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

Yusong Liu[a], Tongxin Wang[b], Ben Duggan[c], Kun Huang[d, e], Jie Zhang*[f],

Xiufen Ye*[a], and Travis S. Johnson*[d]

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

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

[f] Department of Medical and Molecular Genetics, Indiana University School of Medicine, 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

Collage 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

The recently developed spatial transcriptomics (ST) technique has made it possible to view spatial transcriptional heterogeneity in a high throughput manner. It is based on highly multiplexed sequence analysis and uses barcodes to split the sequenced reads into respective tissue locations. However, this type of sequencing technique suffers from high noise and drop-out events in the data, which makes smoothing a necessary step before performing downstream analysis. Traditional smoothing methods used in the similar single cell RNA sequencing (scRNA-seq) data are one-factor methods that can only utilize associations in transcriptome space. Since they do not account for associations in the Euclidean space, i.e. tissue location distances on the ST slide, these one-factor methods cannot take full advantage of all the knowledge in 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 associations from transcriptome and Euclidean space from the ST data. By performing SPCS on 10 ST slides from pancreatic ductal adenocarcinoma (PDAC), smoothed ST slides have better separability, partition accuracy, and biological interpretability than the ones smoothed by pre-existing one-factor smoothing 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

Introduction

Tumors and organ tissues are highly heterogeneous with different phenotypes distributed throughout the tissue [1, 2]. Until recently, this tissue heterogeneity had to be evaluated using histology image analysis or using scRNA-seq. Histology imaging offers a view on spatial tissue heterogeneity whereas scRNA-seq offers insight into the transcriptional heterogeneity. Methods like immunohistochemistry (IHC) labelled histology slides also allow researchers to view protein in tissue images [3-7]. However, IHC is only capable of capturing protein expression level(s) of one or a few proteins at a time. These technologies lacked the ability to view transcriptional heterogeneity spatially in a high throughput manner.

Fortunately, revolutionary new techniques that can be broadly categorized as spatial transcriptomics (ST) have recently become available that allow transcriptome profiles to be measured for different locations in a tissue [8, 9]. These techniques have already been profoundly useful in understanding tumor heterogeneity [10-12] and heterogeneity of brain tissues [13-16]. In recent years, research on this new ST technology has been prolific -- in 2017 there were only five articles mentioning “spatial transcriptomics” in PubMed (https://pubmed.ncbi.nlm.nih.gov/), but by 2020 there were a total of 70 articles, 14 times as many as just three years earlier. However, ST technology is still largely being studied methodologically to improve the library preparation [17], sequencing techniques, and bioinformatic analysis pipelines [18] used in spatial transcriptomic analysis. In comparison, scRNA-seq analysis has now largely established standard practice protocols [19, 20] such that the data preprocessing and initial bioinformatic analyses are consistent between experiments. Once ST techniques has established the standard practice, like single cell technologies, then they will be widely available for any number of disease and normal

tissues.

The most common form of ST data to date, much like the now widely abundant scRNA-seq data, is based on highly multiplexed sequence analysis where barcodes are used to sort the sequenced reads into their respective cells or tissue locations [21]. This type of sequencing suffers from a wide array of noise and drop-out events in the data which has been well studied across scRNA-seq methodologies [22]. In barcoding-based single cell data, a single cell's expression profile on its own is not perfectly representative of the true underlying RNA abundances from that cell due to the noise introduced during the sequencing process. There are techniques such as Smart-seq2 [23] and CEL-seq2 [24] that give more representative expression profiles per cell, but these are much lower throughput, resulting in many fewer cells. As a result, a multitude of techniques have been developed to impute the missing expression values and smooth the noise that come directly from the barcode-based scRNA-seq. These techniques help to smooth the data by using similar cells to correct for technical errors in the sequencing data. Specifically, sets of correlated genes and relative cell similarity in transcriptome space is used to impute the drop-out events and eliminate other types of expression errors via machine learning techniques. Some of these tools include SAVER [25] and MAGIC [26]. When applied, the processed expression values give more accurate representations of the true underlying RNA abundances than the raw read counts.

Because most current ST data are also generated from barcode-based technology, it is not surprising that ST data also suffer from similar noise and drop-out events. This means that the distributions of gene expression in the tissue does not always follow the corresponding phenotypic differences in histopathological images at a high resolution. For drop-out events, it can be assumed that zero expression values have some probability of being spuriously missed in the reads simply due to chance. However, ST data, unlike

scRNA-seq, not only has relative similarity in transcriptome space but also in Euclidean space on the tissue slide. That means locations that are adjacent to one another in the ST array will likely be similar tissue and as a result have similar transcriptome profiles. In addition, if two locations in an ST array share a majority of their transcriptome profiles, then their transcriptional differences can be partially explained by drop-out.

Here we present a novel two-factor ST smoothing method that utilizes both the associations of spatial locations in transcriptome space (expression pattern knowledge) and in Euclidean space (spatial knowledge). This is in contrast to the pre-existing one-factor scRNA-seq techniques that only utilize associations in transcriptome space. We believe that by using both the spatial location of a given ST spot and its transcriptome profile we can better smooth the data such that it better represents the true underlying RNA abundances in the tissue.

Methods

Datasets

The datasets that we use in this study are from an ST study of PDAC [10]. In this dataset, there are a total of 10 ST slides from 5 PDAC patient tumors. These data 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 of the relatively large number of ST slides in a single disease, i.e., 10 slides in PDAC, this dataset is a good testbed for our methods because multiple datasets do not

need to be integrated.

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

For each spot on an ST slide, there exists not only the gene expression but also their 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 expression pattern and spatial location on the ST slide. To achieve this goal, we put forward a KNN-based method, Spatial and Pattern Combined Smoothing (SPCS), to perform the smoothing. We display the procedure of 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 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 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 deleted spatial and pattern neighborhood of object spot i respectively. Unless otherwise stated, the neighborhood in this paper is deleted neighborhood. 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 are spatial and pattern smoothing contribution of corresponding neighbor spot j of object spot i. In this

piece of work, we set $\alpha = 0.6, \beta = 0.4, \tau_s = 2$ and $\tau_p = 16$. We will discuss the selection of these parameters, which define the size of neighborhoods and their contributions, in the Discussion section. Source code of our proposed SPCS method is provided in Github (https://github.com/Usos/SPCS). In the next section, we introduce the concepts of neighborhood and contribution and their detailed mathematical definitions.

Pattern neighborhood

ST data is a type of transcriptomic data, which means it includes 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 spot with those of similarly expressing spots can improve data quality and is common in traditional one-factor smoothing methods [25, 26]. 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 gene expression pattern space of a 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 τ_p-deleted 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 [27]. 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 [28]:

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

In ST data, some genes are expressed at identical or near-identical levels, which lack of variance to establish an accurate pattern neighborhood. Therefore, we used principal component analysis (PCA) [29], to transform the expression of spots into a 10-dimensional principal component (PC) space before smoothing, to extract uncorrelated components with the largest variance. We named this PC space as gene expression pattern space.

Spatial neighborhood

In contrast to scRNA-seq data, ST data can provide spatial position for each spot in the slide. Regions in proximity on histopathology slides are usually 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 τ_s-deleted 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. 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.

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, we used Pearson correlation coefficient distance described in eq. (2) as distance metric, whose range is $[0,2]$. If pattern distance between two spots is greater than 1, the expression of these two spots will have a negative correlation. Smoothing with negative correlation spots is not performed since they are dissimilar, so 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. In addition, when $d_{ij}^{p} > \frac{3\sigma}{\sqrt{2}}$, c_{ji}^{p} will quickly decay to 0 [30, 31]. Hence, setting σ to $\frac{\sqrt{2}}{3}$ will cease the effect of negative correlation neighbors and we can simplify the eq. (3) as:

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

For spatial neighbors, we used Manhattan distance, an integer greater than 0, as spatial

distance to measure their distance to 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.

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 ST datasets. Then we partitioned the smoothed ST slides and the original unsmoothed slides using K-medoid clustering method and judged the effectiveness of multiple smoothing methods by internal and external evaluation of the unsupervised clustering. The internal validation refers to how well the clusters can be separated whereas the external evaluation refers to how well the clusters match the histopathological labels from the corresponding image. As a gene filter, we only kept genes with less than 70% zero expressed spots in our analysis. For normalization, we performed logarithmic count per million (CPM) normalization before smoothing using SPCS, SAVER, and MAGIC. The distance metric used during clustering was Pearson correlation distance as described in eq. (2). Because there were four histopathological labels, we clustered the spots into four clusters for each slide. PCA was performed prior to clustering, then we selected the eigenvectors with eigenvalue greater than 1 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 [32]. Silhouette score, a metric whose range is [-1,1], estimates the average distance between clusters. A greater silhouette score indicates better cluster separations. For a 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:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{6}$$

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

In this piece of 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 based on Pearson correlation distance. 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.

External evaluation

For the external evaluation, we obtained the histopathological labels of the ST spots, which included four different labels: cancer region, pancreatic tissue, duct epithelium, and stroma. The correspondence between smoothed slides and histopathological labels was evaluated at both clustering level and gene expression level. At the clustering level, the imputed ST data is clustered into four clusters using the same clustering algorithm as the

internal evaluation. Next, the concordance was evaluated between the clusters from the imputed ST data and the labels from the corresponding histopathological images. Furthermore, 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 [33]. 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:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{n_i}{2} + \sum_j \binom{n_j}{2}\right] - \frac{\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}}{\binom{n}{2}}} \tag{7}$$

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 the marker gene evaluation, 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 highly expressed in pancreatic cancer among other malignancies [34, 35]. High expression of *TM4SF1* in the cancerous regions of the PDAC dataset has been detected in previous research [10]. 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 a ST sample, we identified differentially expressed genes (DEGs) and performed gene ontology enrichment analysis (GOEA). DEGs were identified by comparing two groups of spots with *TM4SF1*-high and low expression. 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 less than 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 [36]. Only genes with an adjusted p-value less than 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 greater than 1 were considered up-regulated, and down-regulated if the logFC were less than -1.

GOEA was performed on the DEGs using *g:Profiler* [37]. The significance of enriched terms were tested by cumulative hypergeometric test and p-values were corrected by *g:SCS* method [38]. Only terms with an adjusted p-value below 0.05 were reported. All data sources offered by *g:Profiler* were used including Gene Ontology (GO) [39], Reactome [40], KEGG [41], WikiPathways [42], Transfac [43], CORUM [44], Human Protein Atlas [45], and the Human Phenotype Ontology [46]. Heatmaps for each sample were then generated to compare the terms found by each smoothing algorithm.

Results

In this piece of work, we applied our two-factor smoothing method SPCS and two state-of-the-art one-factor smoothing methods (MAGIC and SAVER) to smooth ten PDAC ST slides. To compare the performance of different smoothing methods, we evaluated the quality of generated clusters after performed K-Medoids unsupervised clustering on both smoothed and unsmoothed expression. The generated clusters and the distribution of marker gene expressions were compared with pre-labeled histopathological partitions to check if the smoothed expressions are more accurately reflecting the pathology features of the corresponding images. In addition, the DEGs were identified in different regions with respect to the marker gene *TM4SF1* expression level and GOEA was performed to reveal the biological meanings revealed by the smoothing.

Internal Evaluation

The silhouette score indicates average distance between clusters in a slide. For different smoothing methods, a greater silhouette score usually represents smoothed expressions have better separability. After clustering the spots into four clusters, we calculated the silhouette score of ten slides for each smoothing method in Figure 2(a). Our newly proposed SPCS method outperformed the other smoothing methods in six out of ten slides and was at least second best in the other four slides, indicating that SPCS smoothed ST spots have better separability. However, due to the bad data quality and irregular tissue structure, performing any of the smoothing methods we tested reduced the separability of ST spots in two slides (i.e., PDACA3 and PDACG). This indicates that smoothing generally improves data quality though the quality of the raw ST data and underlying tissue structure may impact the utility of smoothing.

As an example from PDACA1, it was clear smoothing improved the silhouette scores in Figure 2(b). Different smoothing methods improved the separability of original spots to varying degrees. In this case, MAGIC only made the distribution of silhouette width change modestly, which had little effect on improving separability. Alternatively, the other two methods provided significant improvement in distribution of silhouette width and provided more improvement on spot separability. Since there were a smaller number of false clustered spots, the expression smoothed by SPCS got the highest silhouette scores. This result indicates that our proposed SPCS smoothing method can provide better separability than the existing smoothing methods. I don't see PCC distance involved in this evaluation, but it was mentioned in the method.

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 was shown in Figure 3 since the histopathological labels were available for PDACA1. From Figure 3, we observed that, compared to the original ST data, smoothing methods did increase the coherence between the clusters defined by ST data and the original labels of ST spots on the histopathological slides. Moreover, clustering results from the SPCS smoothed data showed a higher correlation with the histopathological labels comparing with the two existing smoothing methods, as shown by ARI.

When examining the results in detail, Figure 3(b) shows that most ST spots in the cancer cells and desmoplasia region can be separated from other clusters even without smoothing on the ST data, but failed to separate the duct epithelium, the interstitium, and the normal pancreatic tissue regions. Similar to this, MAGIC also failed to separate these three different histopathological regions. SPCS and SAVER produced results that could

separate most ST spots in the duct epithelium region, with slight mixing of the interstitium and the normal pancreatic tissue regions. Compared with SAVER, SPCS produced a clearer boundary that separated the interstitium lying within the normal pancreatic tissue and was reflected by the higher ARI score.

Figure 3(c, d) show the influence of different smoothing methods on the two marker genes expression. Compared with other smoothing methods, SPCS generated better marker gene contrast between distinctive histopathological areas, which is due to 1) By performing smoothing, marker gene expressions showed fewer drop-outs and better reflected expression patterns across different tissue regions with SAVER and SPCS (MAGIC does not impute the missing expressions); 2) in contrast to SAVER, SPCS imputed the missing values and kept the distribution of non-missing values unchanged. All of these properties of SPCS contribute to the higher ARI score for SPCS.

Biological Analysis

To biologically verify ST smoothing methods, the biological interpretability of the smoothed results was compared between different methods. When comparing between *TM4SF1 over- and under-expressed region*, the significant DEG numbers were displayed in Figure 4(a). There were no DEGs found in any smoothed or unsmoothed slides of PDACB2 and PDACG while a higher number of DEGs was 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 by Figure 4(b), which are further examined below.

The top ten most significant GO terms found in slide PDACA1 for each algorithm are shown in Figure 4(c) for the up-regulated DEGs and Figure 4(d) for the down-regulated DEGs. 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 activates PTK2 signaling. The GO terms from all smoothed and unsmoothed up-regulated slides can be seen in Figure S1. 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 do not seem related to PDAC, except for slide PDACF which contained cell adhesion and ECM terms for the no smoothing and MAGIC.

The enriched GO terms for the down-regulated DEGs (Figure 4(d)) show a different scenario. 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 showed are in Figure S1 and appear similar results to the terms described above.

Figure 4(e) contains terms enriched in the SPCS smoothed data from up-regulated DEGs that have previously been reported in PDAC [47-50]. 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-ten enriched terms), but were absent from unsmoothed data, reflecting the enhancement of biological interpretability by performing our novel SPCS smoothing method.

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, only 485 of 19738 genes (2.5%) in PDACA1 slide was left, which indicated that drop-out events are highly frequent in the ST dataset. By visualizing the expression distribution of two marker genes, *PRSS1* and *TM4SF1* in Figure 3, the drop-out events can be easily seen, as indicated by the blank spot 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. 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 step 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 on both aspects as compared with the two existing one-factor smoothing methods MAGIC and SAVER. For internal evaluation, from the results shown in Figure 2, SPCS smoothing method produces greater

silhouette scores and lower false clustering rates than MAGIC and SAVER, which means the ST data smoothed by SPCS has better separability.

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. Indeed, as expected, results in Figure 3 reveal SPCS method generates a higher ARI score than existing one-factor methods, which means a more accurate histopathological partition can be acquired by performing two-factor SPCS method. From the marker gene analysis, SPCS method recovers the drop-out events and enhance 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

The GO terms reported in PDACA1 from data using the SPCS smoothing method are more interpretable in the context of pancreatic cancer and the role of *TM4SF1* in cancer than non-smoothed data and the two one-factor smoothed data. *TM4SF1* has been found to be overexpressed in pancreatic cancer and has roles in apoptosis, proliferation, and cell migration [34]. Previous work found that collagen 1 (GO:0030199) binds with DDR1 which then interacts with TM4SF1 to activate the focal adhesion kinase (FAK) [34, 51, 52]. This results in disruption of E-cadherin (GO:0045296) leading to the Wnt signaling pathway and loss of cell-cell adhesion (GO:0050839) [34, 51]. The FAK pathway also increases expression of N-cadherin (GO:0045296) which results in migration of cancer cells [34, 51]. TM4SF1 can also activate the *AKT* pathway (KEGG:04151) leading to anti-

apoptotic effects and angiogenesis [34, 53]. All of these previously identified GO terms are important factors for PDCA 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.

Identifying the DEGs in different histological regions and check their associated GO terms can also help evaluate the biological interpretability of a smoothed slide. Our SPCS method contributed to identify many more differentially expressed genes than the other smoothing methods for most of the slides. Among the DEGs, the up regulation of *IL1RN*, *KRT7*, *LAMB3*, *LAMC2*, *NOTCH3, and S100A16,* which have previously been reported associate with poor survival in PDAC, was also detected in the SPCS smoothed PDACA1 slide [48, 54]. *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. Generally, a high number of DEGs also associate with more GO terms. However, some unsmoothed, SAVER, and MAGIC-smoothed slides contained DEGs but no GO terms, which may indicate the insufficient imputing or smoothing without considering the location information. 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. For this reason, it's possible SPCS found more DEGs because other genes in the slide had more defined expression after smoothing.

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 ten terms in Figure 4(c) and previously reported terms in Figure 4(e) are similar to those found in the literature [47-50] and are consistent between slides. Without applying SPCS, some of the top terms found in Figure 4(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 is not smoothed or not properly smoothed.

Determination of SPCS parameters

There are four parameters in SPCS: τ_s, τ_p, α and β. These parameters affect the performance of SPCS algorithm by adjusting neighborhood size and smoothing intensity and should be tuned according to the quality and distribution of original ST data.

τ_s and τ_p are designed to adjust the size of spatial neighborhood and pattern neighborhood, respectively. Obviously, 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. However, blindly expanding neighborhoods will incorporate some spots that are not similar to the spot being smoothed, reducing the quality of smoothing. In our novel SPCS method, we use pattern and spatial contribution to reduce this effect, but this will increase computational cost. Therefore, we recommend a modest selection of these two parameters to balance the tradeoff. In PDAC dataset, $\tau_s = 2$ and $\tau_p = 16$ is a suitable selection.

The parameters, α and β, are designed to control the degree of smoothing, which has a significant effect on smoothing quality. α is used to balance the expression of spot being

smoothed with the contributions from its neighbors. For clustering problems, a greater α usually lead to a better silhouette score since it will make the expression of a spot trend to more consistency with their neighbors. However, too large α will cause over smoothing and eliminate the uniqueness of each spot. Therefore, we recommend setting α between 0.2 to 0.8. Another intensity parameter β is used to balance the contribution of spatial and pattern neighbors. A greater β adds more contribution from spatial neighbors and usually leads to worse silhouette score. This is mainly caused by the boundary effect of the spatial neighbors. Spatially neighboring spots may not belong to the same category due to the irregular boundaries between different regions of tissue. For example, if a slide contained a straight boundary that clearly separated cancerous versus non-cancerous tissue then this would be a simpler problem than a slide containing a highly complex boundary with mixtures of cancerous and non-cancerous tissue in the same areas. Smoothing a spot using such spatial neighbors will generate error. This effect is especially significant when the tissue is highly heterogeneously spatially. Including pattern neighbors that can remit such boundary effect, but this remission will be voided if β is set too large since large β offsets the effect of loss weight of the spatial neighborhood. Therefore, we recommend β is set less than 0.6 in most cases and should be selected carefully according to data distribution.

Conclusion

Due to the limitation of barcoding based sequencing technologies which result in noisy and high drop-out in transcriptomics data, 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, we have demonstrated that SPCS can recover the missing or erroneous expression in ST data better and improve the separability, partition accuracy, and biological interpretability. This reflects that our proposed SPCS method can effectively improve data quality of ST and contribute to a more accurate and meaningful downstream analysis results. SPCS smoothing is a state-of-the-art ST smoothing algorithm with implications in numerous diseases where ST data is being generated.

Key Points

- Due to the commonly arisen 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 knowledges and spatial knowledges in ST data.
- Compared with traditional expression pattern knowledge-based one-factor smoothing methods, SPCS can provide better separability, partition accuracy and biological interpretability.

Funds

This work is partially supported by IUSM startup funds [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.].

Figures

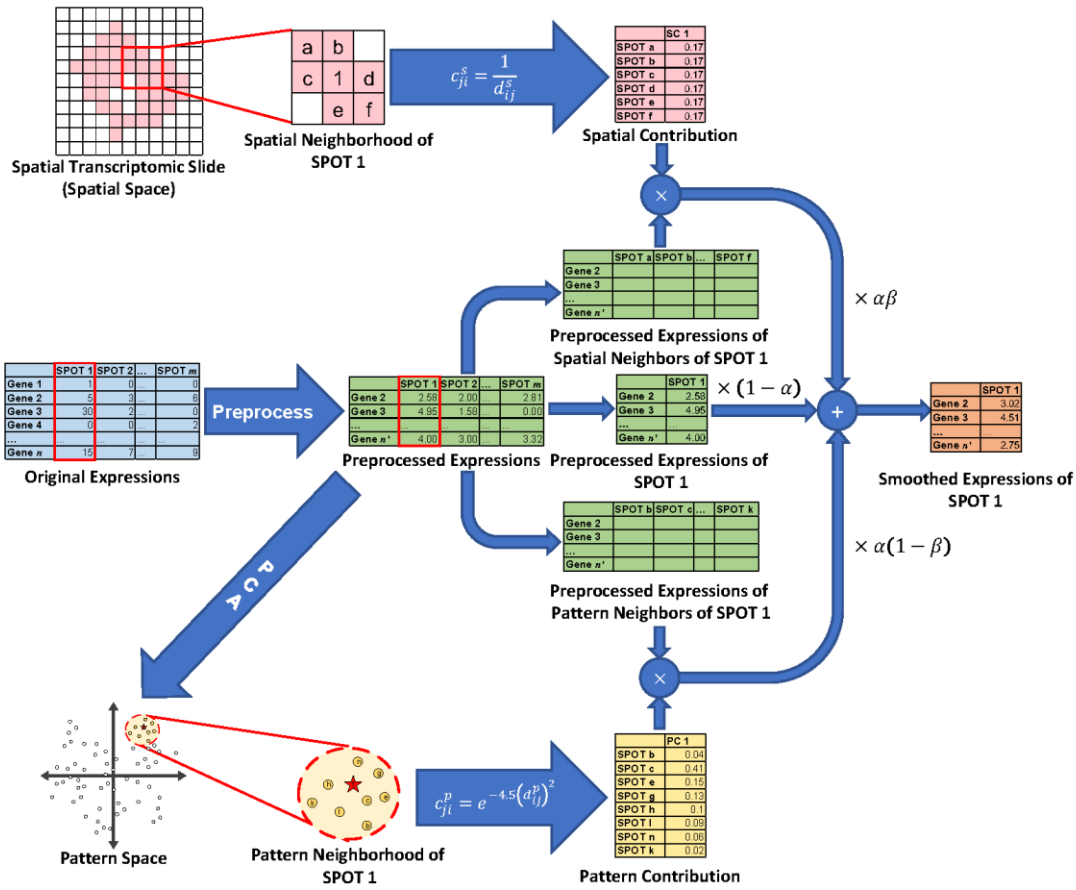

Figure 1 Workflow of our proposed SPCS smoothing method. In this sample, SPOT 1 is the object spot that we are going to smooth. For ST slides, preprocessing steps normalize the expressions and filter out genes with zero expression in most spots. Spatial neighborhood is first order neighborhood of SPOT 1 where have six spots. Pattern neighborhood is the top eight spots with most similar expression pattern to SPOT1. Both spatial and pattern contribution are normalized which means total contribution of all neighbors is equal to 1. Smoothing is performed by integrating parameter- and contribution-weighted expressions of the object spot itself and its both spatial and pattern neighbors.

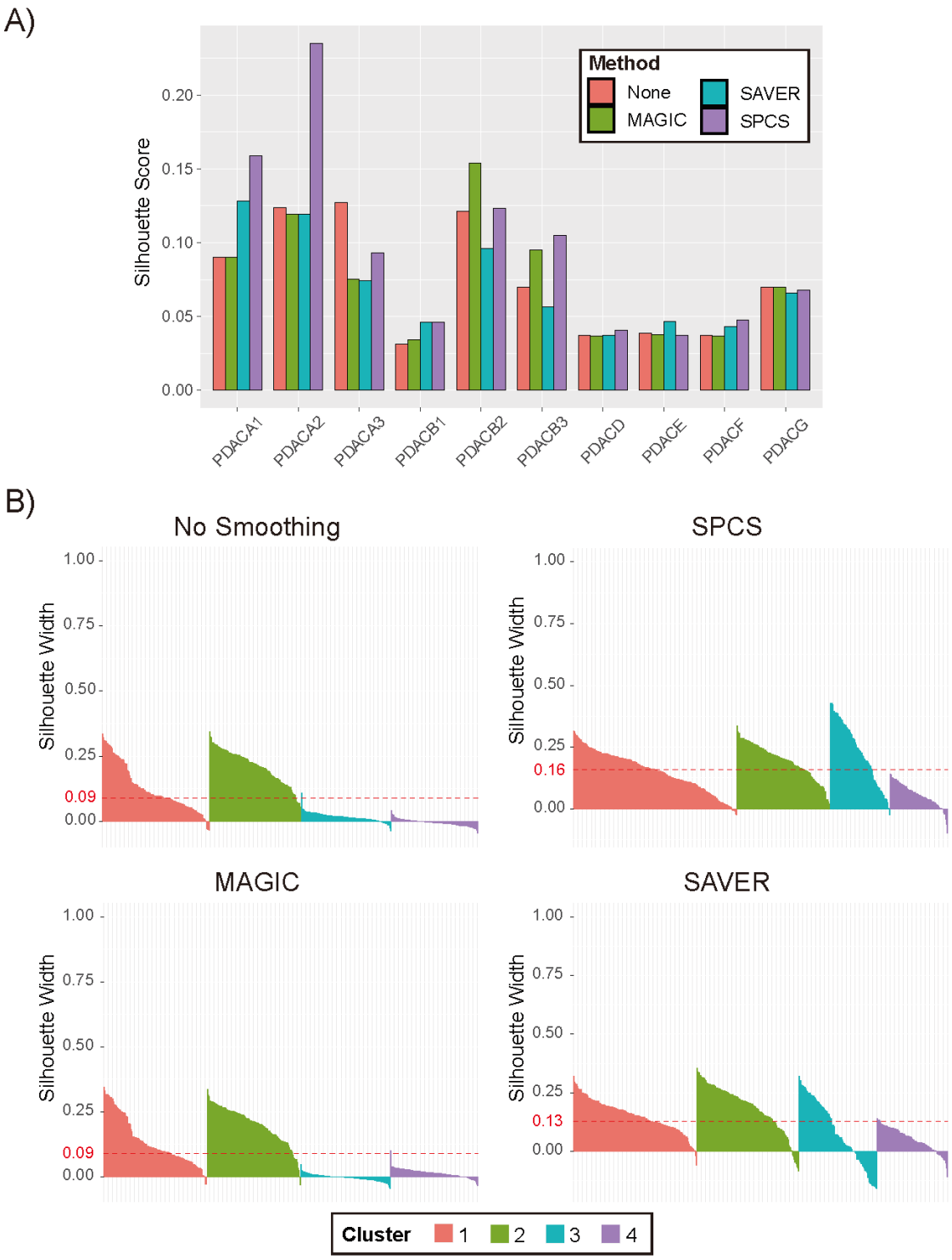

Figure 2 Influence of smoothing on data separability. (a) The average silhouette score of ten PDAC ST slides without smoothing and with smoothing by different methods (MAGIC, SAVER and SPCS). Silhouette plot of unsmoothed and different method smoothed

PDACA1 is shown in (b). Clusters are arranged in descending order of size. Red dashed lines in each figure represent silhouette score of whole smoothed (or unsmoothed) expressions.

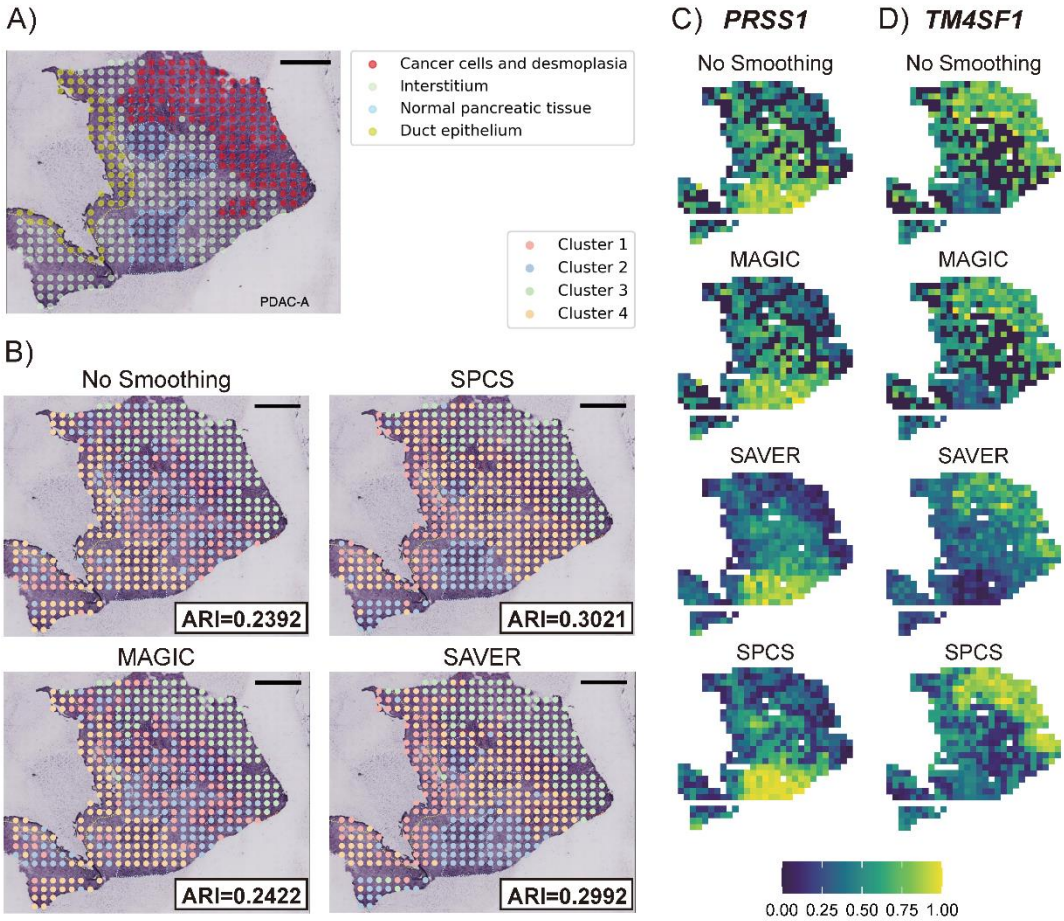

Figure 3 Influence of smoothing on clustering accuracy. (a) displays original ST slide of PDACA1 and its histopathological partitions, (b) illustrates results of K-Medoids clustering on expression smoothed by different methods. ARI score, which reflecting correlation between clusters and histopathological labels, is marked at bottom-right corner of each figure. Heatmap of smoothed expressions 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.

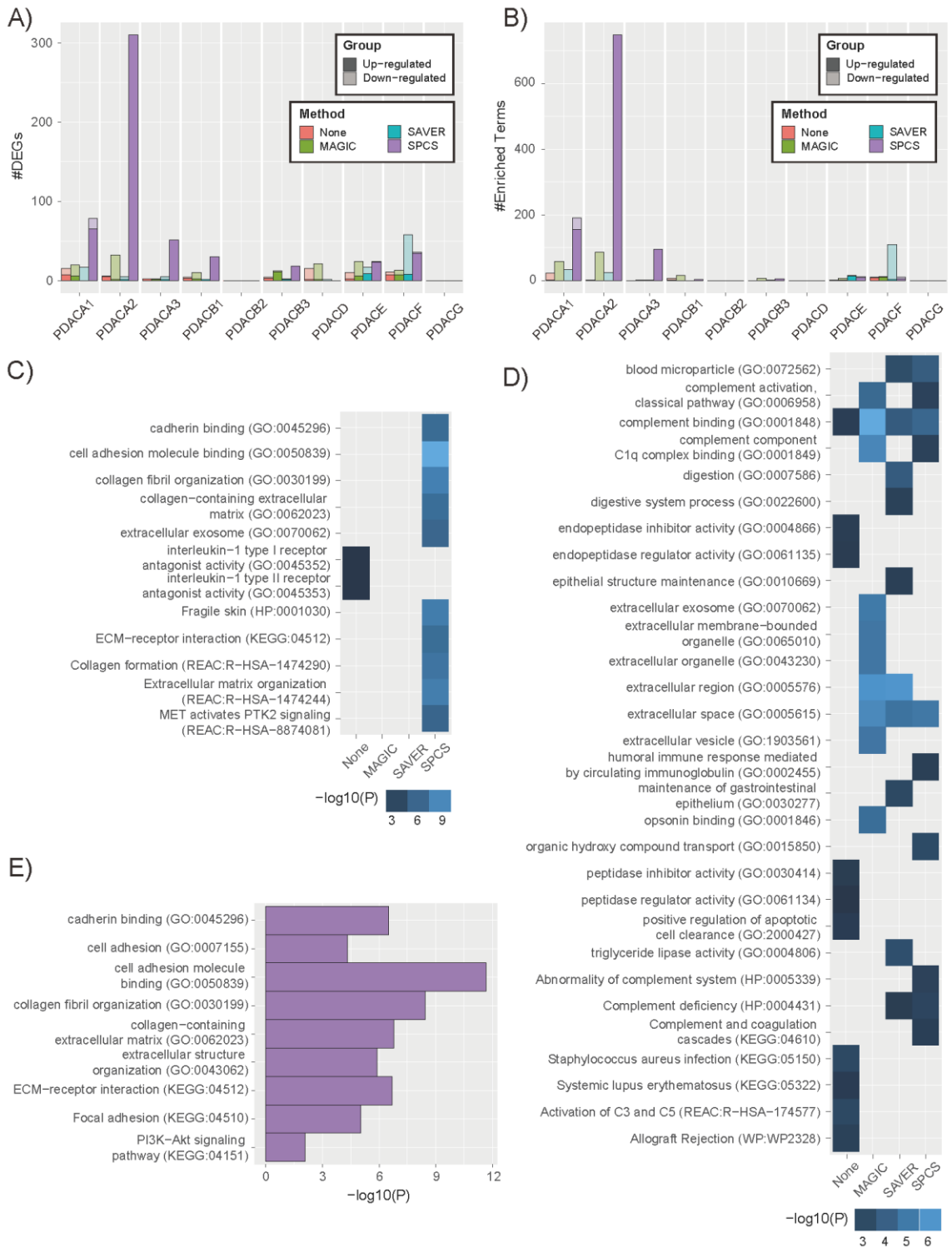

Figure 4 Biological analysis of unsmoothed and different methods (MAGIC, SAVER and SPCS) smoothed slides. (a) Number of DEGs identified in each slide and (b) the number

of GO terms found from GOEA in each slide. The ten most significantly enriched terms 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.

Figure S1 Enriched GO terms for all smoothed and unsmoothed ST slides: (a) PDACA2 (b) PDACA3 (c) PDACB1 (d) PDACB3 (e) PDACE and (f) PDACF.

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

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

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;16:987-990.

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;38:737-746.

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

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;15:539-542.

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

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

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

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

30. 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.

31. Wang Z, Chen Z, Zhao Y et al. A community detection algorithm based on topology

potential and spectral clustering, The Scientific World Journal 2014;2014.

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

33. 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. 2009, p. 175-184. Springer.

34. 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.

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

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

37. 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 research 2019;47:W191-W198.

38. 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.

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

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

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

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

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

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

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

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

47. 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.

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

49. 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;20:1901-1914.

50. 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;28:159-172.

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

52. 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.

53. 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;47:490-498.

54. 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;18:954-961.