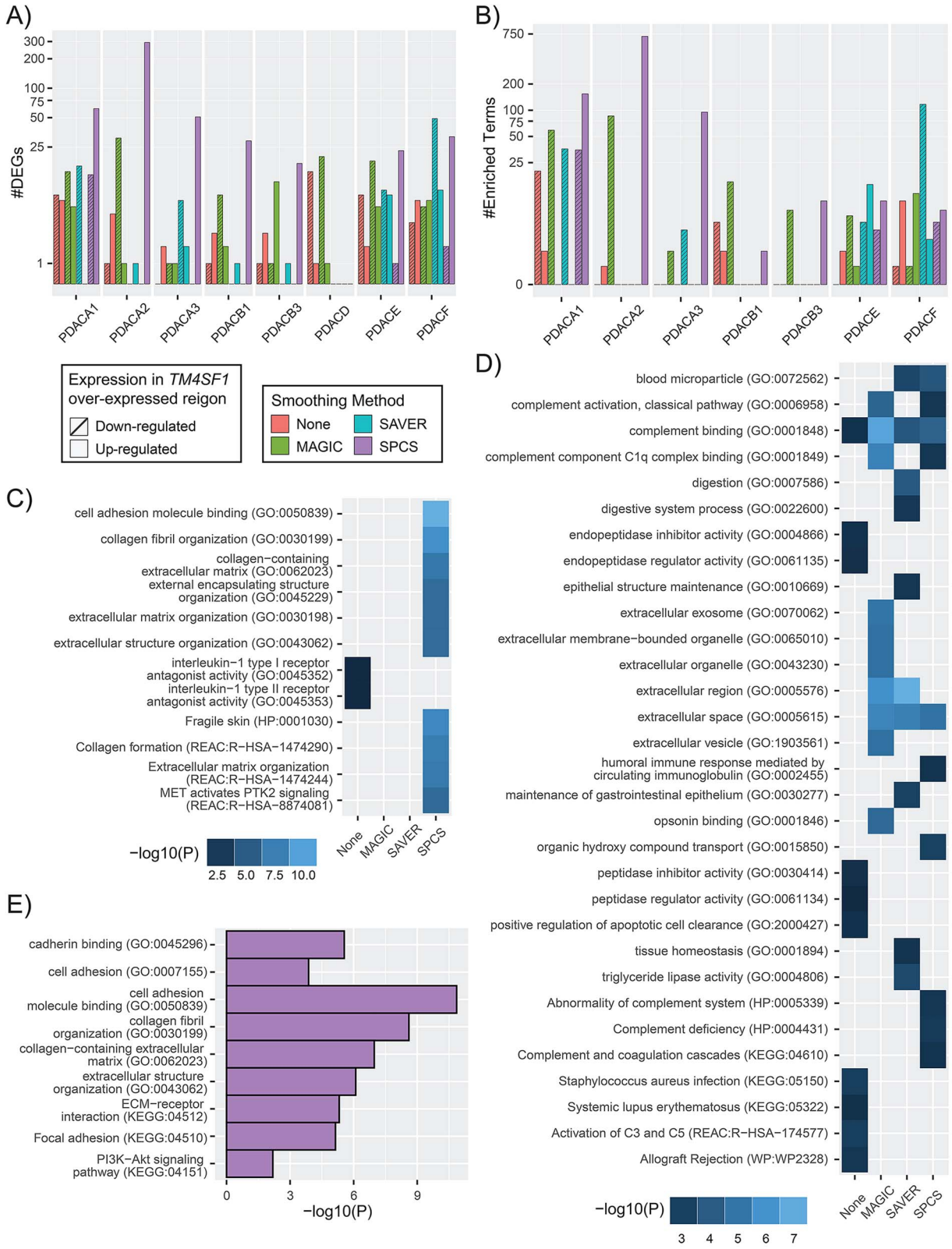

Figure 6. Biological analysis of unsmoothed and different methods (MAGIC, SAVER and SPCS) smoothed PDAC slides. (**A**) Number of DEGs identified in each slide and (**B**) the number of GO terms found from GOEA in each slide shown on a log10 scale. DEGs that are down-regulated in *TM4SF1* over-expressed region and their corresponding GOEA terms are marked as slash-filled texture. The 10 most significantly enriched terms from each smoothing method found in the (**C**) up-regulated DEGs and (**D**) down-regulated DEGs groups in *TM4SF1* over-expressed region for PDACA1 slide are shown, shaded by −log10(*P*-value). (**E**) GO terms enriched in SPCS smoothed slides from up-regulated DEGs that have previously been found in PDAC.

a similar number of DEGs for each algorithm. When comparing the DEGs with those reported by Zeng *et al.* [53], a similar number of genes were shared for each smoothing method, as shown in Supplementary Figure S4**B** available online at https://academic.oup.com/bib. These genes include *MOBP* (an oligodendroctye and white matter marker), *PCP4* (associated with L5 and L6) [14, 53] and *GFAP* (associated with astrocytes and many neurological disorders) [58]. In addition, when comparing with other smoothing methods, SPCS helped to discover five DEGs [i.e. Cell Adhesion Molecule 4 (*CADM4*), *ELOVL5*, *AGR2*, *LGALS3BP* and scavenger receptor class B Member 2 (*SCARB2*)], which the authors have not found to be reported in the DLPFC. These genes were found in the white matter layers of the brain. The expression of *CADM4* and *SCARB2* in sample 151673 is shown in Supplementary Figure S4**C** and **D** available online at https://academic.oup.com/bib, respectively.

Discussion

Importance of smoothing on ST data

ST data are based on highly multiplexed sequence analysis where barcodes are used to split the sequenced reads into their respective tissue locations. However, this type of sequencing suffers from high noise and dropout events. To keep enough genes to perform biological analysis, we set a relaxed filtering threshold (<70% zero-expressed spots) to filter out non-expressed genes in preprocessing steps. Even with this rather relaxed threshold, <10% of genes in slides of both PDAC and DLPFC datasets were left, which indicated that dropout events are highly frequent in ST datasets. By visualizing the expression distribution of two marker genes, *PRSS1* and *TM4SF1* in Figure 3, the gene-level dropout events can be easily seen, as indicated by the black-colored spots on the slides. *PRSS1* and *TM4SF1* dropout events frequently occurred in interstitium regions of the slide, leading to a failure cluster this area on unsmoothed ST data. In addition, there are also entire spots missing in multiple ST slides, which can negatively influence spatial clustering. By performing smoothing methods, the missing and noisy expression values were controlled to a certain extent so that the partition of regions and downstream analysis are greatly improved. Therefore, smoothing is an important and necessary step for analyzing ST data.

SPCS improves data quality

Compared with unsmoothed data, smoothing improves data quality. We have demonstrated that various smoothing algorithms increase the separability and partition accuracy of ST spots. Moreover, by performing internal and external evaluations, we confirm that SPCS-smoothed data show better quality as compared with the two existing one-factor smoothing methods, MAGIC and SAVER. In the internal evaluation, from the results shown in Figure 2, SPCS smoothing method produces greater silhouette scores than MAGIC and SAVER, which means the ST data smoothed by SPCS have better separability. In addition, a more similar silhouette score of SPCS smoothed and original unsmoothed data indicates that SPCS can better preserve the original data distribution, which helps to keep accurate biological analysis results while improving spatial clustering accuracy.

Our external evaluation verifies the partition accuracy of smoothed data. One main objective of slide partition is to identify different histopathological regions in the slide. Hence, we measured the degree of overlap between unsupervised clusters on smoothed data and histopathological partitions using ARI. Since the spots in the same histopathological region are usually connected, incorporating spatial knowledge is expected to boost partition accuracy. Moreover, spatial knowledge can also help to detect and pad missing spots. Indeed, as expected, results in Figures 3–5 reveal that SPCS method generates a higher ARI score than existing one-factor methods, which means a more accurate histopathological partition can be acquired by performing the two-factor SPCS method. In addition, it is also clear that SPCS can be used before various clustering methods to improve clustering accuracy. From the marker gene analysis, SPCS method recovered the dropout events and enhanced the expressions of marker genes in the corresponding regions. This evidence proves that SPCS can improve the accuracy of the ST spot partitions.

SPCS enhances biological interpretability

Our SPCS method identified many more DEGs than the other smoothing methods tested for most of the ST slides. *IL1RN*, *KRT7*, *LAMB3*, *LAMC2*, *NOTCH3 and S100A16* were identified as up-regulated DEGs in the PDACA1 slide after SPCS processing, and each of these genes has been associated with poor survival in PDAC [55, 59]. *IL1RN* and *LAMB3* were also found using unsmoothed data and *LAMB3* was detected with MAGIC smoothed data, but the other genes were uniquely detected by SPCS. A higher number of DEGs associated with GO terms after SPCS processing than unsmoothed, SAVER and MAGIC processing. Specifically, only slides PDACE and PDACF contained up-regulated DEGs and GO terms for SAVER and MAGIC, although PDACD was the only slide that had DEGs and no GO terms for the down-regulated genes. The unsmoothed up-regulated slides PDACA2, PDACB3, PDACD and PDACE and down-regulated slides PDACA3, PDACB3 and PDACD also had DEGs but no GO terms. In contrast, every SPCS smoothed slide that had DEGs also had GO terms. The SPCS data demonstrated a *TM4SF1* expression landscape that matched the original histopathological assignment more accurately than the other smoothing methods.

The GO terms reported in PDACA1 using SPCS related more to pancreatic cancer and to the role of *TM4SF1* than the non-smoothed data and the two one-factor smoothed data. *TM4SF1* has been found to be over-expressed in PDAC and has roles in apoptosis, proliferation and cell migration [40]. Previous work

found that collagen 1 (GO:0030199) binds with DDR1, which then interacts with TM4SF1 to activate the focal adhesion kinase (FAK) [40, 60, 61]. This results in the disruption of E-cadherin (GO:0045296), leading to the Wnt signaling pathway and loss of cell–cell adhesion (GO:0050839) [40, 59]. The FAK pathway also increases the expression of N-cadherin (GO:0045296), which results in the migration of cancer cells [40, 59]. TM4SF1 can also activate the AKT pathway (KEGG:04151), leading to anti-apoptotic effects and angiogenesis [40, 62]. All of these previously identified GO terms are important factors for PDAC survival and metastasis. They were also identified using SPCS-smoothed data only, which would have been missed using other smoothing methods or with the unsmoothed ST data.

The terms found through GOEA were more interpretable in the scenario of PDAC and its pathogenesis when applying SPCS smoothing. Many of the SPCS top 10 terms, as shown in Figure 6**C**, and previously reported terms, as shown in Figure 6**E**, are similar to those found in the literature [54–57] and are consistent between slides. Without applying SPCS, some of the top terms found in Figure 6**D** are involved in typical pancreatic activity, such as peptidase regulator activity (GO:0061134), digestion (GO:0007586) and triglyceride lipase activity (GO:0004806), which indicates that important PDAC pathology-related GO terms may be missed when data are not smoothed or not properly smoothed.

Identifying the DEGs in different histological regions and checking their associated GO terms can help evaluate the biological interpretability of a smoothed slide. It is worth noting that the GO (i.e. gene set database) and the enrichment tool can yield different results. The simplest example would be the use of a hypergeometric test to determine enriched gene sets opposed to gene set enrichment analysis, which accounts for the significance of DEGs. The hypergeometric test has a longer history of use and is easily interpretable whereas gene set enrichment analysis is a newer approach. Furthermore, the gene sets themselves differ between databases such as KEGG and GO. There could potentially be more enriched KEGG pathways for one sample and more enriched GO terms for another. For these reasons, we used g:Profiler since it incorporates many gene set databases in the analysis and relies on the well-established hypergeometric testing approach.

Using the cortical layers in the DLPFC slides, we found a similar number of DEGs and shared genes between the smoothing methods. Most of these genes were previously reported. Due to the enhanced contrast, our proposed SPCS method helped to identify five DEGs in the white matter, while the other methods did not. The *CADM4* and *SCARB2*, to the author's knowledge, have not been identified as white matter markers in the DLPFC. Cell adhesion molecules like *CADM4* play an important role in myelination by oligodendrocytes. Higher expression of *CADM4* leads to many short myelin internodes that disrupt the normal myelination process [63]. *SCARB2* is a lysosomal membrane receptor for the glucocerebrosidase enzyme. It has been associated with Parkinson's disease and Lewy Body disease. Glucocerebrosidase degenerates sphingolipid, which is important for brain development. There is some evidence that decreases in sphingolipids can lead to demyelination [64, 65]. While the authors do not intend to present *CADM4* and *SCARB2* as marker genes for white matter, we believe that SPCS can be used to help aid with this task. The enrichment analysis results for DLPFC are not shown but are similar between all smoothing methods. This is likely because the number of GO terms found is correlated with the number of DEGs. It is possible that using a gene marker instead of layers to identify DEGs could produce different results. Given the generally clearly defined boundaries of the brain, using layer data to get differential gene expression seems more appropriate.

Determination of SPCS parameters

There are four parameters in SPCS:τ_s, τ_p, α and β. The parameters τ_s and τ_p are designed to adjust the size of spatial neighborhood and pattern neighborhood, respectively. Including more information while performing smoothing is beneficial for a more robust result, and increasing the size of neighborhood is a good way to achieve that goal. Blindly expanding neighborhoods will incorporate some spots that are not similar to the spot being smoothed; therefore, SPCS uses contribution weighting to reduce this effect, which gives the size of both spatial and pattern neighborhoods limited influence on data separability, as shown in Figure 2**C**. In addition, as shown in Figure 4**B**, spatial neighborhood will also influence clustering sensitivity. A smaller spatial neighborhood (τ_s) for SPCS can help to capture long narrow regions in slides but may cause over clustering in thicker regions with a similar length and width. Therefore, we recommend a modest selection of these two parameters to balance the trade-off. For most cases, $\tau_p \leq 16$ and $\tau_s \leq 4$ are recommended.

The parameters, α and β, are designed to balance the original expression and corrections from spatial and pattern neighbors, which has a significant effect on smoothing quality. Due to the pattern similarity between the object spot and its pattern neighbors, corrections from pattern neighbors tend to enhance the original data distribution features, which is shown by an increased or stabilized on silhouette score. In contrast, corrections from spatial neighbors make the object spot expression consistent with its spatial neighbors. This is beneficial for spatial clustering but may change the original data distribution, leading to a worse silhouette score. To keep the accuracy of spatial clustering and biological analysis simultaneously, it is important to balance the intensity of correction with the underlying expression signatures. Results in Figure 2**B** indicate that the influence of smoothing strength (α) on data separability is heavily reliant on the proportion of spatial and pattern

correction (β). Hence, we recommend setting α between 0.2 and 0.8 and β <0.6 in most cases, and values should be selected carefully according to data distribution. In addition, as indicated in Figure 5, when combined with spatial clustering methods like BayesSpace, smaller α and β are recommended to avoid erroneous merging of small clusters.

Conclusion

In response to expression noise and dropout events in barcoding-based sequencing technologies, smoothing has become an essential data processing step before performing the downstream analysis on ST data. In this paper, we proposed a novel two-factor ST data smoothing method, SPCS, which can take full advantage of both the expression patterns and the spatial patterns contained in ST data. Compared with traditional one-factor smoothing methods, SPCS improved separability, partition accuracy and biological interpretability of ST experiments. SPCS can effectively improve ST data quality for accurate and meaningful downstream analyses. SPCS is broadly applicable to any barcoding-based ST technology.

Key Points

- Due to the common issue of noise and dropout events in ST data, smoothing has become a necessary step before downstream analysis on ST data.
- SPCS is a novel kNN-based two-factor smoothing method which can fully utilize both expression pattern and spatial knowledge in ST data.
- Compared with traditional expression pattern knowledge-based one-factor smoothing methods, SPCS can provide better separability, partition accuracy and biological interpretability.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgements

We specially thank Dr Edward Zhao and Dr Raphael Gottardo in Fred Hutchinson Cancer Research Center for providing information and for helping in the simulation experiment design. We would also like to acknowledge the support of the Indiana Biosciences Research Institute and the Indiana University Melvin and Bren Simon Comprehensive Cancer Center.

Funding

ACS-IRG Grant Mechanism (Grant No. 19-144-34 to T.S.J.); National Natural Science Foundation of China (Grant No. 41876100 to X.Y.); State Key Program of National Natural Science Foundation of China (Grant No. 61633004 to X.Y.); Indiana University Precision Health Initiative (to K.H. and J.Z.).

Data availability

All of the data used for the analyses in the manuscript are freely available from their original publications.

References

1. Heindl A, Sestak I, Naidoo K, *et al.* Relevance of spatial heterogeneity of immune infiltration for predicting risk of recurrence after endocrine therapy of ER+ breast cancer. *J Natl Cancer Inst* 2018;**110**.
2. Valkonen M, Ruusuvuori P, Kartasalo K, *et al.* Analysis of spatial heterogeneity in normal epithelium and preneoplastic alterations in mouse prostate tumor models. *Sci Rep* 2017;**7**:44831.
3. Yeong J, Tan T, Chow ZL, *et al.* Multiplex immunohistochemistry/immunofluorescence (mIHC/IF) for PD-L1 testing in triple-negative breast cancer: a translational assay compared with conventional IHC. *J Clin Pathol* 2020;**73**:557–62.
4. Gillett C, Fantl V, Smith R, *et al.* Amplification and overexpression of cyclin D1 in breast cancer detected by immunohistochemical staining. *Cancer Res* 1994;**54**:1812–7.
5. Gorsch SM, Memoli VA, Stukel TA, *et al.* Immunohistochemical staining for transforming growth factor beta 1 associates with disease progression in human breast cancer. *Cancer Res* 1992;**52**:6949–52.
6. Liang H, Li H, Xie Z, *et al.* Quantitative multiplex immunofluorescence analysis identifies infiltrating PD1(+) CD8(+) and CD8(+) T cells as predictive of response to neoadjuvant chemotherapy in breast cancer. *Thorac Cancer* 2020;**11**:2941–54.
7. Sun Z, Nyberg R, Wu Y, *et al.* Developing an enhanced 7-color multiplex IHC protocol to dissect immune infiltration in human cancers. *PLoS One* 2021;**16**:e0247238.
8. Burgess DJ. Spatial transcriptomics coming of age. *Nat Rev Genet* 2019;**20**:317.
9. Marx V. Method of the year: spatially resolved transcriptomics. *Nat Methods* 2021;**18**:9–14.
10. Moncada R, Barkley D, Wagner F, *et al.* Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat Biotechnol* 2020;**38**:333–42.
11. He B, Bergenstråhle L, Stenbeck L, *et al.* Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat Biomed Eng* 2020;**4**:827–34.
12. Berglund E, Maaskola J, Schultz N, *et al.* Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat Commun* 2018;**9**:2419.
13. Chen WT, Lu A, Craessaerts K, *et al.* Spatial transcriptomics and in situ sequencing to study Alzheimer's disease. *Cell* 2020;**182**:976–91.e919.
14. Maynard KR, Collado-Torres L, Weber LM, *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat Neurosci* 2021;**24**:425–36.
15. Ortiz C, Navarro JF, Jurek A, *et al.* Molecular atlas of the adult mouse brain. *Sci Adv* 2020;**6**:eabb3446.
16. Vickovic S, Eraslan G, Salmén F, *et al.* High-definition spatial transcriptomics for in situ tissue profiling. *Nat Methods* 2019;**16**:987–90.
17. Berglund E, Saarenpää S, Jemt A, *et al.* Automation of spatial transcriptomics library preparation to enable rapid and

robust insights into spatial organization of tissues. *BMC Genomics* 2020;**21**:298.
18. Dries R, Zhu Q, Dong R, *et al.* Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol* 2021;**22**:78.
19. Haque A, Engel J, Teichmann SA, *et al.* A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med* 2017;**9**:75.
20. Linnarsson S, Teichmann SA. Single-cell genomics: coming of age. *Genome Biol* 2016;**17**:97.
21. Ståhl PL, Salmén F, Vickovic S, *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 2016;**353**:78–82.
22. Ding J, Adiconis X, Simmons SK, *et al.* Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat Biotechnol* 2020;**38**:737–46.
23. Picelli S, Björklund ÅK, Faridani OR, *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* 2013;**10**:1096–8.
24. Hashimshony T, Senderovich N, Avital G, *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol* 2016;**17**:77.
25. Huang M, Wang J, Torre E, *et al.* SAVER: gene expression recovery for single-cell RNA sequencing. *Nat Methods* 2018;**15**:539–42.
26. van Dijk D, Sharma R, Nainys J, *et al.* Recovering gene interactions from single-cell data using data diffusion. *Cell* 2018;**174**:716–29.e727.
27. Izar B, Tirosh I, Stover EH, *et al.* A single-cell landscape of high-grade serous ovarian cancer. *Nat Med* 2020;**26**:1271–9.
28. Zhao E, Stone MR, Ren X, *et al.* Spatial transcriptomics at subspot resolution with BayesSpace. *Nat Biotechnol* 2021;**39**:1375–84.
29. Jiang DX, Tang C, Zhang AD. Cluster analysis for gene expression data: a survey. *IEEE Trans Knowl Data Eng* 2004;**16**:1370–86.
30. Jaskowiak PA, Campello RJ, Costa IG. On the selection of appropriate distances for gene expression data clustering. *BMC Bioinformatics* 2014;**15**:S2.
31. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemom Intel Lab Syst* 1987;**2**:37–52.
32. Liu Y, Ye X, Zhan X, *et al.* TPQCI: a topology potential-based method to quantify functional influence of copy number variations. *Methods* 2021;**192**:46–56.
33. Wang Z, Chen Z, Zhao Y, *et al.* A community detection algorithm based on topology potential and spectral clustering. *Sci World J* 2014;**2014**:329325.
34. Struyf A, Hubert M, Rousseeuw P. Clustering in an object-oriented environment. *J Stat Softw* 1997;**1**:1–30.
35. Waltman L, van Eck NJ. A smart local moving algorithm for large-scale modularity-based community detection. *Eur Phys J B* 2013;**86**:471.
36. Scrucca L, Fop M, Murphy TB, *et al.* mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R J* 2016;**8**:289–317.
37. Hu J, Li X, Coleman K, *et al.* SpaGCN: integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat Methods* 2021;**18**:1342–51.
38. Shahapure KR, Nicholas C. Cluster quality analysis using silhouette score. In: *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, Sydney, NSW, Australia. Manhattan, NY, US: IEEE, 2020, 747–8.
39. Santos JM, Embrechts M. On the use of the adjusted Rand index as a metric for evaluating supervised classification. In: *International Conference on Artificial Neural Networks*, Limassol, Cyprus. Berlin, Germany: Springer, 2009, 175–84.
40. Fu F, Yang X, Zheng M, *et al.* Role of transmembrane 4 L six family 1 in the development and progression of cancer. *Front Mol Biosci* 2020;**7**:202.
41. Liu Q, Guo L, Zhang S, *et al.* PRSS1 mutation: a possible pathomechanism of pancreatic carcinogenesis and pancreatic cancer. *Mol Med* 2019;**25**:44.
42. Dal Molin A, Baruzzo G, Di Camillo B. Single-cell RNA-sequencing: assessment of differential expression analysis methods. *Front Genet* 2017;**8**:62.
43. Raudvere U, Kolberg L, Kuzmin I, *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* 2019;**47**:W191–8.
44. Reimand J, Kull M, Peterson H, *et al.* g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* 2007;**35**:W193–200.
45. Consortium GO. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res* 2021;**49**:D325–d334.
46. Jassal B, Matthews L, Viteri G, *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res* 2020;**48**:D498–503.
47. Kanehisa M, Furumichi M, Sato Y, *et al.* KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* 2020;**49**:D545–51.
48. Martens M, Ammar A, Riutta A, *et al.* WikiPathways: connecting communities. *Nucleic Acids Res* 2020;**49**:D613–21.
49. Wingender E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform* 2008;**9**:326–32.
50. Giurgiu M, Reinhard J, Brauner B, *et al.* CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res* 2019;**47**:D559–d563.
51. Colwill K, Gräslund S. A roadmap to generate renewable protein binders to the human proteome. *Nat Methods* 2011;**8**:551–8.
52. Köhler S, Gargano M, Matentzoglu N, *et al.* The human phenotype ontology in 2021. *Nucleic Acids Res* 2021;**49**:D1207–17.
53. Zeng H, Shen Elaine H, Hohmann John G, *et al.* Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell* 2012;**149**:483–96.
54. Zhou J, Hui X, Mao Y, *et al.* Identification of novel genes associated with a poor prognosis in pancreatic ductal adenocarcinoma via a bioinformatics analysis. *Biosci Rep* 2019;**39**:BSR20190625.
55. Atay S. Integrated transcriptome meta-analysis of pancreatic ductal adenocarcinoma and matched adjacent pancreatic tissues. *PeerJ* 2020;**8**:e10141.
56. Liu L, Wang S, Cen C, *et al.* Identification of differentially expressed genes in pancreatic ductal adenocarcinoma and normal pancreatic tissues based on microarray datasets. *Mol Med Rep* 2019;**20**:1901–14.
57. Shang M, Zhang L, Chen X, *et al.* Identification of hub genes and regulators associated with pancreatic ductal adenocarcinoma based on integrated gene expression profile analysis. *Discov Med* 2019;**28**:159–72.
58. Brenner M. Role of GFAP in CNS injuries. *Neurosci Lett* 2014;**565**:7–13.
59. Ye J, Wen J, Ning Y, *et al.* Higher notch expression implies poor survival in pancreatic ductal adenocarcinoma: a systematic review and meta-analysis. *Pancreatology* 2018;**18**:954–61.
60. Weniger M, Honselmann KC, Liss AS. The extracellular matrix and pancreatic cancer: a complex relationship. *Cancers (Basel)* 2018;**10**:316.
61. Yang JC, Zhang Y, He SJ, *et al.* TM4SF1 promotes metastasis of pancreatic cancer via regulating the expression of DDR1. *Sci Rep* 2017;**7**:45895.

62. Zheng B, Ohuchida K, Cui L, *et al.* TM4SF1 as a prognostic marker of pancreatic ductal adenocarcinoma is involved in migration and invasion of cancer cells. *Int J Oncol* 2015;**47**: 490–8.
63. Elazar N, Vainshtein A, Golan N, *et al.* Axoglial adhesion by Cadm4 regulates CNS myelination. *Neuron* 2019;**101**:224–31.e225.
64. Alcalay RN, Levy OA, Wolf P, *et al.* SCARB2 variants and glucocerebrosidase activity in Parkinson's disease. *npj Parkinson's Disease* 2016;**2**:16004.
65. Giussani P, Prinetti A, Tringali C. The role of sphingolipids in myelination and myelin stability and their involvement in childhood and adult demyelinating disorders. *J Neurochem* 2021;**156**: 403–14.