SPCS: A Spatial and Pattern Combined Smoothing Method for Spatial Transcriptomic Expression

Yusong Liu[a], Tongxin Wang[b], Ben Duggan[c], Michael Sharpnack[d], Kun Huang[e, f], Jie Zhang*[g], Xiufen Ye*[a], and Travis S. Johnson*[e,h]

[a] College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin, Heilongjiang 150001, China

[b] Department of Computer Science, Indiana University Bloomington, Bloomington, IN 47408, USA

[c] Department of Medicine, Indiana University School of Medicine, Indianapolis, IN 46202, USA

[d] Department of Pathology, University of California San Francisco, San Francisco, CA 94143, USA

[e] Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN 46202, USA

[f] Regenstrief Institute, Indianapolis, IN 46202, USA

[g] Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN 46202, USA

[h] Indiana Biosciences Research Institute, Indianapolis, IN 46202, USA

* Indicate corresponding authors.

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

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

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

E-mail: jizhan@iu.edu

Declarations of Competing Interests:

None

Abstract

High dimensional, localized 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 suffers from high noise and drop-out 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 ST data. In this study, we present a novel two-factor smoothing technique, Spatial and Pattern Combined Smoothing (SPCS), that employs k-nearest neighbor 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 pre-existing 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

57 # Introduction

58 Mamallian tissue is highly heterogeneous with phenotypes that depend on their spatial
59 distribution [1, 2]. Until recently, studies of tissue heterogeneity have either sacrificed
60 spatial relationships (e.g. scRNAseq) or produced low-dimensional measurements (e.g.
61 IHC) [3-7]. Novel spatial transcriptomics (ST) techniques allow whole transcriptome
62 profiles to be measured while preserving spatial relationships [8, 9]. These techniques
63 have already been profoundly useful in understanding tumor [10-12] and non-tumor tissue
64 [13-16] heterogeneity. However, improvements in ST library preparation [17], sequencing
65 techniques, and bioinformatic analysis pipelines [18] are still necessary and ongoing, in
66 contrast to more establish scRNA-seq standard practice protocols [19, 20].

67 The most widely utilized ST technologies are based on highly multiplexed sequence
68 barcoding, which suffers from expression noise and drop-out events [21, 22]. Barcoding-
69 based scRNAseq data suffers from similar limitations, while plate- and in vitro
70 transcription-based techniques such as Smart-seq2 [23] and CEL-seq2 [24], respectively,
71 provide more representative expression profiles per cell at the cost of fewer cells
72 measured per experiment. As a result, a multitude of techniques have been developed to
73 impute the missing expression values and smooth the noise that come directly from the
74 barcode-based non-spatial scRNA-seq. SAVER [25] and MAGIC [26] use sets of
75 correlated genes and relative cell similarity in transcriptome space to impute the drop-out
76 events and eliminate other types of expression errors via machine learning techniques.
77 These methods are termed “one-factor methods” given that they only incorporate
78 expression values. The smoothened expression values give more accurate
79 representations of the true underlying RNA abundances than the raw read counts. ST data
80 has 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 (SPCS), specifically designed for ST data that 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 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 pre-existing one-factor methods.

Methods

Datasets

The datasets that we use in this study including two real world ST dataset, PDAC [10] and DLPFC [27], and a simulating dataset generated from HGSOC single-cell dataset [28]. For the two real world datasets, PDAC include 10 ST slides sourced from 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 wise 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 sourced from different ST platform, i.e. traditional ST platform and new developed Visium platform, we can explore the influence

104 of smoothing methods more comprehensively. A detailed summary statistics of the data
105 we used is provided in Supplementary Table S1.

106 To better explore the ability of smoothing methods to deal with outlier spots, we designed
107 a simulation experiment based on those used in the BayesSpace study [29]. Simulated
108 ST data is based on HGSOC single-cell dataset and an immunofluorescence stained
109 image of an ovarian cancer (OC) biopsy. In the original single-cell analysis of the HGSOC
110 dataset, all the cells were divided into 15 clusters by the DBSCAN clustering method and
111 annotated [28]. Considering the limited number of cells, we only used some of the slide.
112 Ground-truth cluster labels were derived from single-cell level annotation of tumor and
113 stroma compartments within the image. To make the simulated data reflect biology, we
114 separated the slide into four clusters: intratumor (including dendritic and fibroblast cells),
115 stroma (corresponding to macrophage cells) and two tumor clusters (associated with two
116 different maliganent cell clusters). Detailed information about ground truth clusters is
117 provided in Supplementary Table S2. To test the ability of different smoothing methods to
118 find outlier spots, we randomly mixed 5% of other cell types as perturbation for each spatial
119 cluster. We generated 10 sets of simulating data in the simulation analysis.

120 Spatial and Pattern Combined Smoothing (SPCS) of Spatial 121 Transcriptomic Expression

122 For each spot on an ST slide, there exists not only the gene expression but also their
123 spatial positions. This means we can improve the quality of the expression values within
124 each specific spot using the relative similarity to the other spots based on both expression
125 pattern and spatial location on the ST slide. To achieve this goal, we propose a KNN-
126 based method, Spatial and Pattern Combined Smoothing (SPCS), to perform the
127 smoothing and padding. We display the procedure of SPCS method in Figure 1.

Figure 1 placement

In our method, we obtained the smoothed expression of each spot by integrating the contribution-weighted expression of its pattern and spatial neighbors. Let X_i be a vector of gene expression values for spot i, smoothed expression X_i' can be calculated by:

$$X_i' = (1-\alpha)X_i + \alpha\left(\beta\frac{\sum_{j\in N_s(i)} c_{ji}^s X_j}{\sum_{k\in N_s(i)} c_{ki}^s} + (1-\beta)\frac{\sum_{j\in N_p(i)} c_{ji}^p X_j}{\sum_{k\in N_p(i)} c_{ki}^p}\right) \quad (1)$$

In eq. (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. $N_s(i)$ and $N_p(i)$ are the spatial and pattern neighborhood of the object spot i respectively. The size of neighborhoods can be determined by parameter τ_s and τ_p which also should be specified in advance. c_{ji}^s and c_{ji}^p represent the spatial and pattern contributions of a corresponding neighbor 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 seprability 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 is a type of transcriptomic data that measures gene expression patterns similar to scRNA-seq. Like some scRNA-seq data, where cells can be localized to tissue location of origin, spots which 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 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): S_p is the gene expression pattern space of an ST slide, $i, j, k \in S_p$ are different spots, d_{ij}^p and d_{ik}^p are pattern distance between spots i, j and i, k respectively. $N_p(i)$ is the τ_p pattern neighborhood of spot i if $|N_p(i)| = \tau_p, \forall j \in N_p(i), \forall k \in S_p - (N_p(i) \cup \{i\})$ s.t. $d_{ij}^p < d_{ik}^p$.

For gene expression data, the overall shapes of gene expression patterns are of greater interest than the individual magnitudes of each feature [30]. Hence, we used Pearson correlation distance to measure pattern distance between different spots. Let ρ_{ij} represent 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 [31]:

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

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) [32] to transform the expression of spots into a 10-dimensional

principal component (PC) space before smoothing. These uncorrelated components with the largest variance from our PCA are considered 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 named the group of spots that are spatially near a given spot the "spatial neighborhood" of that spot, which is defined explicitly below.

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

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 of spatial pattern of spots, the shape of spatial neighborhood could be different in different platform. Figure 1(b) illustrated 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 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 Pearson correlation distance described in eq. (2) as the distance metric, whose range is $[0,2]$. If the expression of two spots has a negative correlation, the distance based on eq. (2) will become greater than 1. However, smoothing with negative correlation spots is not performed since they are dissimilar. Therefore, we set the contribution to 0 if pattern distance between the object spot and one of its neighbors is greater than 1. We used an exponential transformation to achieve this goal. For object spot i and its pattern neighbor j, pattern contribution c_{ji}^{p} can be defined as:

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

The exponential function in eq. (3) limits the range of c_{ji}^{p} to [0,1] and ensures 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 $d_{ij}^{p} > \frac{3\sigma}{\sqrt{2}}$, c_{ji}^{p} will quickly decay to 0 [33, 34]. Hence, setting σ to $\frac{\sqrt{2}}{3}$ will remove the effect of negative correlation neighbors and we can simplify the eq. (3) to:

$$c_{ji}^{p} = e^{-4.5\left(d_{ij}^{p}\right)^{2}} \tag{4}$$

For spatial neighbors, we used Manhattan distance, an integer greater than 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.:

$$c_{ji}^{s} = \frac{1}{d_{ij}^{s}} \tag{5}$$

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 values are missing genes, which means the spot itself is located in tissue region, but some specific gene expression is missing. This is the main form of drop-out events and also frequently happens in single-cell data, which most of the smoothing methods can handle. The second type are missing spots, i.e. expression of all genes in the spot is 0. This kind of absent data is 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 need to be padded. As shown in Figure 1(c), for each blank spot in a slide, SPCS will only pad the ones whose pre-determined spatial neighborhood has more than 50% non-blank spots. This criteria ensures that boundary of the tissue will not be erroneously expanded. Since there's 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, its expression X_i can be estimated by:

$$X_i = \frac{\sum_{j \in N_s(i)} c_{ji}^{s} X_j}{\sum_{k \in N_s(i)} c_{ki}^{s}} \tag{6}$$

236 In Eq.(6), $N_s(i)$ is τ_s-order spatial neighborhood of spot i, c_{ji}^s and c_{ki}^s are spatial
237 contribution of spot j and k to spot i. Spot padding is performed after non-missing spot
238 smoothing. To evalulate ARI score for padding spots, we assign ground truth labels for
239 them. The ground truth labels of the padded spots are determined by spatial contribution
240 weighted major voting of their spatial neighbors and manually inspected by a pathology
241 resident.

242 Performance evaluation

243 To evaluate the effectiveness of our proposed SPCS method, we first performed SPCS
244 and other one-factor smoothing methods (SAVER and MAGIC) on PDAC, DLPFC and
245 simulated ST datasets. SAVER and MAGIC are two representative one-factor smoothing
246 methods using different techniques (i.e., statistical model-based and kNN-based). Then
247 we partitioned both smoothed and the original unsmoothed real world slides using K-
248 medoids clustering method [35] and judged how well the clusters were separated after
249 smoothing by internal evaluation of the unsupervised clustering. In addition, we also
250 explored how the parameters in SPCS will influence the smoothing. Next, we performed
251 multiple unsupervised clustering methods, including K-medoids, Louvain [36], mclust [37],
252 BayesSpace [29] and spaGCN [38], on all the smoothed and unsmoothed slides to find
253 out how well the clusters match the histopathological labels from the corresponding image
254 as an external evaluation. Simulating data was also used here to detect outliers for each
255 smoothing method. As a gene filter, we only kept genes with less than 70% zero expressed
256 spots in our analysis. For normalization, we performed logarithmic count per million (CPM)
257 normalization before smoothing using SPCS, SAVER, and MAGIC. The distance metric
258 used during clustering was Pearson correlation distance as described in eq. (2). For each
259 dataset, number of clusters is pre-determined by ground truth which is provided in

Supplementary Table S1-S2. PCA was performed prior to clustering and the eigenvectors with top 20 eigenvalues were selected to reduce the dimensions of expression matrices and enhance the clustering results.

Internal evaluation

For the internal evaluation, silhouette score was used as the evaluation indicator [39]. 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:

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

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

In this work, Pearson correlation based measurment in Eq. (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 Eq. (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 level. At the clustering level, the imputed ST data is clustered using multiple clustering algorithms. Technical details of the clustering methods is shown in Supplementary Table S4. 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 Adjusted Rand Index (ARI) was used to evaluate the similarity between the clustering results from the imputed ST data and the histopathological labels [40]. For clusters from the imputed ST data $\{\boldsymbol{I}_1, \boldsymbol{I}_2, \boldsymbol{I}_3, \boldsymbol{I}_4\}$ and the histopathological categories of ST spots $\{\boldsymbol{H}_1, \boldsymbol{H}_2, \boldsymbol{H}_3, \boldsymbol{H}_4\}$, we denoted n_{ij} as the number of ST spots that are both in $\boldsymbol{I}_i$ and $\boldsymbol{H}_j$. The ARI is defined as:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{n_i}{2} + \sum_j \binom{n_j}{2}\right] - \frac{\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}}{\binom{n}{2}}} \tag{8}$$

where n_i is the number of ST spots in $\boldsymbol{I}_i$ and n_j is the number of ST spots in $\boldsymbol{H}_j$. A higher ARI value indicates that the imputed ST clusters and the histopathological labels are more similar. In this analysis, padded spots 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

303 *TM4SF1*, were used to compare their expression distribution spatially. Both genes are
304 protein-coding genes. *PRSS1* encodes a trypsinogen, which is often highly expressed in
305 normal pancreatic tissues while *TM4SF1* is a common proto-oncogene and highly
306 expressed in pancreatic cancer among other malignancies [41, 42]. High expression of
307 *TM4SF1* in the cancerous regions of the PDAC dataset has been detected in previous
308 research [10]. For the DLPFC dataset, three other marker genes, *MOBP*, *PCP4* and
309 *SNAP25*, were used in the analysis. Previous research has reported that these genes can
310 delineate different cortical layers [27].The expression distribution of these representative
311 genes can stratify the histopathological regions and can be used to measure the accuracy
312 of that partition.

313 Biological analysis

314 To examine how our algorithm aids in informing the biology of a ST sample, we identified
315 differentially expressed genes (DEGs) and performed gene ontology enrichment analysis
316 (GOEA). DEGs were identified in the PDAC slides by comparing two groups of spots with
317 *TM4SF1*-high (neoplastic tissue) and low expression (non-neoplastic tissue). To define
318 the two groups, we first linearly transformed the expression values of *TM4SF1* between
319 [0,1] by dividing each value by the maximum expression value. The spots on each slide
320 were split into two groups with one having a transformed expression value less than 0.7
321 (*TM4SF1* under-expressed) and the other greater (*TM4SF1* over-expressed). The cutoff
322 0.7 was chosen as it can reflect the boundary of the cancer region accurately.

323 A Kruskal-Wallis test was performed between the over- and under-expressed groups and
324 the p-values were adjusted using the Benjamini-Hochberg method to account for the
325 multiple comparisons [43]. Only genes with an adjusted p-value less than 0.05 were
326 included. In addition, we used logarithmic foldchange (logFC) to determine the up-

327 regulated and down-regulated events for genes. Genes with a logFC greater than 1 were
328 considered up-regulated, and down-regulated if the logFC were less than -1.

329 GOEA was performed on the DEGs using *g:Profiler* [44]. The significance of enriched
330 terms were tested by cumulative hypergeometric test and p-values were corrected by
331 *g:SCS* method [45]. Only terms with an adjusted p-value below 0.05 were reported. All
332 data sources offered by *g:Profiler* were used including Gene Ontology (GO) [46],
333 Reactome [47], KEGG [48], WikiPathways [49], Transfac [50], CORUM [51], Human
334 Protein Atlas [52], and the Human Phenotype Ontology [53]. Heatmaps for each sample
335 were then generated to compare the terms found by each smoothing algorithm.

336 The DLPFC slides were processed slightly differently than the PDAC slides. To find the
337 differentially expressed genes, one layer of the cortex was compared with the other six
338 layers. For example, the expression values for all the sports in Layer 1 (L1) were compared
339 to the expression values for the spots in layers L2, L3, L4, L5, L6, and WM (white matter).
340 A Kruskal-Wallis test was performed between one layer of the cortex and all other layers,
341 p-values corrected using the Benjamini-Hochberg method, and logFC calculated. This
342 resulted in 7 sets of comparisons. The same cuttoff of p-value less than 0.05 and logFC
343 greater than one were used. The significant DEGs were then compared to a set of DLPFC
344 layer marker genes published by Zeng et al. [54]. Enrichment analysis was performed
345 using *g:Profiler* in the same way as the PDAC slides were processed.

346 # Results

347 In this study, we applied our two-factor smoothing method SPCS and two state-of-the-art
348 one-factor smoothing methods (MAGIC and SAVER) to smooth ST slides. To compare
349 the performance of different smoothing methods, we evaluated the quality of generated

clusters after performing K-Medoids unsupervised clustering on both smoothed and unsmoothed expression. 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 with respect to the marker gene *TM4SF1* expression level and GOEA was performed to reveal the biological meaning from the smoothed data.

Figure 2 placement

Internal Evaluation

The silhouette score indicates 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 seperable. After clustering the spots into corresponding clusters, we calculated the silhouette score of ten PDAC slides and eight DLPFC slides with seven clusters for each smoothing method in Figure 2(a) and Figure S1. In most of the slides from these two datasets, SPCS and MAGIC got 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 charactristics of original data distribution resulting in better data separablity.

374 Figure 2 (b)-(c) illustrated how the parameters influenced the data separability of SPCS
375 smoothed PDACA1 slide. For the two ratio parameters α and β, while $\beta \leq 0.4$, silhouette
376 score were not affected much by the changing of α, but the result became more
377 sensentive to α once β is large. This result indicated that pattern based smoothing can
378 help keep the original data distribution. In addition, the results reveal that the data
379 distribution of spatial neighbors is different from the pattern neighbors. For the
380 neighborhood size parameters, the results show that τ_s and τ_p, have no significant
381 influence on data separability. However, while $\tau_s = 2$ and $\tau_p = 4$, silhouette score is
382 significantly lower than other parameter combinations, which indicates neighborhood size
383 parameters also should be well tuned according to the dataset to avoid unexpected results.

384 External Evaluation

385 In the external evaluation, different smoothing methods were evaluated based on the
386 consistency of unsupervised clusters with histopathological labels. The result of PDACA1
387 alone was shown in Figure 3 since the histopathological labels were not available for other
388 PDAC slides. This slide can be well clustered by K-Medoids clustering even without
389 smoothing. When examining the results in detail, Figure 3(b) shows that most ST spots in
390 the cancer cells and desmoplasia region were separated from other clusters. Most ST
391 spots in the duct epithelium region were also well spearated, with slight mixing of the
392 interstitium and the normal pancreatic tissue regions. It is worth mentioning that by
393 including spatial position information, SPCS can pad missing spots, which other one-factor
394 smoothing method cannot do. Figure 3(c, d) show the influence of different smoothing
395 methods on marker genes expression. Compared with other smoothing methods, SPCS
396 generated better marker gene contrast between distinctive histopathological areas due to
397 two reasons. First, marker gene expressions showed fewer drop-outs and better reflected

398 expression patterns across different tissue regions with SAVER and SPCS (MAGIC does
399 not impute the missing expressions). Second and in contrast to SAVER, SPCS imputed
400 the missing values and kept the spatial distribution of non-missing values stable.

401 **Figure 3 placement**

402 Figure 4 illustrates external evaluation results from 12 slides in DLPFC dataset. Figure
403 4(a) shows the influence of smoothing methods on different clustering methods. To make
404 the evaluation comprehensive, we chose three commonly used general clustering
405 methods (i.e. k-Medoids, Louvain and mclust) and two state-of-the-art clustering methods
406 developed specifically for ST slides (i.e BayesSpace and SpaGCN). The results revealed
407 that smoothing achieved a higher ARI score using various clustering methods and SPCS
408 outperform other one-factor smoothing methods for every clustering method. When
409 combining SPCS with Louvain or mclust clustering methods on the DLPFC dataset, we
410 got an ARI score near BayesSpace, indicating that the combination of SPCS with general
411 clustering methods improves the performance of spatial clustering on ST slides. In addition,
412 combining SPCS and BayesSpace got the highest ARI score among all the combinations.
413 Due to the difference of preprocessing steps, the results of SpaGCN clustering were
414 moderately lower than the original publication [38], which indicates that hyperparameter
415 settings of SpaGCN is sensitive to preprocessing steps and we hope to optimize this in
416 the future. Figure 4(b) and Figure S2(a) illustrated clustering ground truth and BayesSpace
417 clustering results combined with different smoothing methods on slide 151675 and 151673.
418 Compared with other one-factor smoothing or without smoothing, SPCS can achieve more
419 clear and accurate boundaries between different regions, which contributes to a higher
420 ARI score. In addition, smaller spatial neighborhood (τ_s) for SPCS can help to capture
421 long narrow regions in the slide but may cause over clustering in thicker regions with
422 similar length and width. Similar to the PDAC dataset, SPCS also helps to obtain a better

423 marker gene contrast between distinctive cortical layers in DLPFC dataset as shown in
424 Figure S2(b-d).

425 **Figure 4 placement**

426 We also performed simulation analysis to test the clustering accuracy for different
427 smoothing methods. Louvain and BayesSpace, the two best general and ST dedicated
428 clustering methods in the DLPFC experiment, were used in this analysis. The clustering
429 ground truth is shown in Figure 5(a). For the 10 simulated slides, Figure 5(b) illustrated
430 distribution of ARI score of each combination of smoothing and clustering method. In
431 general, since the original single cell dataset was well clustered, every combination of
432 smoothing and clustering methods got high ARI scores in this experiment. With Louvain
433 clustering, there was only a small difference in ARI scores on different slides. For different
434 smoothing methods, SPCS gets a slightly higher average ARI score which indicates that
435 spatial information helps general clustering methods to cluster ST slides. However,
436 Louvain seperates the spots into five clusters instead of four, which means Louvain tends
437 to cluster the spots according to cell types. For the BayesSpace experiments, the ARI
438 scores were lower than Louvain on average and greater in variance. To better review the
439 higher variance of BayesSpace, we illustrated BayesSpace clustering results of two
440 simulating slides in Figure 5(c) and (d). Obviously, BayesSpace tends to merge small
441 clusters into nearby larger ones which leads to oversmoothing in the simulated dataset,
442 and smoothing with SAVER and SPCS will aggravate this problem. Overall all methods
443 performed well on the simulated data (ARI > 0.85) but it is worth noting that SPCS has
444 higher potential ARI as evaluated by the 75th quantile. This is surprising considering the
445 other smoothing methods are specifically designed for scRNA-seq data from which these
446 simulated ST slides are generated.

447 **Figure 5 placement**

448 ## Biological Analysis

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

457 The top ten most significant GO terms found in slide PDACA1 for each algorithm are
458 shown in Figure 6(c) for the up-regulated DEGs and Figure 6(d) for the down-regulated
459 DEGs. More than ten terms are shown in the heatmap because most GO terms were not
460 shared between smoothing methods. Without smoothing the slides, the up-regulated
461 DEGs identified are related to interleukin-1. Instead, the terms found by SPCS smoothing
462 are related primarily to cell adhesion, extracellular matrix (ECM) organization and MET
463 activates PTK2 signaling. The GO terms from all smoothed and unsmoothed up-regulated
464 slides can be seen in Figure S3. *g:Profiler* could not find any enriched terms for slides
465 PDACB2, PDACD, and PDACG. Up-regulated enriched terms for all SPCS slides except
466 PDACB1 appear similar to PDACA1. The results for the other methods had a small
467 number of terms which do not seem related to PDAC, except for slide PDACF which
468 contained cell adhesion and ECM terms for the no smoothing and MAGIC.

469 **Figure 6 placement**

470 The enriched GO terms for the down-regulated DEGs (Figure 6(d)) tell a different story.
471 Without smoothing, GO terms related to digestion as well as infection and autoimmune
472 related pathways and complement cascade components (C3, C5) were identified. GO
473 terms related to digestion were seen with SAVER as well. After performing smoothing, we
474 found many more GO terms related to the ECM, more complement binding, as well as
475 apoptosis regulation, but the infection and autoimmune-related pathways were absent.
476 Applying MAGIC and SPCS also helped to find other complement cascade components
477 such as C1q complex binding. The other slides showed are in Figure S3 and have similar
478 results to the terms described above.

479 Figure 6(e) contains terms enriched in the SPCS smoothed data from up-regulated DEGs
480 that have previously been reported in PDAC [55-58]. These terms, such as cell adhesion,
481 cadherin binding, PI3K-Akt signaling pathway and focal adhesion, are all significantly
482 enriched in DEGs from SPCS-smoothed data (albeit not among the top-ten enriched
483 terms), but were absent from unsmoothed data, reflecting the enhancement of biological
484 interpretability by performing our novel SPCS smoothing method.

485 The results of the biological analysis of the DLPFC slides are shown in Figure S4. The
486 significant DEGs are shown in Figure S4(a) and show a similar number of DEGs for each
487 algorithm. When comparing the DEGs to those reported by Zeng et al. [54], a similar
488 number of genes were shared for each smoothing method as shown in Figure S4(b).
489 These genes include *MOBP* (an oligodendroctye and white matter marker), *PCP4*
490 (associated with L5 and L6) [27, 54], and *GFAP* (associated with astrocytes and many
491 neurological disorders) [59]. In addition, comparing with other smoothing methods, SPCS
492 helped to discover five DEGs (i.e., *CADM4*, *ELOVL5*, *AGR2*, *LGALS3BP* and *SCARB2*)
493 which the authors have not found reported in the DLPFC. These genes were found in the
494 white matter layers of the brain. The expression of *CADM4* and *SCARB2* in sample

151673 are shown in Figure S4(c) and Figure S4(d), respectively.

Discussion

Importance of smoothing on ST data

ST data is 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 drop-out events. To keep enough genes to perform biological analysis, we set a relaxed filtering threshold (less than 70% zero expressed spots) to filter out non-expressed genes in preprocessing steps. Even with this rather relaxed threshold, less than 10% of genes in slides of both PDAC and DLPFC datasets were left, which indicated that drop-out 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 drop-out events can be easily seen, as indicated by the black-colored spots on the slides. For both genes, drop-out events frequently occurred in interstitium regions of the slide, leading to a failure cluster this area on unsmoothed ST data. In addition, there's 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

517 accuracy of ST spots. Moreover, by performing internal and external evaluations, we
518 confirm that SPCS-smoothed data show better quality as compared with the two existing
519 one-factor smoothing methods MAGIC and SAVER. For internal evaluation, from the
520 results shown in Figure 2, SPCS smoothing method produces greater silhouette scores
521 than MAGIC and SAVER, which means the ST data smoothed by SPCS has better
522 separability. In addition, a more similar silhouette score of SPCS smoothed and original
523 unsmoothed data indicates that SPCS can better preserve the original data distribution
524 which helps to keep accurate biological analysis results while improving spatial clustering
525 accuracy.

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

SPCS Enhances Biological Interpretability

Our SPCS method identified many more differentially expressed genes than the other smoothing methods tested for most of the ST slides. *IL1RN*, *KRT7*, *LAMB3*, *LAMC2*, *NOTCH3, and S100A16* were identified as upregulated DEGs in the PDACA1 slide after SPCS processing, and each of these genes has been associated with poor survival in PDAC [56, 63]. *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 for SAVER and MAGIC; although PDACD was the only slide which 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 slides 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 the role of *TM4SF1* than the non-smoothed data and the two one-factor smoothed data. *TM4SF1* has been found to be overexpressed in PDAC and has roles in apoptosis, proliferation, and cell migration [41]. Previous work found that collagen 1 (GO:0030199) binds with DDR1 which then interacts with TM4SF1 to activate the focal adhesion kinase (FAK) [41, 60, 61]. This results in disruption of E-cadherin (GO:0045296) leading to the Wnt signaling pathway and loss of cell-cell adhesion (GO:0050839) [41, 60]. The FAK pathway also increases expression of N-cadherin (GO:0045296) which results in migration

564 of cancer cells [41, 60]. TM4SF1 can also activate the *AKT* pathway (KEGG:04151)
565 leading to anti-apoptotic effects and angiogenesis [41, 62]. All of these previously
566 identified GO terms are important factors for PDAC survival and metastasis. They were
567 also identified using SPCS-smoothed data only, which would have been missed using
568 other smoothing methods or with the unsmoothed ST data.

569 The terms found through GOEA were more interpretable in the scenario of PDAC and its
570 pathogenesis when applying SPCS smoothing. Many of the SPCS top ten terms in Figure
571 4(c) and previously reported terms in Figure 4(e) are similar to those found in the literature
572 [55-58] and are consistent between slides. Without applying SPCS, some of the top terms
573 found in Figure 4(d) are involved in typical pancreatic activity such as peptidase regulator
574 activity (GO:0061134), digestion (GO:0007586), and triglyceride lipase activity
575 (GO:0004806) which indicates that important PDAC-pathology related GO terms may be
576 missed when data is not smoothed or not properly smoothed.

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

589 Using the cortical layers in the DLPFC slides, we found a similar number of DEGs and
590 shared genes between the smoothing methods. Most of these genes were previously
591 reported. Due to the enhanced contrast, our proposed SPCS method helped to identify
592 five DEGs in the white matter while the other methods did not. The Cell Adhesion Molecule
593 4 (*CADM4*) and scavenger receptor class B Member 2 (*SCARB2*), to the author's
594 knowledge, have not been identified as white matter markers in the DLPFC. Cell adhesion
595 molecules (CAMs) like *CADM4* play an important role in myelination by oligodendrocytes.
596 Higher expression of *CADM4* leads to many short myelin internodes that disrupt the
597 normal myelination process [64]. *SCARB2* is a lysosomal membrane receptor for the
598 glucocerebrosidase enzyme. It has been associated with Parkinson's disease and Lewy
599 Body Disease. Glucocerebrosidase degenerates sphingolipid, which is important for brain
600 development. There is some evidence that decreases sphingolipids can lead to
601 demyelination [65, 66]. While the authors do not intend to present *CADM4* and *SCARB2*
602 as marker genes for white matter, we believe that SPCS can be used to help aid with this
603 task. The enrichment analysis results for DLPFC are not shown but are similar between
604 all smoothing methods. This is likely because the number of GO terms found is correlated
605 with the number of DEGs. It's possible that using a gene markers instead of layers to
606 identify DEGs could produce different results. Given the generally clearly defined
607 boundries of the brain, using layer data to get differential gene expression seems more
608 appropriate.

609 Determination of SPCS parameters

610 There are four parameters in SPCS: τ_s, τ_p, α and β. The parameters τ_s and τ_p are
611 designed to adjust the size of spatial neighborhood and pattern neighborhood respectively.
612 Including more information while performing smoothing is beneficial for a more robust

613 result, and increasing the size of neighborhood is a good way to achieve that goal. Blindly
614 expanding neighborhoods will incorporate some spots that are not similar to the spot being
615 smoothed; therefore, SPCS uses contribution weighting to reduce this effect, which gives
616 the size of both spatial and pattern neighborhoods limited influence on data separability
617 as shown in Figure 2(e). In addition, as shown in Figure 4 (b) and (c), spatial neighborhood
618 will also influence clustering sensitivity. A smaller spatial neighborhood (τ_s) for SPCS can
619 help to capture long narrow regions in slides but may cause over clustering in thicker
620 regions with similar length and width. Therefore, we recommend a modest selection of
621 these two parameters to balance the tradeoff. For most cases, $\tau_p \leq 16$ and $\tau_s \leq 4$ is
622 recommended.

623 The parameters, α and β, are designed to balance the original expression and corrections
624 from spatial and pattern neighbors, which has a significant effect on smoothing quality.
625 Due to the pattern similarity between the object spot and its pattern neighbors, corrections
626 from pattern neighbors tend to enhance the original data distribution features shown by
627 an increased or stabilized on silhouette score. In contrast, corrections from spatial
628 neighbors make the object spot expression consistent with its spatial neighbors. This is
629 beneficial for spatial clustering but may change the original data distribution leading to a
630 worse silhouette score. To keep the accuracy of spatial clustering and biological analysis
631 simultaneously, it is important to balance the intensity of correction with the underlying
632 expression signatures. Results in Figure 2(d) indicate that the influence of smoothing
633 strength (α) on data separability is heavily reliant on the proportion of spatial and pattern
634 correction (β). Hence, we recommend setting α between 0.2 to 0.8 and β less than 0.6 in
635 most cases, and values should be selected carefully according to data distribution. In
636 addition, as indicating in Figure 5, when combined with spatial clustering methods like
637 BayesSpace, smaller α and β are recommended to avoid erroroneous merging of small

clusters.

Conclusion

In response to expression noise and drop-out 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 on any barcoding-based ST technology.

Key Points

- Due to the common issue of noise and drop-out 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 that 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.

Data Availability

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

original publications.

Funds

This work is partially supported by ACS-IRG Grant Mechanism [Grant No. 19-144-34, to T.S.J.], the National Natural Science Foundation of China [Grant No. 41876100, to X.Y.], the State Key Program of National Natural Science Foundation of China [Grant No. 61633004, to X.Y.], the Indiana University Precision Health Initiative [to K.H. and J.Z.].

Acknowledgement

We special thanks Dr. Edward Zhao and Dr. Raphael Gottardo in Fred Hutchinson Cancer Research Center for providing information and help on simulation experiment design.

Description of Authors

Yusong Liu is a PhD student in the College of Intelligent Systems Science and Engineering, Harbin Engineering University. His research is focus 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 Indiana University School of Medicine. His research interests include bioinformatics, machine learning, and using computing to improve patient care.

679 Dr. Michael Sharpnack is a pathology resident in the Department of Pathology at the
680 Univeristy of California San Francisco. His research is primarily concerned with novel
681 immunotherapies, antigen presentation, and bioinformatics methods development as it
682 applies to lung cancer.

683 Dr. Kun Huang is a Professor and Chair of the Department of Biostatistics and Health Data
684 Science, Indiana University School of Medicine, Indiana University School of Medicine
685 Precision Health Initiative Chair for Genomic Data Science, Director of Data Science and
686 Informatics for the Precision Health Initiative, Associate Director of Data Science for the
687 Indiana University Simon Comprehensive Cancer Center, and an Investigator at the
688 Regenstrief Institute. His main research interests are in medical image analysis, multi-
689 omics, and machine learning.

690 Dr. Jie Zhang is an Assistant Professor of Medical and Molecular Genetics who is also a
691 member of the Center for Computational Biology and Bioinformatics. Her research
692 interests are applied translational bioinformatics and systems biology for cancer and
693 neurological disease.

694 Dr. Xiufen Ye is a professor of College of Intelligent Systems Science and Engineering,
695 Harbin Engineering University, IEEE Senior Member. Her main research interest includes
696 image processing, pattern recognition and artificial intelligent.

697 Dr. Travis S. Johnson is an Assistant Research Professor in the Department of
698 Biostatistics and Health Data Science, Indiana University School of Medicine. His research
699 interests include applied machine learning for use with high dimensional omic data as it
700 applies to cancer and dementia research.

701

Figures

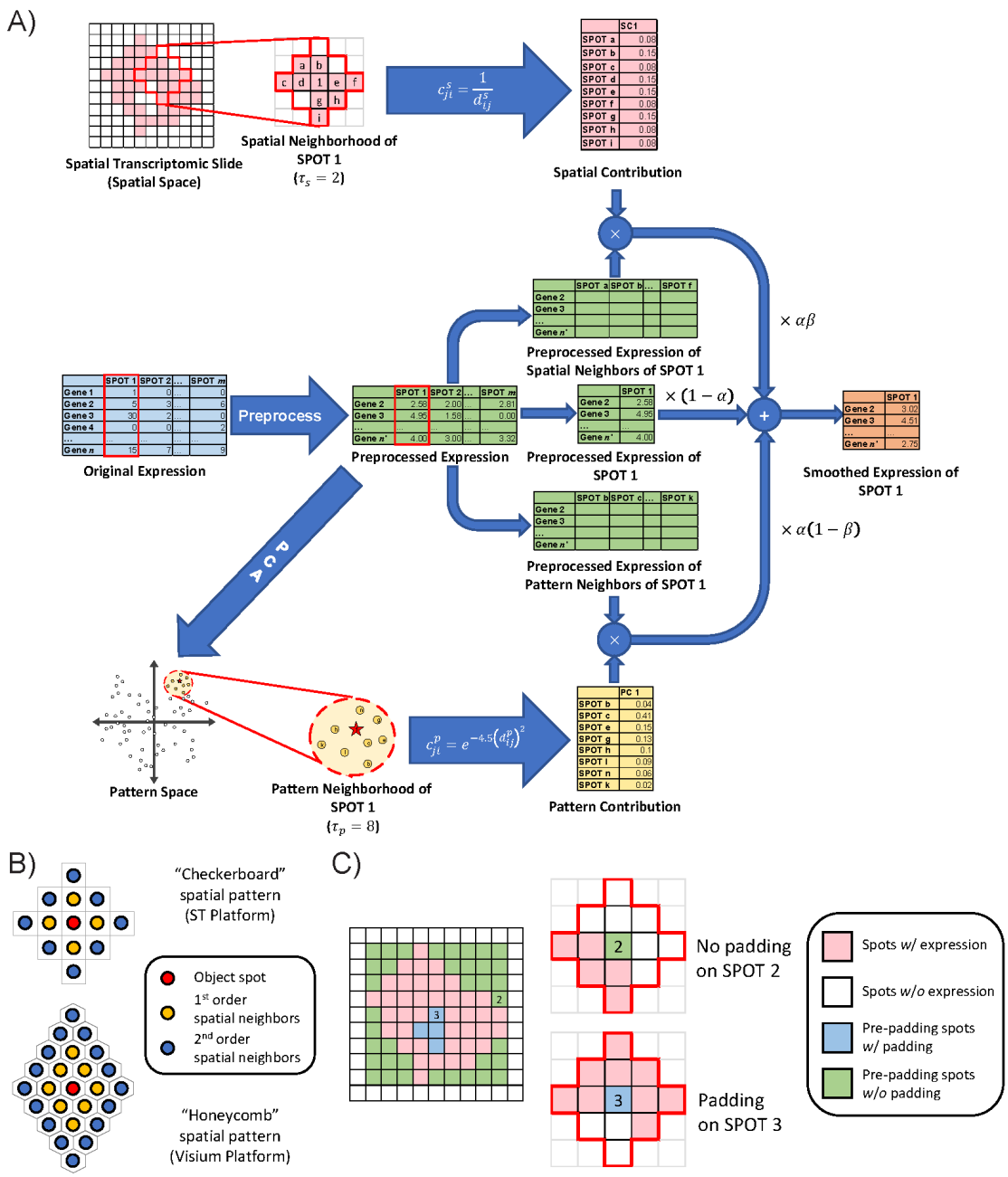
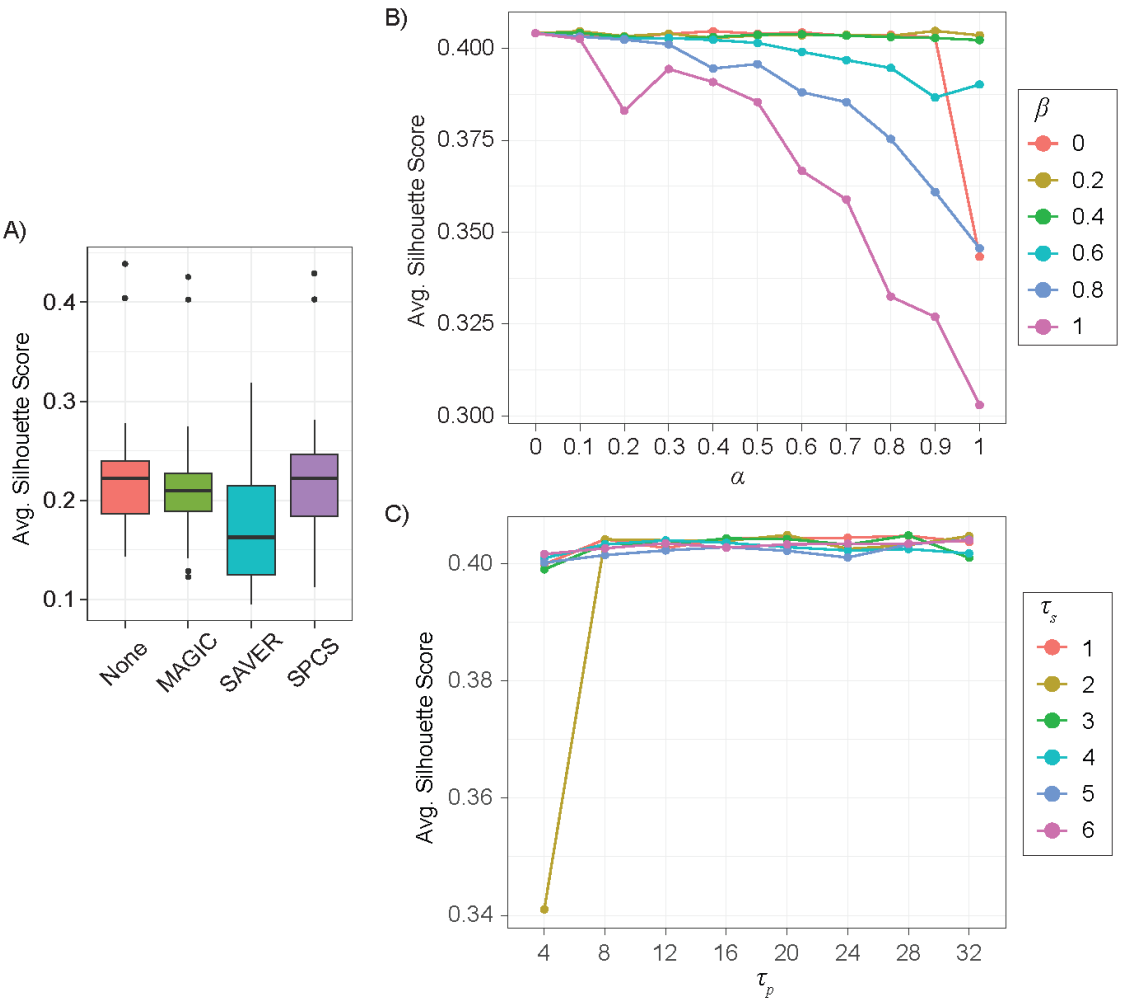


Figure 1 Workflow of our proposed SPCS smoothing method. (a) An example of smoothing. In this sample, SPOT 1 is the object spot that is going to be smoothed. For ST slides, preprocessing steps normalize the expressions and filter out genes with zero

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 reigons and excluded from spatial neighborhood. Pattern neighborhood consists of the top eight spots with most similar expression pattern to SPOT1. Both spatial and pattern contributions are normalized which means total contribution of all neighbors is equal to 1. Smoothing is performed by integrating parameter- and contribution-weighted expression of the object spot itself and both it's 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 same Manhattan distance to object spot belong to 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 more than 50% of unmissing spots inside its spatial neighborhood.



720

721 **Figure 2** Influence of smoothing on data separability and data distribution. (a) Box plot of
722 average silhouette score of ten PDAC samples and eight 7-layer DLPFC samples without
723 smoothing and with smoothing by different methods (MAGIC, SAVER and SPCS). (b)
724 Influence of parameters α and β on average silhouette score for SPCS smoothed
725 PDACA1 slide when $\tau_p = 16, \tau_s = 2$. (c) Influence of parameters τ_p and τ_s on average
726 silhouette score for SPCS smoothed PDACA1 slide when $\alpha = 0.6, \beta = 0.4$.

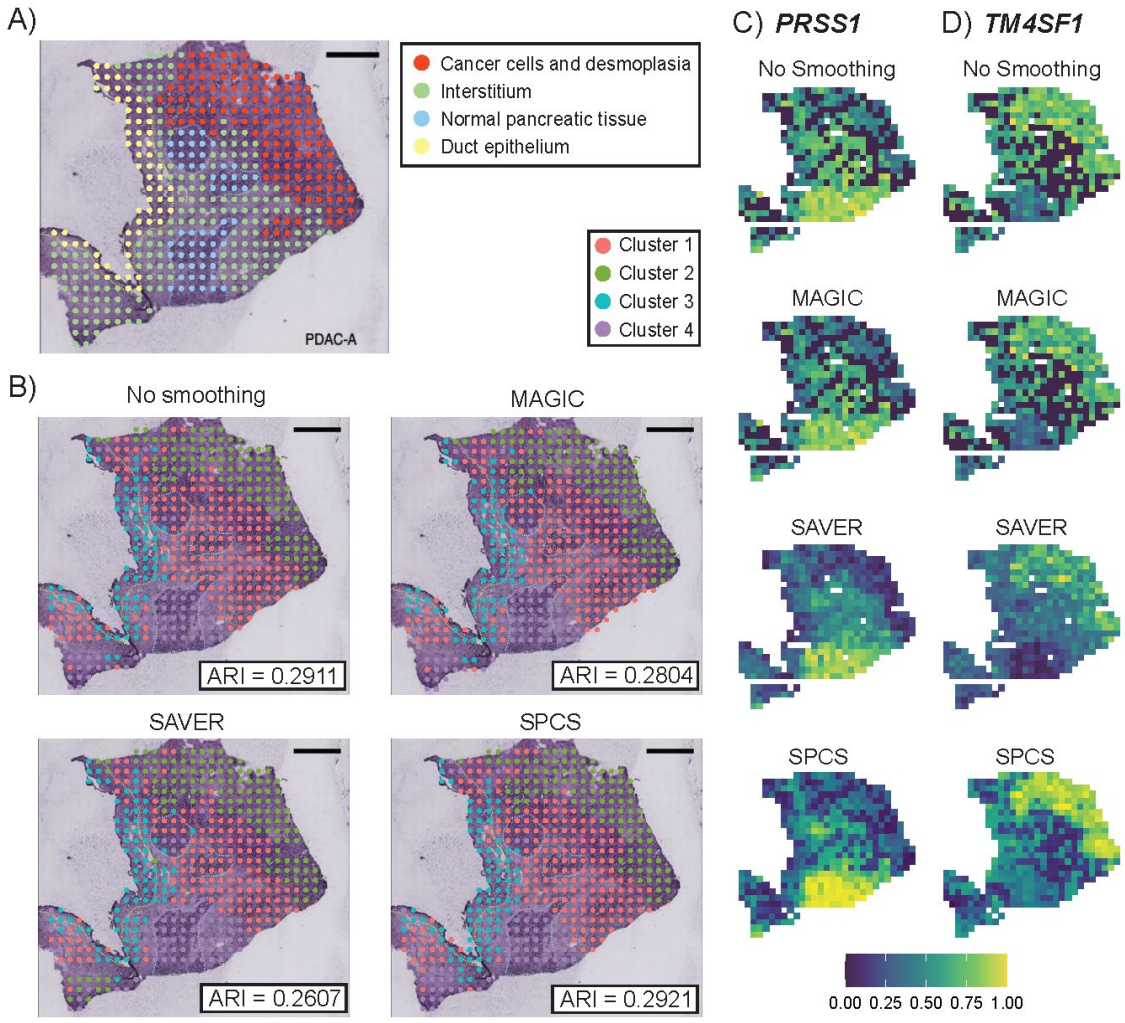


Figure 3 Influence of smoothing on clustering accuracy in PDACA1. (a) Original ST slide of PDACA1 and its histopathological partitions. (b) Results of K-Medoids clustering on expression smoothed by different methods. ARI score, which reflects the correlation between clusters and histopathological labels, 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 linear transformed into range of 0 to 1 as normalization.

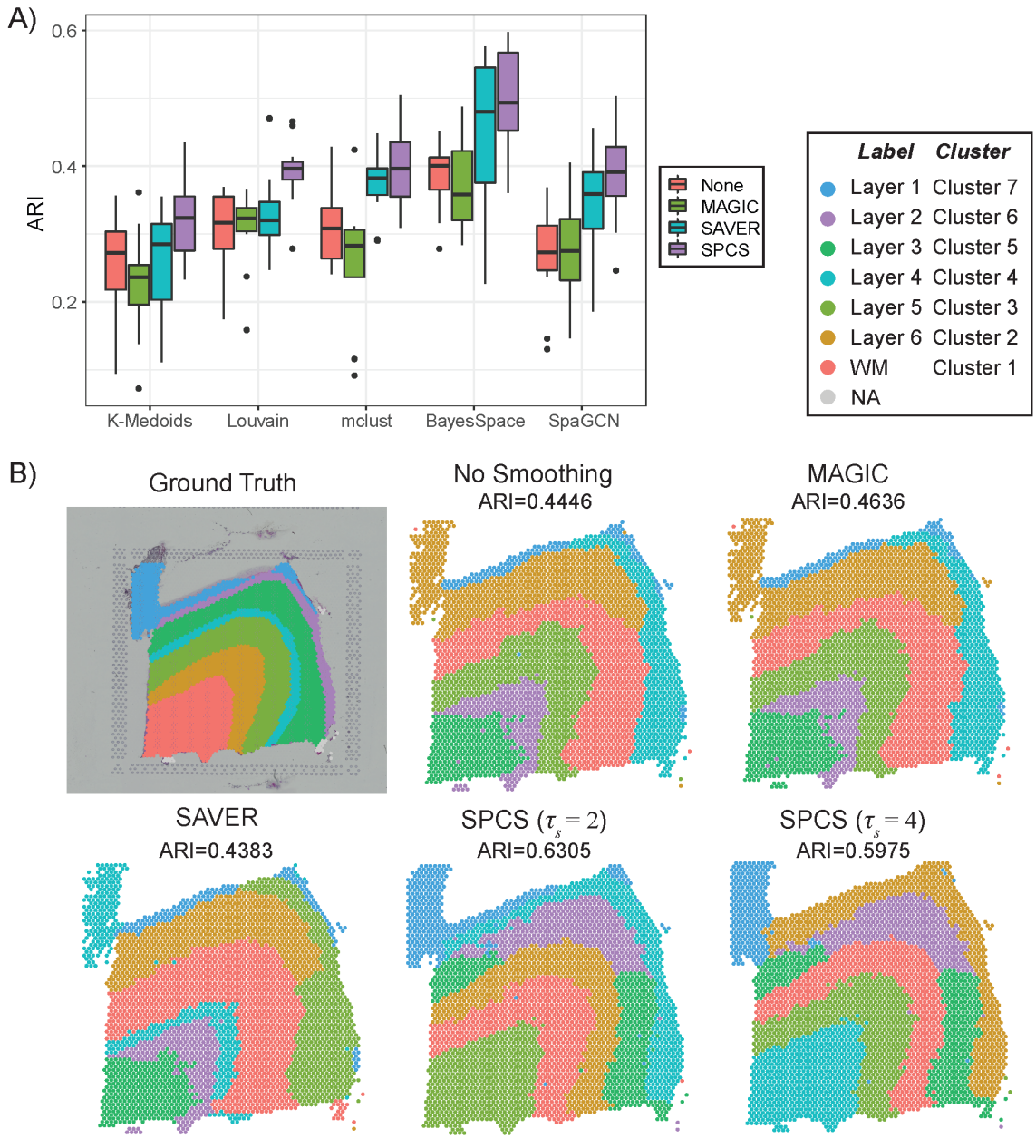


Figure 4 Influence of smoothing on clustering accuracy in DLPFC dataset. (a) Box plot of ARI score on various combination of smoothing (No smoothing, MAGIC, SAVER and SPCS) and clustering methods (k-medoids, Louvain, mclust, BayesSpace and SpaGCN) in 12 DLPFC samples. (b) Ground truth label and BayesSpace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

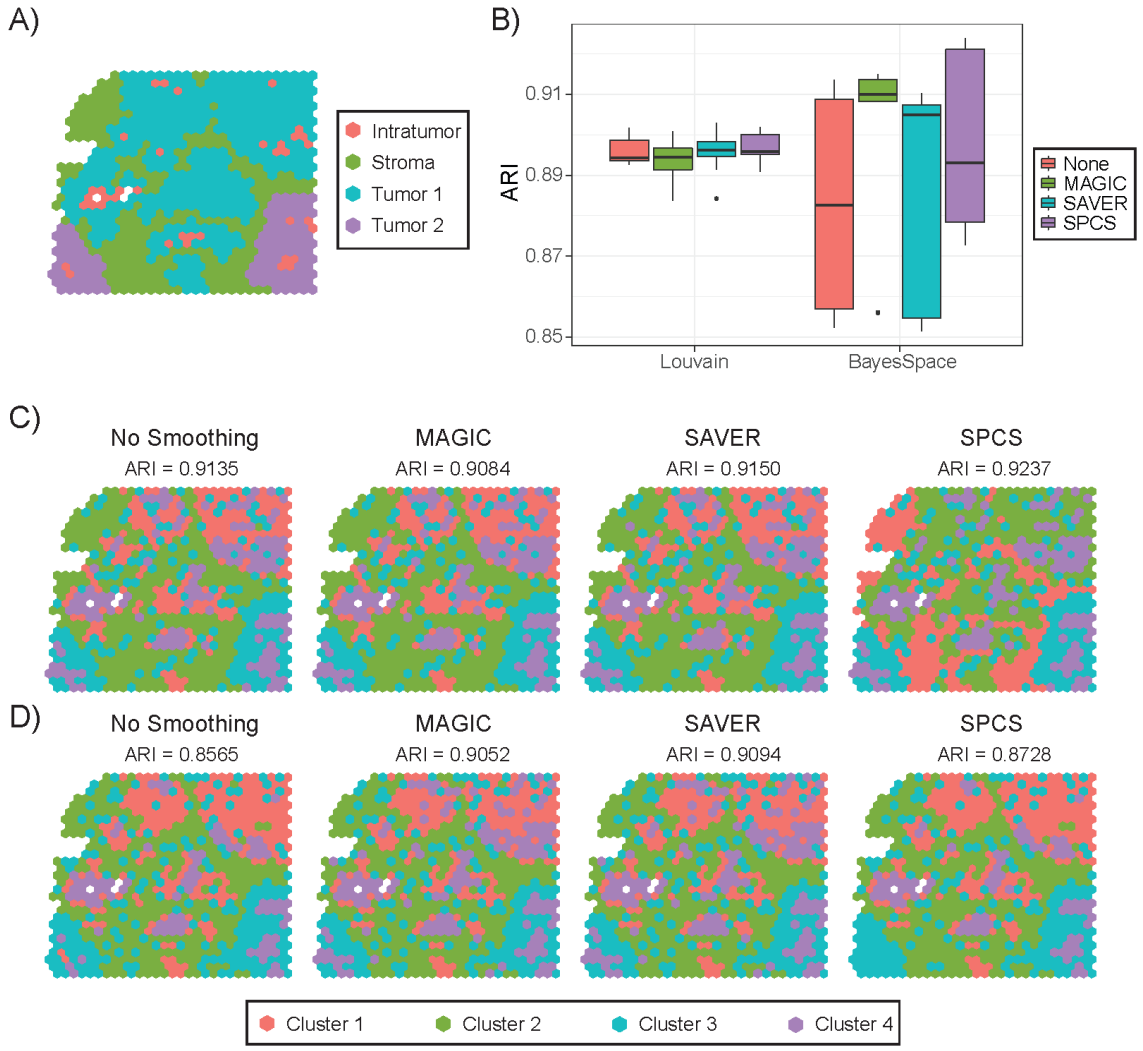
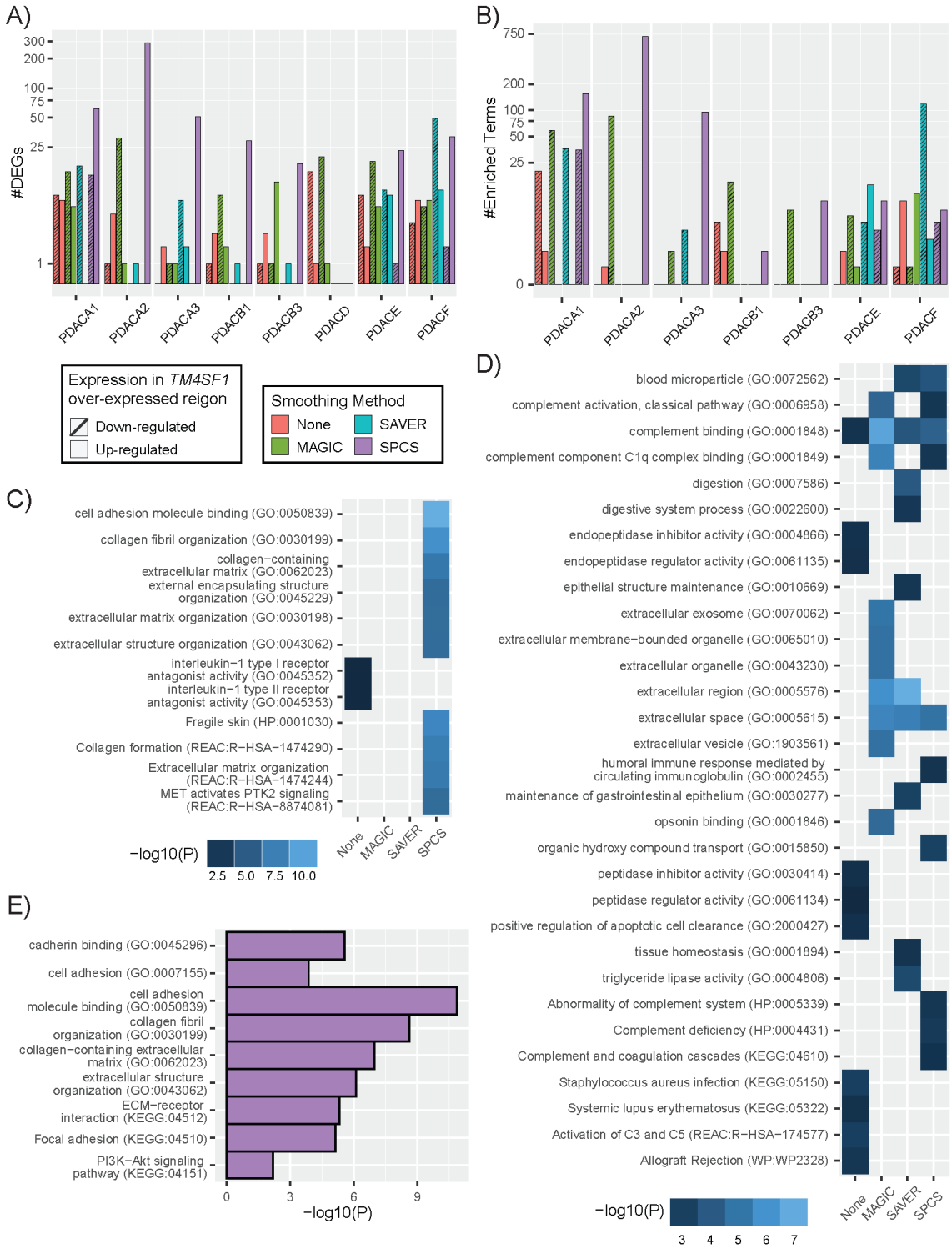


Figure 5 Influence of smoothing on clustering accuracy in HGSOC simulating dataset. (a) Ground truth labels of a simulated ST slide. (b) Box plot of ARI score on Louvain and BayesSpace clustered 10 smoothed simulated ST slides. (c) and (d) illustrate BayesSpace clustering results of two of the simulated ST slides. Clusters are ordered by their size.



746

747 **Figure 6** Biological analysis of unsmoothed and different methods (MAGIC, SAVER and

748 SPCS) smoothed PDAC slides. (a) Number of DEGs identified in each slide and (b) the

749 number of GO terms found from GOEA in each slide shown on a log10 scale. DEGs which
750 are down-regulated in *TM4SF1* over-expressed region and their corresponding GOEA
751 terms are marked as slash filled texture. The ten most significantly enriched terms from
752 each smoothing method found in the (c) up-regulated DEGs and (d) down-regulated DEGs
753 groups in *TM4SF1* over-expressed region for PDACA1 slide are shown, shaded by -
754 log10(p-value). (e) GO terms enriched in SPCS smoothed slides from up-regulated DEGs
755 that have previously been found in PDAC.

Reference

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;73:557-562.

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;54:1812-1817.

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;52:6949-6952.

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;11:2941-2954.

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;38:333-342.

11. He B, Bergenstråhle L, Stenbeck L et al. Integrating spatial gene expression and breast tumour morphology via deep learning, Nat Biomed Eng 2020;4:827-834.

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;182:976-991.e919.

14. Maynard KR, Collado-Torres L, Weber LM et al. Transcriptome-scale spatial gene

792 expression in the human dorsolateral prefrontal cortex, Nat Neurosci 2021;24:425-436.

793 15. Ortiz C, Navarro JF, Jurek A et al. Molecular atlas of the adult mouse brain, Sci
794 Adv 2020;6:eabb3446.

795 16. Vickovic S, Eraslan G, Salmén F et al. High-definition spatial transcriptomics for in
796 situ tissue profiling, Nat Methods 2019;16:987-990.

797 17. Berglund E, Saarenpää S, Jemt A et al. Automation of Spatial Transcriptomics
798 library preparation to enable rapid and robust insights into spatial organization of tissues,
799 BMC Genomics 2020;21:298.

800 18. Dries R, Zhu Q, Dong R et al. Giotto: a toolbox for integrative analysis and
801 visualization of spatial expression data, Genome Biol 2021;22:78.

802 19. Haque A, Engel J, Teichmann SA et al. A practical guide to single-cell RNA-
803 sequencing for biomedical research and clinical applications, Genome Med 2017;9:75.

804 20. Linnarsson S, Teichmann SA. Single-cell genomics: coming of age, Genome Biol
805 2016;17:97.

806 21. Ståhl PL, Salmén F, Vickovic S et al. Visualization and analysis of gene expression
807 in tissue sections by spatial transcriptomics, Science 2016;353:78-82.

808 22. Ding J, Adiconis X, Simmons SK et al. Systematic comparison of single-cell and
809 single-nucleus RNA-sequencing methods, Nat Biotechnol 2020;38:737-746.

810 23. Picelli S, Björklund Å K, Faridani OR et al. Smart-seq2 for sensitive full-length
811 transcriptome profiling in single cells, Nat Methods 2013;10:1096-1098.

812 24. Hashimshony T, Senderovich N, Avital G et al. CEL-Seq2: sensitive highly-
813 multiplexed single-cell RNA-Seq, Genome Biol 2016;17:77.

814 25. Huang M, Wang J, Torre E et al. SAVER: gene expression recovery for single-cell
815 RNA sequencing, Nat Methods 2018;15:539-542.

816 26. van Dijk D, Sharma R, Nainys J et al. Recovering Gene Interactions from Single-
817 Cell Data Using Data Diffusion, Cell 2018;174:716-729.e727.

818 27. Maynard KR, Collado-Torres L, Weber LM et al. Transcriptome-scale spatial gene
819 expression in the human dorsolateral prefrontal cortex, Nature Neuroscience
820 2021;24:425-436.

821 28. Izar B, Tirosh I, Stover EH et al. A single-cell landscape of high-grade serous
822 ovarian cancer, Nature Medicine 2020;26:1271-1279.

823 29. Zhao E, Stone MR, Ren X et al. Spatial transcriptomics at subspot resolution with
824 BayesSpace, Nature Biotechnology 2021;39:1375-1384.

825 30. Jiang DX, Tang C, Zhang AD. Cluster analysis for gene expression data: A survey,
826 Ieee Transactions on Knowledge and Data Engineering 2004;16:1370-1386.

827 31. Berthold MR, Höppner F. On clustering time series using euclidean distance and
828 pearson correlation, arXiv preprint arXiv:1601.02213 2016.

829 32. Wold S, Esbensen K, Geladi P. Principal component analysis, Chemometrics and
830 intelligent laboratory systems 1987;2:37-52.

831 33. Liu Y, Ye X, Zhan X et al. TPQCI: A topology potential-based method to quantify
832 functional influence of copy number variations, Methods 2021;192:46-56.

833 34. Wang Z, Chen Z, Zhao Y et al. A community detection algorithm based on topology
834 potential and spectral clustering, The Scientific World Journal 2014;2014.

835 35. Struyf A, Hubert M, Rousseeuw P. Clustering in an Object-Oriented Environment,
836 Journal of Statistical Software 1997;1:1 - 30.

837 36. Waltman L, van Eck NJ. A smart local moving algorithm for large-scale modularity-
838 based community detection, The European Physical Journal B 2013;86:471.

839 37. Scrucca L, Fop M, Murphy TB et al. mclust 5: Clustering, Classification and Density
840 Estimation Using Gaussian Finite Mixture Models, R Journal 2016;8:289-317.

841 38. Hu J, Li X, Coleman K et al. SpaGCN: Integrating gene expression, spatial location
842 and histology to identify spatial domains and spatially variable genes by graph
843 convolutional network, Nature Methods 2021;18:1342-1351.

844 39. Shahapure KR, Nicholas C. Cluster Quality Analysis Using Silhouette Score. In:
845 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA).
846 2020, p. 747-748.

847 40. Santos JM, Embrechts M. On the use of the adjusted rand index as a metric for
848 evaluating supervised classification. In: International conference on artificial neural
849 networks. 2009, p. 175-184. Springer.

850 41. Fu F, Yang X, Zheng M et al. Role of Transmembrane 4 L Six Family 1 in the
851 Development and Progression of Cancer, Front Mol Biosci 2020;7:202.

852 42. Liu Q, Guo L, Zhang S et al. PRSS1 mutation: a possible pathomechanism of
853 pancreatic carcinogenesis and pancreatic cancer, Mol Med 2019;25:44.

854 43. Dal Molin A, Baruzzo G, Di Camillo B. Single-Cell RNA-Sequencing: Assessment
855 of Differential Expression Analysis Methods, Frontiers in Genetics 2017;8.

856 44. Raudvere U, Kolberg L, Kuzmin I et al. g: Profiler: a web server for functional
857 enrichment analysis and conversions of gene lists (2019 update), Nucleic acids research
858 2019;47:W191-W198.

859 45. Reimand J, Kull M, Peterson H et al. g:Profiler--a web-based toolset for functional
860 profiling of gene lists from large-scale experiments, Nucleic Acids Res 2007;35:W193-200.

861 46. Consortium GO. The Gene Ontology resource: enriching a GOld mine, Nucleic
862 Acids Res 2021;49:D325-d334.

863 47. Jassal B, Matthews L, Viteri G et al. The reactome pathway knowledgebase,
864 Nucleic Acids Res 2020;48:D498-d503.

865 48. Kanehisa M, Furumichi M, Sato Y et al. KEGG: integrating viruses and cellular
866 organisms, Nucleic acids research 2020;49:D545-D551.

867 49. Martens M, Ammar A, Riutta A et al. WikiPathways: connecting communities,
868 Nucleic acids research 2020;49:D613-D621.

869 50. Wingender E. The TRANSFAC project as an example of framework technology
870 that supports the analysis of genomic regulation, Brief Bioinform 2008;9:326-332.

871 51. Giurgiu M, Reinhard J, Brauner B et al. CORUM: the comprehensive resource of
872 mammalian protein complexes-2019, Nucleic Acids Res 2019;47:D559-d563.

873 52. Colwill K, Gräslund S. A roadmap to generate renewable protein binders to the
874 human proteome, Nat Methods 2011;8:551-558.

875 53. Köhler S, Gargano M, Matentzoglu N et al. The Human Phenotype Ontology in
876 2021, Nucleic Acids Res 2021;49:D1207-d1217.

877 54. Zeng H, Shen Elaine H, Hohmann John G et al. Large-Scale Cellular-Resolution
878 Gene Profiling in Human Neocortex Reveals Species-Specific Molecular Signatures, Cell
879 2012;149:483-496.

880 55. Zhou J, Hui X, Mao Y et al. Identification of novel genes associated with a poor
881 prognosis in pancreatic ductal adenocarcinoma via a bioinformatics analysis, Biosci Rep
882 2019;39.

883 56. Atay S. Integrated transcriptome meta-analysis of pancreatic ductal
884 adenocarcinoma and matched adjacent pancreatic tissues, PeerJ 2020;8:e10141.

885 57. Liu L, Wang S, Cen C et al. Identification of differentially expressed genes in
886 pancreatic ductal adenocarcinoma and normal pancreatic tissues based on microarray
887 datasets, Mol Med Rep 2019;20:1901-1914.

888 58. Shang M, Zhang L, Chen X et al. Identification of hub genes and regulators
889 associated with pancreatic ductal adenocarcinoma based on integrated gene expression
890 profile analysis, Discov Med 2019;28:159-172.

891 59. Brenner M. Role of GFAP in CNS injuries, Neuroscience Letters 2014;565:7-13.

892 60. Weniger M, Honselmann KC, Liss AS. The Extracellular Matrix and Pancreatic
893 Cancer: A Complex Relationship, Cancers (Basel) 2018;10.

894 61. Yang JC, Zhang Y, He SJ et al. TM4SF1 Promotes Metastasis of Pancreatic
895 Cancer via Regulating the Expression of DDR1, Sci Rep 2017;7:45895.

896 62. Zheng B, Ohuchida K, Cui L et al. TM4SF1 as a prognostic marker of pancreatic
897 ductal adenocarcinoma is involved in migration and invasion of cancer cells, Int J Oncol
898 2015;47:490-498.

899 63. Ye J, Wen J, Ning Y et al. Higher notch expression implies poor survival in
900 pancreatic ductal adenocarcinoma: A systematic review and meta-analysis,
901 Pancreatology 2018;18:954-961.

902 64. Elazar N, Vainshtein A, Golan N et al. Axoglial Adhesion by Cadm4 Regulates
903 CNS Myelination, Neuron 2019;101:224-231.e225.

904 65. Alcalay RN, Levy OA, Wolf P et al. SCARB2 variants and glucocerebrosidase
905 activity in Parkinson's disease, npj Parkinson's Disease 2016;2:16004.

906 66. Giussani P, Prinetti A, Tringali C. The role of Sphingolipids in myelination and
907 myelin stability and their involvement in childhood and adult demyelinating disorders,
908 Journal of Neurochemistry 2021;156:403-414.

909