De-noising Spatial Expression Profiling Data Based on *in situ* Position and Image Information

Yunguan Wang, Ph.D.[1,+], Bing Song, Ph.D.[1,+], Shidan Wang, Ph.D.[1], Mingyi Chen, M.D.[2], Yang Xie, Ph.D.[1,3], Guanghua Xiao, Ph.D.[1,3], Li Wang, Ph.D.[4,*], and Tao Wang, Ph.D.[1,5,*]

1 Quantitative Biomedical Research Center, Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA, 75390.
2 Department of Pathology, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA
3 Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA
4 Department of Mathematics and Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019, USA
5 Center for the Genetics of Host Defense, University of Texas Southwestern Medical Center, Dallas, TX, USA, 75390.
+ Co-first authors
* Corresponding authors: (1) Tao Wang, Ph.D., Quantitative Biomedical Research Center, Department of Population and Data Sciences, UT Southwestern Medical Center, Dallas, TX, 75390, USA; E-mail: Tao.Wang@UTSouthwestern.edu **(Lead contact)**; (2) Li Wang, Ph.D., Department of Mathematics and Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019, USA; E-mail: li.wang@uta.edu.

Short title: Sprod for Spatial Expression Profiling De-noising

Key words: Visium, Slide-seq, noise removal, drop-out, gene expression, Sprod

ABSTRACT

Spatial expression profiling (SEP) technologies provide gene expression close to or even superior to single-cell resolution while retaining the physical locations of sequencing and often also providing matched pathology images. However, the expression data captured by SEP technologies suffer from high noise levels, including but not limited to drop-outs as in regular single-cell RNA-sequencing (scRNA-seq). The extra experimental steps for preserving the spatial locations of sequencing could result in even more severe noises, compared to regular scRNA-seq. Fortunately, such noises could be largely reduced by leveraging information from the physical locations of sequencing and the tissue and cellular organization reflected by corresponding pathology images. In this work, we demonstrated the extensive levels of noise in SEP data. We developed a mathematical model, named Sprod, to reduce such noises based on latent space and graph learning of matched location and imaging data. We comprehensively validated Sprod and demonstrated its advantages over prior methods for removing drop-outs in scRNA-seq data. We further showed that, after adequately de-noising by Sprod, differential expression analyses, pseudotime analyses, and cell-to-cell interaction inferences yield significantly more informative results in various biological application settings. In particular, with Sprod, we discovered 3-4 times more RNA transcripts that were actively transported in mouse hippocampus neurons. We also showed that the tumor cells at the tumor-stroma boundaries demonstrate differential transcriptomic features from the tumor cells in the central regions, caused by their interactions with the stroma/immune cells. Overall, we envision denoising by Sprod to become a key first step to empower SEP technologies for biomedical discoveries and innovations.

INTRODUCTION

Spatially resolved expression profiling (SEP) technologies have recently bloomed, such as 10X Visium, Slide-Seq[1,2], HDST[3], Seq-scope[4], XYZeq[5], *etc*. SEPs provide gene expression information for the whole transcriptome at near- or sub-single cell resolution, in tandem with matching spatial information and, for many techniques, also the matched pathology images stained by Hematoxylin and eosin (H&E) or immunofluorescence (IF). The units of sequencing are called spots or beads by different techniques but are essentially a small cluster of cells (up to a few dozens), or even parts of a single cell for some newer techniques with higher spatial resolution, such as HDST. The 10X Visium technique has been commercialized and popular, while the other newer techniques are anticipated to catch up in popularity soon. These powerful techniques have enabled researchers to localize cell types within the tissues, to characterize spatial expression patterns, to define the local ecosystems of cells, and to resolve the spatiotemporal order of cellular development.

However, as we demonstrate below, such technologies suffer from severe noises in the gene expression measurements. The noises come from random variations due to the shallow nature of sequencing for each spot/bead, similar to regular single-cell RNA-sequencing (scRNA-seq). But

they could also be further complicated by the extra and delicate experimental steps needed to preserve the spatial locations of the gene expression measurements. Thus, data pre-processing steps of SEP data are necessary before any downstream analyses, in order to reduce the noises in the SEP data. The methodologies developed for scRNA-seq drop-outs (loss of expression)[6,7] are likely insufficient for this purpose, as the expressional noises in SEPs are more than just drop-outs. Moreover, such methods only rely on the expression data themselves to correct for drop-outs in measurement, and such "bootstrapping" methodologies are limited in the extent to which the drop-outs can be reliably corrected. In other words, they ignored the spatial and imaging features of the spots/beads provided by SEPs, which can potentially guide and improve the correction of noises in SEPs with useful external information.

In this work, we demonstrated the existence of extensive noises in the SEP data. We developed Sprod, short for **S**patial **Pro**filing **D**e-noising, to reduce such noises, by leveraging the location information of each measurement and the corresponding imaging data, which are available for most of SEP techniques. By testing on a variety of Visium and Slide-seq datasets, we systematically validated Sprod's accuracy and robustness in reducing the noises in SEP data. We also showed its superiority to existing drop-out removal methods of scRNA-seq data. With Sprod-corrected SEP data, we discovered 3-4 times more RNA transcripts that were transported from the soma of the neurons into the dendrites in the mouse hippocampus, compared to the un-corrected SEP data. With Sprod, we better delineated the different transcriptomic features of the tumor cells near the boundaries of the tumor-stroma/immune regions compared with those in the centers of the tumor regions, *via* pseudotime analyses. These differences are, at least partially, driven by tumor cell-stroma/immune cell interactions at their interfaces, according to our results. Overall, careful handling of technical noises in SEP data is a critical first step to the unbiased discovery of new biological knowledge.

RESULTS

Extensive Noises Exist in SEP Data

We demonstrated the existence of extensive noises in SEP data. First, we investigated a 10X Visium ovarian cancer dataset, with matched immunofluorescence (IF) images of CD45/LCA (Leukocyte Common Antigen), keratin, and DNA. **Fig. 1a** presented the CD45 protein expression obtained by IF staining, and the RNA expression level of *PTPRC* (the gene encoding CD45), obtained from both the target panel sequencing and whole transcriptome sequencing by Visium. Strikingly, there is a very poor correlation between CD45 protein IF and *PTPRC* RNA expression (by either target panel or whole transcriptome). When the spots were subset by keeping only those with higher overall sequencing depth, which are potentially of higher quality, the correlation improved by a large extent (**Sup. Fig. 1**). One of the many possible sources of inaccuracies results from drop-outs, as in regular scRNA-seq[6,7]. In **Fig. 1b**, it is clear that severe drop-outs exist in the expression of *PTPRC* (excessive zeros at X=0). We created a gene signature of *PTPRC* RNA expression by including highly correlated genes of *PTPRC*, which

essentially removed drop-outs by *ad hoc* averaging. We found that this *PTPRC* RNA signature's correlation with CD45 protein IF intensity is further improved (**Sup. Fig. 1**), compared with the single gene of *PTPRC*.

Indeed, we observed that drop-outs are a severe issue in both Visium and Slide-seq datasets. In **Fig. 1c**, we showed the percentages of zero counts of all captured genes, as a function of the average expression of each individual gene across all beads/spots, for expression measurements from Visium, Slide-seq, regular scRNA-seq, and regular bulk RNA-seq (from TCGA breast cancer cohort). In all techniques, the percentages of 0 counts increased with lower average gene expression levels, as expected. Visium and Slide-seq both have drop-out levels comparable to regular scRNA-seq. Visium's drop-outs rates are lower than regular scRNA-seq, as each spot of Visium sequencing usually contains a group of cells. Slide-seq has a much higher resolution (very close to single-cell resolution), and a much higher rate of drop-outs than regular scRNA-seq, possibly due to its lower per-cell sequencing depth.

However, the noises from SEP data are more than simply drop-outs. We examined another 10X Visium dataset of human benign reactive lymph nodes. In the normal lymph nodes, follicles and perifollicular "ring-like" structures exist (the darker rings in the H&E stained tissues images of **Fig. 1d** and **Sup. Fig. 2**), which represent reactive germinal centers (GCs) surrounded by well-defined mantle zones[8–10]. Sometimes, another outer layer, called marginal zones, also exists. IgD is a specific marker of the mantle zone development[2]. However, examination of the IgD expression in this Visium dataset shows only a weak correlation with the ring-like mantle zone structures on the H&E stained tissue sides (**Fig. 1d**). In the "good" example, we show that the brighter green color (IgD expression) mostly exists on the ring of the mantle zone (darker purple staining), while in the "bad" example, the green color is mostly uniform and only weakly correlates with the purple staining. More importantly, drop-outs can not completely explain the observed weak correlation, as it appears that IgD expressions are strong in some places (such as the GCs) where they should not be.

The Slide-seq data was also investigated. Slide-seq data do not have the matched imaging data for us to cross-reference as a gold standard. Therefore, we investigated another aspect of the Slide-seq data. We noticed that Slide-seq has a much shallower sequencing depth compared with Visium and regular scRNA-seq. This could result in severe noises in the gene expression measurement, where the measured expression could be either artificially higher or lower than the truth. Here we focused on the problem of inflated high expression (false positive), as the problem of drop-outs (false negative) has already been investigated in our previous example. We showed in **Sup. Fig. 1** that the total sequencing depth of each bead can be used as a surrogate of bead sequencing quality. We hypothesized that the beads with lower sequencing quality are more likely to demonstrate overly inflated expression. We divided all beads into 4 bins, based on the total sequencing depth. Then for each gene, we investigated the probability that the beads would show a higher expression than a given cutoff in each of the bins, among the beads that have the non-zero expression of this gene (**Fig. 1e**). Since the bead sequencing quality is a technical issue,

we expected the beads in different bins to follow similar distributions of gene expression. However, as we showed in **Fig. 1f**, there is a dramatic difference between beads in different quality bins in terms of their gene expression distribution. Furthermore, the bins with lower quality are more likely, for all cutoffs employed (X-axis), to yield expression counts that are larger than the cutoffs (Y-axis). This unexpected distributional difference among bins of different technical qualities could confer an alarming level of bias in gene expression analyses, where highly expressed marker genes can be wrongly identified.

Sprod for De-noising SEP Data Based on Latent Graph Learning

To correct the significant level of noises in the SEP expression data, we developed Sprod (**Fig. 2a**), which corrects the noises in expression data guided by the location and imaging information. Sprod operates in two stages. In the first stage, Sprod leverages the spatial locations of the beads/spots to determine the neighborhoods of the spots/beads to borrow information. Gene expression of neighboring spots are assumed to be more similar to each other than those from distant spots. In the case of drop-outs, for example, the lost expression can be recovered by referring to the neighboring beads/spots. However, it is imperative to consider the cell type heterogeneity of the spots/beads in this neighborhood. Therefore, borrowing information is only restricted to the cells of the same type. For the Visium platform, SEP data are provided with the matched pathology images, from which textures and channel intensities can be derived to inform the cell type heterogeneity. On other hand, for the Slide-seq platform, no matched images are provided. We leverage the fact that the overall transcriptomic profiles of the beads should be sufficient for a simple cell type clustering task, which essentially averages out individual genes' noises in an *ad hoc* manner. Therefore, we use the overall gene expression profiles of the beads to perform unsupervised clustering to detect different types of cells. Then Sprod creates a "pseudo-image", whose "image channels" are essentially the detected cell types and their assignment probabilities. In the second stage of the model, de-noised gene expression is generated by capturing the local information on the manifold of the learned similarity graph, namely by borrowing information from expression data across beads/spots through the graph edges. The technical details of Sprod are specified in the method section and **Sup. File 1**.

We performed simulation analysis to evaluate and showcase the performance of Sprod. We simulated a dataset of 5,000 spots (**Sup. File 2**), similar to the number of spots in a typical Visium dataset. The spots were divided into three cell types, A, B, and C. The three blobs represented cell type A and the two blobs represented cell type B, while cells of type C filled the remaining spaces in between. The details of the simulation were specified in the method section. We applied Sprod to this simulation dataset. We first checked the graph of spatial/image similarity constructed by Sprod and found that the graph generally correctly connected spots of the same cell types that are also in spatial proximity (one example shown in **Fig. 2b**, the graph edges were randomly down-sampled to avoid cluttering). We evaluated the denoising performance using the sum of absolute error (SAE) between the denoised/raw expression and the clean expression (with no simulated noise). The metric for our evaluation is defined as

denoising%=1-SAE(denoised)/SAE(raw). We tried the combinations of all tuning parameters on this dataset (all tested parameters described in the method section). **Fig. 2c** showed that Sprod can decently reduce the noises in the SEP data across a wide range of parameter combinations, though to different extents of denoising efficiencies. Our exploration of these parameter combinations shows that three tuning parameters, R (the radius to define the neighborhood of the spots), K (the dimension of the latent space), and Lambda (a scaling parameter to control clustering sparsity), have a critical contribution to the denoising%, while the other tuning parameters (such as the tSNE-type perplexity for building the similarity graph) have minimum influences. Overall, K=10 works the best. In **Fig. 2d**, we showed a surface plot of the performances of Sprod (denoising%) with respect to the choices of R and Lambda, given K=10, which stably exceeds 80%. We arrived at a set of optimal tuning parameters for this simulation dataset (**Sup. Table 1**), which serves as a reference for the application of Sprod on the real datasets below, with necessary adjustments.

In practice, we acknowledge the simulation data are still different from the real datasets, and real datasets are also different from each other, inevitably calling for tuning of the R, K, and Lambda parameters. We recommend the users to start from the optimal parameter set in our simulation dataset above and use two diagnostic plots that we provided in the Sprod software (**Fig. 2e**, the 10X breast cancer dataset shown as an example) to choose the optimal parameter set. The first diagnostic plot displays the spots/beads and the edges of the similarity graph on the physical X-Y coordinates, which can inform whether the graph has correctly captured the pattern of tissue organization. The other two diagnostic plots display the spots/beads and the similarity graph in the imaging feature space (dimension reduction performed through t-SNE and uMap, only the uMap result is presented in **Fig 2e**). These plots inform whether the detected graph has correctly connected spots/beads that are similar in imaging appearances.

Validation of Sprod Shows Its capability to Reduce the Noises in SEP Data

Sprod was first applied on a 10X Visium ovarian cancer dataset. For the matched images, we included the Keratin and DAPI IF channels and left out the CD45 IF channel for an independent validation. **Fig. 3a** shows the denoised *PTPRC* expression on the slide. The "discrete" appearances of the original *PTPRC* expression (**Fig. 1a**) disappeared and the adjusted *PTPRC* expression demonstrates a more continuously changing pattern, and more importantly, is consistent with the pattern of CD45 IF intensities (**Fig. 1a**). We showed a scatterplot of the CD45 IF intensities and the denoised *PTPRC* expression for all spots, and observed an overall good correlation (**Fig. 3b**, Spearman cor.=0.7 and Pearson cor.=0.42), in contrast to the original expression (**Sup. Fig. 4a**). We also overlaid the differences between CD45 IF and *PTPRC* gene expression for each spot with respect to their physical locations. As is shown in **Sup. Fig. 4b**, Sprod has indeed significantly reduced the deviances between CD45 IF intensities and *PTPRC* gene expression for the majority of spots. We also performed the same analyses with SAVER and scImpute as comparisons[6,7], which are software tools for drop-out removal in regular scRNA-seq data. **Fig. 3c** showed that SAVER and scImpute only achieved modest improvement in terms of

correlation. In addition, another control analysis with Sprod was performed by randomly permuting the image/spatial information. This "scrambled" control resulted in a very low correlation between CD45 IF intensities and the "denoised" *PTPRC* (**Fig. 3c**), confirming that Sprod reduced the noise in SEP data through correctly borrowing information from external image/location information, rather than through merely smoothing of the expression data themselves.

We next investigated a Visium human lymph node dataset. After the correction by Sprod, the adjusted IgD expression demonstrates a spatial pattern more concordant with the H&E stained images (IgD counts shown in **Sup. Fig. 5** and H&E stained image in **Sup. Fig. 2**). We highlighted one mantle zone in **Fig. 3d** (red circle in **Sup. Fig. 5** and the yellow circle in **Sup. Fig. 2**). The Sprod-corrected IgD expression formed a more distinctive ring-like pattern, compared with the original IgD expression (**Fig. 3d**), and is more consistent with the structures of mantle zones. To quantify the improvement of denoising by Sprod, we computed the expression correlations of IgD with several other genes. CD1c and CD20/MS4A1 are also markers of mantle zones and should be positively correlated with IgD[2]. In contrast, CD3 spans the perifollicular/interfollicular T cell regions and should be negatively correlated with IgD[3]. We correlated IgD with CD1c, CD20, and CD3 for the original expression data (Pearson cor.: **Fig. 3e**, Spearman cor.: **Sup. Fig. 6**), and observed positive correlations for CD1c and CD20 as expected, and weak negative correlations for CD3. For the Sprod-corrected expression, CD1c/CD20 showed a much stronger positive correlation with IgD, and CD3 also showed a clearer negative correlation (**Fig. 3e** and **Sup. Fig. 6**). This indicates that the expression data are of higher quality after noise removal. Again, comparative analyses with SAVER and scImpute were performed, and also the "scrambling" control, which validated the superiority of denoising by Sprod in capturing morphology-related gene expression patterns.

Finally, we applied Sprod in the Slide-seq mouse brain dataset used in **Fig. 1**. As in **Fig. 1f**, we calculated whether the over-inflation of highly expressed genes still persists after the Sprod-correction from the beads of lower sequencing quality. Unlike the original Slide-Seq data (**Fig. 1f**), the Sprod-corrected data have almost equal probabilities of observing highly expressed genes across the whole range of cutoffs for the four bins of beads (**Fig. 3f**). We performed the same analysis with SAVER (scImpute fails due to large data size), and observed similar but slightly worse performance, in terms of removing artificial differences between the four bins (**Fig. 3g**). We also calculated the percentage of cells with zero counts (reflective of the level of drop-outs) for each gene. In **Fig. 3h**, we showed that Sprod has significantly reduced the levels of drop-outs in the Slide-Seq data. However, it's important to note that although the apparent false positives and zero counts have been removed, by both Sprod and SAVER, it does not necessarily imply that the "corrected" expression is closer to the true transcriptomic status of the cells, which is further examined below.

Detection of Spatially Variable Genes is More Accurate after Sprod Denoising

We tested whether Sprod denoising improves the downstream analyses of the SEP data. The detection of spatially variable genes (*i.e.* genes whose expression demonstrate certain spatial patterns) is one of the most prevalent analyses for SEP data. The mouse hippocampal Slide-Seq data[1] was examined once again. The hippocampus comprises several regions, including Cornu Ammonis (CA) 1, CA2, CA3 neurons, and dentate gyrus (DG). Stickels *et al* found several clusters of genes enriched in the dendritic region of CA1 by directly comparing beads in the proximal neuropil to the soma of neurons[1]. We tested whether the detection of spatially variable genes after the Sprod correction is more consistent with biological knowledge. Importantly, we note that, due to the higher resolution of Slide-Seq and the nature of the neuronal cells in the mouse hippocampus regions, the Slide-Seq beads in this study captured different parts of the neurons, rather than different individual neurons. We normalized each bead's expression data by per-bead library size, with the understanding that the normalized gene counts reflect the relative enrichment of mRNA transcripts within different physical regions.

In **Fig. 4a**, the Slide-Seq beads that correspond to the soma, proximal neuropil, basal neuropil of the hippocampus neurons are highlighted in different colors, where these regions were identified using the same methods by Stickels *et al*. In **Fig. 4b**, the spatial expression of the *Camk2a* gene is displayed in the order of basil neuropil, soma, and proximal neuropil on the X-axis. **Fig. 4b** shows that there are severe drop-outs in *Camk2a*'s expression (dots concentrated at Y=0), and the beads unnaturally break into two groups by expression of *Camk2a*. The *Camk2a* expression in beads without drop-outs demonstrates a spatial gradient pattern of lower expression in the soma and higher expression in the neuropils. This spatial gradient reflects the fact that *Camk2a* is one of the genes that are actively transported from the soma to the neuropils for local translation[11]. The other group of beads has *Camk2a* expression being strictly 0 and confounds the interpretation of the *Camk2a* expression pattern. To remove the drop-outs, we first used SAVER. However, as shown in **Fig. 4b**, SAVER only minimally recovered the non-zero expression of those spots with drop-outs (still concentrated close to Y=0), and the expression of *Camk2a* remains unnaturally dichotomized. In contrast, Sprod-corrected *Camk2a* expression has nearly completely eliminated the drop-out effects and the beads now demonstrate an overall consistent pattern of high expression of *Camk2a* in the neuropils and lower expression in the soma.

We also examined the expression of another gene, *Hpca*, whose expression is higher in the soma but lower in the neuropils. This gene's expression also demonstrated a severe artificial dichotomization caused by drop-outs in the original un-corrected data (**Fig. 4b**). SAVER is still unable to sufficiently remove the drop-outs. In contrast, in the Sprod-corrected data, drop-outs have been mostly removed and *Hpca*'s expression follows a more continuous and clearer gradient along the X-axis. These results prove that Sprod has indeed addressed drop-outs and corrected noises in the SEP expression data in a biologically meaningful way.

Next, we evaluated the effect of Sprod correction on the detection of spatially variable genes in a genome-wide manner. SpatialDE[12], which is developed specifically for the detection of spatially variable genes, was used to detect the genes with differential expression between the soma and

proximal neuropil regions in both the original data and the Sprod-denoised data (**Fig. 4c** and **Sup. File 2**). In the uncorrected data, SpatialDE yielded 28 genes with differential expression at an FDR (False Discovery Rate)-adjusted P-value cutoff of 0.05, while in the Sprod-corrected data, SpatialDE identified 124 genes, showing that the Sprod correction has greatly increased the sensitivity of differential expression analyses. To validate whether these genes are in consistency with those previously reported, we cross-referenced Tushev *et al*[13] and Ainsley *et al*[14], who identified dendritically localized transcripts *via* microdissection and ribosomal-RNA enrichment, respectively. In **Fig. 4d**, the differentially expressed gene set from Sprod-corrected data exhibits a stronger enrichment with the union of the Tushev *et al* and Ainsley *et al* genes (Hypergeometric P-value: 6.98E-57), in comparison to the raw data (P-value: 2.72E-19). This observation has also been confirmed by employing a simple Wilcoxon rank test to detect the differentially expressed genes (FDR-adjusted P-value <0.05, Fold-change >2). The Wilcoxon test seems to be much more conservative than SpatialDE, but similarly, the Wilcoxon test yielded only 16 genes with differential expression in the uncorrected data, while in the Sprod-corrected data, the Wilcoxon test identified 42 genes. These differentially expressed genes from the corrected data also have a much more significant overlap with Tushev *et al* and Ainsley *et al* (P-value: 1.10E-26), than the raw uncorrected data (P-value: 1.07E-11).

We further evaluated the pathways in which these spatially variable genes are enriched. Examination of enriched GO pathways in the spatially variable genes from SpatialDE shows that the Sprod-adjusted expression data lead to the discovery of more genes that are indicative of dendritic functions (*e.g.* “dendrite”), or synapse functions (*e.g.* "neurotransmitter uptake") (**Fig. 4e**), consistent with the enrichment of these genes in the dendritic areas of the neurons. The dendritically enriched mRNAs are transported from the soma bodies of the neurons, which is a process dependent on RNA binding proteins (RBPs)[15,16]. The CLIP-Seq technology is an experimental technique for profiling RNA-RBP interactions[17–19]. We downloaded CLIP-Seq datasets of several RBPs with known roles in dendritic mRNA transport or cytoplasmic stability from the ENCODE database[20], which includes TDP-43[21], SMNDC1[22], IGF2BP1[23], RBFOX[24], FMR1[25], CPEB[26], and PUM2[27]. We calculated the percentage of the dendritically enriched genes that are binding targets of each RBP. In **Fig. 4f**, we show that the Sprod-corrected expression data yield dendritically enriched genes that are much more likely to be known targets of these RBPs according to CLIP-Seq (T test P value= 0.0037).

In the dendritically-enriched genes, we also discovered over-representation of astrocyte terms (“astrocyte projection”), which conforms to the known existence of astrocytes and other glial cells around the neuropil regions of neurons[28–30]. We performed GO analyses with specialized glial cell signatures from PangloDB[31], and again confirmed stronger enrichment of glial cell signatures in the dendritic regions, only in the Sprod-corrected data (**Sup. Fig. 7a**). Finally, this analysis on pathways from DisGeNET[32] shows that the dendritically enriched genes from Sprod are also more strongly associated with neurological diseases (**Sup. Fig. 7b**). Overall, our

analyses prove that the Sprod-correction of the SEP data is critical for the accurate detection and interpretation of spatially variable genes and pathways.

Sprod Facilitates Inference of More Accurate Spatial Pseudo-time and Cellular Communications

We further tested whether Sprod's denoising improved the accuracy of the pseudotime inference. Pseudotime aims to infer the relative ordering of cells along a lineage based on the cells' gene expression[33,34]. The breast cancer Visium dataset previously used in **Fig. 2e** was examined. We performed expression clustering of the spots and defined four tumor cell regions (**Fig. 5a**), with the remaining areas filled with stromal/immune cells. Given each Visium sequencing spot usually contains more than one cell, in order to ensure the inferred pseudotimes are more specific to the tumor cells, we removed the "contamination" from stromal/immune cells in the tumor region spots. To do this, we calculated a stromal/immune "contamination score" for each spot in the tumor region using a combined stroma/immune gene signature from Wang *et al*[48] and Wang *et al*[35]. We removed spots with high "contamination scores", and for the remaining spots, we also removed genes in this signature. Next, we predicted the pseudotimes of the tumor cells for each of the four regions, using SlingShot[33] and Monocle3[34]. As shown in **Fig. 5b** and **Sup. Fig. 8**, the pseudotimes inferred from the Sprod-corrected data appear to have a smoother spatial gradient, and the centers of the tumor regions and the boundaries of the tumor regions demonstrate a more consistent differential pattern in pseudotimes.

To objectively prove that the pseudotimes predicted from the Sprod-corrected data are more biologically meaningful, we examined the genes that demonstrate differential gene expression along the pseudotime trajectories (two example genes, *CALML5*[36], and *CAPRIN1*[37], shown in **Fig. 5c**). The Gene Set Enrichment Analysis (GSEA)[38] was performed using the KEGG, GO, and cancer Hallmark gene sets (https://www.gsea-msigdb.org/gsea/msigdb/) (**Fig. 5d**) to detect the enriched pathways in the genes that demonstrate significant variation along pseudotime trajectories. We analyzed the concordance between the significant pathways detected by GSEA, among the four regions, as we reasoned that the pattern of tumor differentiation and phenotypic evolution should be similar within this small local region. As illustrated in **Fig. 5e** and **Sup. File 2**, the differentially expressed genes from each of the four regions are enriched in many more shared significant pathways for the Sprod-corrected data, than for the raw expression data. This holds true for both the Monocle3- and the Slingshot-inferred pseudotimes.

Next, the top common pathways between the four tumor regions from the Sprod-corrected data (ordered by the absolute enrichment score of GSEA) but not from the raw expression data were inspected more closely (**Fig. 5f**). Among those pathways, we identified the pathway of Hallmark_Epithelial_Mesenchymal_Transitionis, which is highly relevant for tumorigenesis[39]. More interestingly, many of these pathways are highly relevant for cell-to-cell communications, such as Collagen_Containing_Extracellular_Matrix and Integrin_Mediated_Signalling_Pathway.

Given this and the observation that the pseudotimes reveal a difference between the tumor boundaries and the tumor centers (**Fig. 5b**), CellChat was used to examine whether the interactions between the tumor cells and immune/stromal cells at the tumor boundaries could be driving the differential transcriptomic status between the center and the boundaries of the tumor regions and if Sprod facilitates identifying such interactions. We further divided (**Fig. 6a**) the tumor and non-tumor regions into 4 sections: region A (tumor cells not adjacent to stromal/immune (SI) cells), region B (tumor cells adjacent to SI cells), region C (SI cells adjacent to tumor cells), and region D (SI cells not adjacent to tumor cells). The classification process was described in the method section. Cellchat was deployed to infer cellular communications among the cells in these four regions. We hypothesize that long-distance tumor-SI cellular communications (B->D or C->A) should be substantially less active than the cellular communications at the boundaries of the tumor regions (B->C or C->B). Indeed, in the Sprod-corrected data, the number of significant ligand-receptor pairs inferred by CellChat is higher for the B->C communication, than B->D, and higher for C->B than C->A (**Fig. 6b**). In contrast, the raw expression data yields a significantly reduced number of interacting pathways overall and less clarity in the trend of the difference for B->C *vs.* B->D, and for C->B *vs.* C->A.

In particular, we noticed that the pathway of *PD-L1* and *PD-1* in the significant interacting pathways detected from the denoised expression data, but not in the original expression data. The role of the antagonizing interactions between *PD-L1* and *PD-1* in breast cancers has been well established[40–42]. In the denoised data, this pathway was observed to be significant in the direction of A->C and B->C, but with stronger confidence in B->C (CellChat Probability Strength=2.21E-8, P value<1E-10) than in A->C (Probability=1.72E-8, P value<1E-10). We visualized the expression of the ligand *PD-L1*/*CD274* in the tumor cells and the receptor *PD1*/*PDCD1* in the stroma/immune cells in **Fig. 6c**, for both the raw and denoised data. Overall, the denoised data seemed to demonstrate a more obvious co-expression of *PD-L1* and *PD1* around the interfaces of the tumor-stroma/immune regions, compared with the raw expression. However, it is also evident that the expression of *PD-L1*/*PD-1* is not uniformly high along the interfaces but rather possesses a local enrichment pattern. To objectively quantify the co-expression of *PD-L1*/*PD1* in the B/C regions, we defined close pairs of spots, with one spot in the B region and the other in the C region (**Fig. 6d**). In the pairs from the Sprod denoised data, the expression of tumor region *PD-L1* becomes significantly higher when neighboring stroma/immune regions demonstrate higher PD-1 expression (**Fig. 6e**). But when *PD-L1* becomes higher in the tumor cell regions, the expression of *PD-1* is only minimally higher in the stroma/immune regions. This uni-directional observation is intriguing but also reasonable, as we should anticipate tumor cells to up-regulate *PD-L1* expression in response to the cytotoxic pressure from $PD1^+$ T cells, but not the other way around. In other words, this analysis reveals the causal relationships between the interaction of the PD-L1/PD-1 pathway, in alignment with existing biological knowledge of how this pathway functions. With the raw expression data, however, we cannot make this observation. Overall, our above analyses suggest that Sprod also enables more accurate inference of cell-to-cell communications using SEP data.

DISCUSSION

We developed Sprod to reduce noises in SEP data. The existence of extensive noises have been demonstrated in data from different SEP technologies, which will impact downstream analysis severely and result in significant biases and misleading conclusions. Sprod took a novel approach of leveraging the physical location and matched imaging data of SEP to reduce such noises, rendering analysis and interpretation of SEP data much more robust and accurate. Technically, the location/imaging similarity graph is obtained by an innovative sparse graph construction method based on a probabilistic density-based approach. This modeling strategy can tolerate high-dimensional data noise, preserve the pairwise metric, and integrate the imaging and positional features in a unified framework[43]. This allows Sprod to denoise the expression of spots/beads in a dynamic neighborhood structure defined by imaging and location information. We systematically validated Sprod and its performance was demonstrated to be superior to algorithms designed solely for the removal of drop-outs in scRNA-seq data. Sprod is computationally efficient and user-friendly, and it is capable of readily handling data generated from a wide variety of SEP technologies and being seamlessly integrated into bioinformatics pipelines for spatial transcriptomic analyses.

Drop-outs, as in regular scRNA-seq data, are one source of the inaccuracies in the SEP data. Concerns have been raised in the field that drop-out correction methods for scRNA-seq data may introduce over-smoothing and thus erroneous signals in the scRNA-seq data[44,45]. Since regular scRNA-seq data only provide the gene expression of individual cells, these drop-out removal methods inevitably have to only rely on the cells with similar expressions to correct the drop-outs in each cell of interest. This essentially bootstraps the expression data, which leads to the over-smoothing among cells. For SEP data, Visium, Slide-seq, or similar technologies provide the spatial locations and often imaging features associated with the spots/beads, which serve as external information for reliable denoising and alleviate the concerns of over-smoothing. Besides, the latent graph building process in Sprod enforces a neighborhood constraint (R) in denoising, further preventing the potential problem of over-smoothing. Moreover, it is worth noting that Sprod can, in principle, address other potential sources of noise in SEP data (such as the false positive shown above).

Sprod is a universal application that can be used with all SEP technologies. However, there are differences between different SEP technologies, so considerations should be given to how to best apply Sprod for each technology. This study focused on Slide-Seq and 10X Visium data, between which one distinction is the level of resolution. Each bead in Slide-seq likely profiles a 10um by 10um region while each spot in Visium profiles a 55um by 55um region. The sequenced cells and the multicellular structures of the tissues, on the other hand, are also of various size scales depending on the tissue and cell types. These factors together determine the degree of spatial dependency between spots/beads. The R parameter (physical scope to examine for spatial dependency) should be set larger for tissue types and SEP technologies that impart a weaker

spatial dependency, to introduce fewer constraints on spatial dependency in the graph building process. Our practical experiences using Sprod on datasets have indeed confirmed this intuition.

The current project focuses on RNA expression. Future spatial sequencing technologies are likely to be able to incorporate more modalities in the sequencing, such as protein expression, copy number variations, methylation, *etc*. When the new modalities are available, the Sprod model can be easily extended to them. 3-dimensional sequencing technologies are also emerging, such as STARmap[46]. Currently, STARmap is only available for a small number of genes. However, whole transcriptome 3-dimensional sequencing will most certainly become available in the future. With little modifications, the graph building model employed by the current Sprod model can be easily extended to consider spatial dependency in the 3-dimensional space.

Despite the great advancements of SEP technologies, the field must pay close attention to the quality issues of SEP data, which are more challenging than regular scRNA-seq. Though the current study considered Visium and Slide-seq, Sprod is also applicable to newer techniques such as HDST[3] and Seq-scope[4]. It is important to note that HDST and Seq-scope (and, most likely, future SEP techniques as well) achieve much higher resolution than Visium and Slide-seq. Noises in the SEP data are likely to be more prevalent, necessitating even more cautious preprocessing before drawing any conclusions. Overall, we envision Sprod, developed specifically for this purpose, to become a key first step to empower SEP technologies for biomedical discoveries and innovations.

MATERIALS AND METHODS

Overview of Sprod

Sprod takes the positional and image features as input, calculates a latent graph from them, and finally uses the latent graph to smooth out noise in the original SEP expression matrix. In our study, the SEP expression matrices were transformed to the CPM scales and log-transformed. For future users, pre-processing of expression matrices (use of either counts/CPM/RPKM/TPM, normalization, batch correction, log transformation) is assumed to be the responsibility of the users and conducted before Sprod analysis, as necessary. The Sprod model will only use the expression matrices as is. The final output includes the denoised expression matrix, and the graph of spatial/location similarity of the spots/beads. The core Sprod model was implemented in the R language, and image manipulation and input/output interfaces were implemented in the Python language. The detailed model description was provided in **Sup. File 1**.

Image processing and feature extraction

The Python ‘skimage’ package (version 0.18) was used for all image processing and feature extraction operations. Immunofluorescence (IF) and H&E images were normalized into 8-bit format first. For H&E images, additional steps including stain separation (skimage.colors.separate_stains) and adaptive histogram equalization

(skimage.exposure.equalize_adapthist) were used to ensure the channels in the normalized image are minimally correlated.

Sprod relies on two types of image features based on the input data type. For datasets with matching IF or H&E images, two sets of features were extracted. For both sets of features, the image region where the features were extracted from can be the spot itself ('spot') or a box covering both the spot and its neighboring regions ('block'). We calculated the 20th, 30th, 40th, 50th, 60th, 70th, and 80th percentiles of intensity values among all the pixels in each sequencing spot/block in each channel. These values were used as the intensity features for Sprod. We also calculated six sets of Haralick's texture features[47], including contrast, dissimilarity, homogeneity, ASM, energy, and correlation, for each sequencing spot/block in each channel. Specifically, the 'skimage.feature.greycomatrix' were used to extract the texture features with 'offset = [1]' and 'angles = [0, π/4, π/2, 3π/4]'. We enable the option of choosing block- *vs.* spot-level and intensity *vs.* texture image features to the user. For the datasets used in our study, our choices were shown in **Sup. Table 1**.

For datasets without matching images, such as those generated by the Slide-Seq, pseudo-images were created. These features were generated in the following steps. Highly variable genes were selected using the 'scanpy.preprocessing._highly_variable_gene' method (https://github.com/theislab/scanpy/blob/f7279f6342f1e4a340bae2a8d345c1c43b2097bb/scanpy/preprocessing/_highly_variable_genes.py). Umap transformation was applied on the normalized dataset with only the highly variable genes, and beads in the transformed data were partitioned into clusters using the Dirichlet process clustering method implemented in the 'dirichletprocess' package in R. The pseudo-image features were then assigned by the possibility of each bead belonging to each cluster.

Scaling up in large datasets

Sprod is fast for datasets of thousands of spots/beads. However, for large datasets with tens of thousands of spots, special operations must be performed so that Sprod can run smoothly. In this work, we employed a splitting-stitching scheme to facilitate large dataset processing. Each Slide-Seq dataset was randomly (not dependent on spatial location) divided into n (10 by default) equal-sized subsets, and this process was repeated b (10 by default) times. Sprod denoising was performed on each of the n x b subsets and the denoised results were concatenated. Each spot was exactly denoised b times, and the concatenated denoised data from the n sampling batches were averaged so that the randomness resulting from the sub-sampling was averaged out.

Generation of simulation data

In the simulation analyses (**Fig. 2**), three matrices were simulated as the input data: the "Expression matrix" (E), "Spots_metadata" (C), and "Image features" (IF). E had 100 genes and 5,000 spots and comprised 282 spots of cell type A (3 clusters), 232 spots of cell type B (2 clusters), and 4,486 spots of cell type C. $C$is the X/Y coordinates of the spots. A Dirichlet

Process clustering of the expression matrix was performed on the expression data to detect the cell types, and the cell type labels and their assignment probabilities formed the IFmatrix, in the same way as Sprod created the "pseudo-images".

In particular, the expression matrix, E, was a sum of three parts: (1) Expressional variation of cell types, E_1. It was generated from a multivariate normal distribution with three different means corresponding to the three cell types, and the covariance matrix was chosen such that three clusters of cells are visually discernible in the first two components of a Principal Component Analysis transformation. (2) Expressional variation of spatial dependency, E_2. Here, we reasoned that spots of the same cell types that were nearby should have more similar expressions. We simulated $E_1$from a multivariate normal distribution with $\mu = 0$ and the covariance matrix calculated based on the exponential of the negative of the Euclidean distances between all spots. The covariance among different cell types is set to zero so spatial dependency only happens for spots of the same cell types. $E_2$was scaled so that its variation was comparable to E_1. (3) White noises in expression, N. $N$was a matrix of white noises generated from independent normal distributions with a mean of 0 and equal variance, which is controlled at several different levels (**Fig. 2c**). The summed expression matrix is then $E = E_1 + E_2 + N$. Finally, the E matrix was transformed by an exponential function and scaled such that its distribution mimics the distribution of typical SEP data.

The details of our simulation, especially all the numeric settings, can be found in our simulation script made available at: https://github.com/yunguan-wang/SPROD.

Hyperparameter optimization

We employed grid search for determining an optimal set of tuning parameters for the simulated dataset. We evaluated combinations of these possible values: R (0.04-0.24, step size = 0.01), K (3-10, step size = 1), U (250, 500, and 1000), Lambda (0.1-1, step size = 0.1, plus 5 and 20), and L_E (0.3125, 0.625, 1.25, and 2.5). R is the radius to define the neighborhood of the spots. K is the dimension of the latent space. Lambda is a scaling parameter to control clustering information. U is the perplexity of the tSNE-like distance function of the input image data. L_E is a penalty parameter adjusting for the relative weight between the original expression matrix and the information from the spots in the neighborhood on the graph. In the real data applications, the parameters were selected by referring to the best parameter set according to the simulation dataset, which was determined by grid search, and are adjusted based on the diagnostic plots (**Fig. 2e**). The parameters used in all datasets involved in this work are listed in **Sup. Table 2**.

Defining tumor and immune/stroma spots/regions that are close neighbors (Fig. 6)

In the breast cancer 10X Visium dataset, the spots were clustered based on gene expression, and manually examined and merged to the tumor and stroma/immune regions. Four tumor regions were defined (left, mid, mid-right, top-right) based on the cluster and spatial information. The tumor spots and their gene expression were cleansed for stroma/immune cell contamination. Specifically, a stromal gene module and an immune gene module were defined by the union of the gene signatures for stromal and immune cells defined by Wang *et al*[48] and Wang *et al*[35]. Each spot was given a score for the stromal cells and a score for the immune cells, based on the average normalized expression levels of the genes in each module. Then, the tumor region spots with the top 5% of the module scores were filtered out. For the remaining tumor region spots, these immune- and stromal-related genes were also filtered out.

In **Fig. 6a**, we further split the tumor regions into two subsets: tumor regions not close to immune/stromal cells (A) or close to immune/stromal cells (B). For each spot in the tumor region, we will draw a circle based on a radius of twice the size of the spacing between two neighboring spots in the Visium spot lattice. We will count the number of stroma/immune region spots in this circle. This spot will be classified as "B" if its closest neighboring spots in the stroma/immune regions are more than a given cutoff. The cutoff is 15 for the left region, 25 for the mid region, 20 for the top-right region, and 15 for the mid-right region. This cutoff is slightly adjusted to ensure a smooth appearance of the classified "A" and "B" tumor regions. The stroma/immune region spots were classified similarly into "C" (close to tumor region) and "D" (not close to tumor region).

In **Fig. 6d**, we examined all pairwise connections between the type B spots and type C spots and calculated their distances. When the distance between a type B spot and a type C spot is smaller than three times the minimum of all distances, the two spots were counted as a close pair.

Statistical analyses

All computations are conducted in the R or Python programming languages. Umap and tSNE were conducted by the umap v0.2.7.0 or Rtsne v0.15 R packages. Drop-out removals were conducted by SAVER v1.1.2 and scimpute v0.0.9. For all boxplots appearing in this study, box boundaries represent interquartile ranges, whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range, and the line in the middle of the box represents the median. All figures were made using the Python 'matplotlib' and 'seaborn' packages or the R 'ggplot2' package. In **Fig. 4**, functional enrichment analysis using hypergeometric tests was performed using the 'gseapy' package in Python. In **Fig. 5**, the immune/stroma-contaminating spots/genes were cleansed as described above. For the remaining tumor region spots, pseudotimes were inferred using the Slingshot v1.8.0 or Monocle3 R packages. To be comparable, the roots, which were assumed to be the cells with the "earliest" developmental stages by Monocle3 were chosen to be also the spots with the earliest pseudotimes in Slingshot, in each of the tumor regions. It's important to note that, at least in the case of linear lineage which is most common, the pseudotime by itself only determines which cells appear in similar

developmental stages but the cells with higher pseudotimes could actually emerge earlier or later on in “real-time”. The differential expression genes along pseudotime trajectories were detected by the ‘pseudotimeDE v0.9.0’ package. Pre-rank GSEA analyses were conducted by the ‘clusterprofiler v3.18.1’ package. For GSEA, the 9,000 top most variable genes were extracted from each tumor region and analyzed, to include as many genes as possible, but also to avoid including too many lowly variable genes that introduce ties in differential expression comparisons and thus difficulties in the clusterprofiler computation. The Visium tumor region spots were divided into "early" and "late" by the median of their pseudotimes. The “ranks” of the GSEA analysis for the genes were calculated as the log fold change of gene expression between the "early" and "late" groups. In **Fig. 6**, the cell-to-cell communications were inferred by the CellChat v1.1.3 package.

Data availability

The Visium datasets are obtained from the public 10X resources/datasets website: https://www.10xgenomics.com/resources/datasets. The IDs of the datasets are: human-lymph-node-1-standard-1-1-0, Human-ovarian-cancer-whole-transcriptome-analysis-stains-dapi-anti-pan-ck-anti-cd-45-1-standard-1-2-0, human-ovarian-cancer-targeted-pan-cancer-panel-stains-dapi-anti-pan-ck-anti-cd-45-1-standard-1-2-0, and human-breast-cancer-block-a-section-1-1-standard-1-1-0. The ID of the regular 10X scRNA-seq dataset used in **Fig. 1c** is 10-k-peripheral-blood-mononuclear-cells-pbm-cs-from-a-healthy-donor-single-indexed-4.0.0. The Slide-Seq data are available from the publicly archived data by Stickels et *al*[1]. Specifically, we used the Puck_200115_08 data from https://singlecell.broadinstitute.org/single_cell/study/SCP815/highly-sensitive-spatial-transcriptomics-at-near-cellular-resolution-with-slide-seqv2.

Code availability

The Sprod software is available at: https://github.com/yunguan-wang/SPROD

FIGURE AND TABLE LEGENDS

Fig. 1 Extensive noise in spatial expression profiling data. (a) The spatial expression levels of CD45 (IF), *PTPRC* (targeted panel sequencing) and *PTPRC* (whole transcriptomic sequencing) of the 10X Visium Ovarian Cancer dataset. (b) Severe drop-outs in the *PTPRC* gene expression for both the targeted sequencing and whole transcriptomic sequencing. X-axis shows the *PTPRC* RNA expression level (unit=log2(CPM+1)). (c) Severe drop-outs in scRNA-seq, Slide-Seq, Visium and bulk RNA-Seq. The X-axis is the average RNA expression levels of each gene profiled by each technique/dataset, and the Y axis shows the percentages of counts of exactly 0 for each gene. (d) Overlaying IgD gene expression on the H&E stained image in the 10X Visium human lymph node dataset. Three examples of good, intermediate, and poor concordance

between IgD and mantle zone structures were shown. (e) A cartoon describing the schema of the calculation of Adjusted Exp.%. (f) The Adjusted Exp.% of beads in the four bins of different sequencing qualities, in the mouse Slide-Seq dataset, and with the cutoffs to define the Adjusted Exp.% shown on the X axis.

Fig. 2 Sprod for de-noising of spatial expression profiling data. (a) A cartoon describing the schema of the Sprod model, from data preparation, graph building, to expressional de-noising. (b) The simulated dataset. The figure shows the location of the simulated spots, and the cell types they are assigned to. The light gray bars show the graph built by Sprod. A bar connects two dots as long as the graph connects two spots, regardless of edge weights. However, the dots have been subsampled to avoid over-cluttering of the presentation. (c) The denoising% of all parameter combinations tested. The X axis shows all parameter combinations, ordered from lower to higher denoising%. (d) Visualizing the denoising% with respect to specific choices of parameters. K=10. The X and Y axes show all the choices tried for each of R and Lambda. The Z axis shows the denoising%, which is the average of the denoising% over all the other parameters' choices. Noise level=0.5. (e) An example of the two diagnostic plots generated by Sprod from the 10X breast cancer Visium dataset. Left, the spots and edges of the detected graph on the X-Y coordinates. Right, the spots and edges of the detected graph in the Umap space of the image features.

Fig. 3 Validation of Sprod on real Visium and Slide-Seq datasets. (a) The Sprod-corrected *PTPRC* expression. Darker colors refer to higher expression. (b) Scatterplots showing the correlation between CD45 IF and *PTPRC* gene expression, corrected by Sprod. (c) The Spearman and Pearson correlations between CD45 IF and *PTPRC* expression, from the original expression data, Sprod-corrected expression data, expression data with drop-out removal performed by SAVER and scImpute, and the Sprod "scrambling" control. (d) The spatial IgD expression of the mantle zone marked in **Sup. Fig. 2** and **5**, for both the original Slide-Seq expression data and the Sprod-adjusted expression data. (e) Pearson correlations between IgD and CD3/CD20/CD1c, for each analysis group. (f,g) The Adjusted Exp.% of beads in the four bins of different sequencing qualities, in the Sprod-corrected expression data (f) and SAVER-corrected expression data (g), with the cutoffs to define the Adjusted Exp.% shown on the X axis. (h) The level of drop-outs in the expression data of each analysis group. Each gene is grouped into one of 8 bins according to their average expression across all spots. B1 has the lowest expression and B8 has the highest expression. For each gene, the percentages of spots with zero counts in that gene is calculated and shown on the Y axis.

Fig. 4 Detection of spatially differentially expressed genes is more accurate after de-noising. (a) The Slide-Seq beads that were defined to be in the basal neuropils, soma, and proximal neuropils of the CA1 region, following the definition of Stickels *et al*. (b) The expression of *Camk2a* and *Hpca* in the CA1 region Slide-Seq beads, ordered along the soma-proximal axis. The X axis shows the location of the beads with respect to the soma. Results for the raw expression matrix, SAVER-adjusted matrix, and Sprod-adjusted matrix are shown. (c) Barplots of the expression of

genes with significant spatial variation between neuropils and the soma, detected by the SpatialDE test. The result for the raw expression matrix is shown as an example. (d) Venn diagrams showing the overlap of the genes with differential expression detected from the raw expression data or the Sprod-adjusted data, with the genes that show dendritic enrichment from Tushev *et al* and Asinley *et al*. (e) Enrichment of GO pathways in the genes with stronger expression in the proximal neuropil regions, detected from the raw expression data and the Sprod-corrected data. (f) Dendritically enriched genes detected from the Sprod-corrected expression data are more likely to be RBP binding targets than genes detected from the raw expression data. Left: the numbers of dendritically enriched genes that are RBP targets (any RBP) or are not RBP targets. Right: the percentages of dendritically enriched genes that are binding targets of each RBP.

Fig. 5 Sprod facilitates inference of more informative spatial pseudo-time in SEP data. (a) The four tumor regions (blue, green, orange, and red) that were extracted for pseudotime analyses, according to expressional clustering and concordance with the H&E stained slide. (b) The pseudo-times that were inferred by Slingshot for each of the four regions, respectively. Results for both the raw expression matrix (left) and the Sprod-corrected matrix (right) were shown. (c) An example gene, *CALML5*, with significant differential gene expression along the pseudotime trajectory, inferred by pseudotimeDE. Results for both the raw expression and the Sprod-corrected matrices, as well as for Slingshot-inferred pseudotimes and Monocle3-inferred pseudotimes were shown. (d) GSEA enrichment plots of the genes that demonstrate differential expression in the "early" and "late" (pseudotime) groups of spots in the left tumor region. (e) Boxplots counting the number of pathways that showed an absolute enrichment score >1.5 in the GSEA analyses in at least 3 out of all 4 tumor regions. (f) The pathways inferred from the Sprod-corrected data that showed an absolute enrichment score >1.5 in the GSEA analyses in at least 3 out of all 4 tumor regions, for both the Monocle-inferred and Slingshot-inferred pseudotimes. The enriched pathways from the raw expression data from the same analyses were excluded from this list.

Fig. 6 Inference of cell-to-cell communication is more accurate with Sprod-corrected expression data. (a) The tumor and stroma/immune regions were both split into sub-regions that are either close or not close to the tumor-stroma/immune boundaries. These sub-regions were named as A, B, C, and D. (b) The numbers of CellChat-inferred significantly interacting pathways, in the raw and Sprod-corrected expression matrices. (c) The expression of *PD1* in the stroma/immune regions and *PD-L1* in the tumor regions, for the raw and Sprod-corrected expression matrices, respectively. (d) The close pairs of spots with one spot in the tumor region and another one in the tumor/stroma region. (e) The expression of *PD1* in the stroma/immune-side spots in the pairs of spots from (d), dichotomized by the expression of *PD-L1* in the corresponding tumor-side spots, and *vice versa*. Dichotomization was performed on the 75% percentile of *PD-L1* or *PD-1*.

Sup. Fig. 1 Correlation between *PTPRC* RNA expression (targeted) and CD45 protein expression (IF). Red arrows mean that the results are limited to spots of higher quality. Yellow

arrows mean that the single gene of *PTPRC* is replaced by a signature of *PTPRC* by including correlated genes.

Sup. Fig. 2 Overlaying the un-corrected IgD gene expression on the H&E stained image in the 10X Visium human lymph node dataset. The whole slide is shown, with the three examples of **Fig. 1d** picked from this slide. The yellow circle marks the mantle zone to be highlighted in **Fig. 3d**.

Sup. Fig. 3 Gene expression clustering of the beads in the mouse brain Slide-Seq dataset reflects the multi-cellular structures of mouse brain hippocampus.

Sup. Fig. 4 Correction of noise in the Ovarian Cancer Visium dataset. (a) Scatterplots showing the correlation between CD45 IF and the original *PTPRC* gene expression. (b) Deviances between CD45 IF intensities and the expression levels of *PTPRC* (left: original, right: denoised). CD45 IF intensities and *PTPRC* expression values were normalized and distributionally warped to the same scale so they can be directly compared. The differences between CD45 IF and *PTPRC* on each spot are denoted by color. Red refers to small differences and green refers to larger differences.

Sup. Fig. 5 Spatial IgD expression of the raw Visium data (left) and the Sprod-adjusted data (right). The red circles mark the mantle zone to be highlighted in **Fig. 3d**.

Sup. Fig. 6 Spearman correlations between IgD and CD3/CD20/CD1c for the human lymph node Visium dataset. Results are shown for the original expression data, SAVER/scImpute-corrected data, the Sprod-corrected data, and the Sprod-corrected data with image/location information scrambled.

Sup. Fig. 7 Enrichment of gene signatures and pathways in the genes with stronger expression in the proximal neuropil regions, detected from the raw expression data and the Sprod-corrected data. Left: PangloDB pathways; right: DisGeNET pathways.

Sup. Fig. 8 The pseudo-times that were inferred by Monocle3 for each of the four regions, respectively. Results for both the raw expression matrix (left) and the Sprod-corrected matrix (right) were shown.

Sup. Fig. 9 An example gene, *CAPRIN1*, with significant differential gene expression along the pseudotime trajectory, inferred by pseudotimeDE. Results for both the raw expression matrix and the Sprod-corrected matrix, as well as for Slingshot-inferred and Monocle3-inferred pseudotimes were shown.

Sup. File 1 Mathematical details of the Sprod model.

Sup. File 2 Additional analyses involving Sprod.

Sup. Table 1 Tuning parameters used in Sprod for each dataset

Sup. Table 2 Dendritically enriched genes in the mouse brain that are identified *via* different methods

FUNDING

This study was supported by the National Institutes of Health (NIH) [CCSG 5P30CA142543/TW, GX, YX, 1R01CA258584/TW, U01AI156189/TW, YX, R01DE030656/GX, R01GM141519/GX, R01GM140012/GX, U01CA249245/GX, R35GM136375/YX, 2P50CA070907/TW, YX, GX], and Cancer Prevention Research Institute of Texas [CPRIT RP190208/TW, RP190107/GX].

ACKNOWLEDGMENTS

We acknowledge the ENCODE Consortium and the ENCODE production laboratories that generated the eCLIP datasets used in our study. We acknowledge Dr. Jane Johnson for providing input on the interpretation of the mouse Slide-Seq data.

AUTHOR CONTRIBUTIONS

Y.W. and B.S. implemented the software and contributed bioinformatics analyses. L.W. and T.W. designed the model. M.C. provided input on the interpretation of the pathology analyses. Y.X., S.W., and G.X. provided input on the analyses and the writing. T.W. supervised the whole study.

COMPETING INTERESTS

The authors declare no competing interests

Bibliography

1. **Stickels, R. R. *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* 39, 313–319 (2021).**
2. **Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* 363, 1463–1467 (2019).**
3. **Vickovic, S. *et al.* High-definition spatial transcriptomics for in situ tissue profiling. *Nat. Methods* 16, 987–990 (2019).**
4. **Cho, C.-S. *et al.* Microscopic examination of spatial transcriptome using Seq-Scope. *Cell* 184, 3559-3572.e22 (2021).**

5. **Lee, Y. *et al.* XYZeq: Spatially resolved single-cell RNA sequencing reveals expression heterogeneity in the tumor microenvironment. *Sci. Adv.* 7, (2021).**

6. **Li, W. V. & Li, J. J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* 9, 997 (2018).**

7. **Huang, M. *et al.* SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods* 15, 539–542 (2018).**

8. **Nakagawa, T., Yamada, M. & Suzuki, Y. 18F-FDG uptake in reactive neck lymph nodes of oral cancer: relationship to lymphoid follicles. *J. Nucl. Med.* 49, 1053–1059 (2008).**

9. **Weller, S. *et al.* Human blood IgM "memory" B cells are circulating splenic marginal zone B cells harboring a prediversified immunoglobulin repertoire. *Blood* 104, 3647–3654 (2004).**

10. **Agbay, R. L. M. C. *et al.* Characteristics and clinical implications of reactive germinal centers in the bone marrow. *Hum. Pathol.* 68, 7–21 (2017).**

11. **Mayford, M., Baranes, D., Podsypanina, K. & Kandel, E. R. The 3'-untranslated region of CaMKII alpha is a cis-acting signal for the localization and translation of mRNA in dendrites. *Proc Natl Acad Sci USA* 93, 13250–13255 (1996).**

12. **Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nat. Methods* 15, 343–346 (2018).**

13. **Tushev, G. *et al.* Alternative 3' UTRs Modify the Localization, Regulatory Potential, Stability, and Plasticity of mRNAs in Neuronal Compartments. *Neuron* 98, 495-511.e6 (2018).**

14. **Ainsley, J. A., Drane, L., Jacobs, J., Kittelberger, K. A. & Reijmers, L. G. Functionally**

diverse dendritic mRNAs rapidly associate with ribosomes following a novel experience. *Nat. Commun.* 5, 4510 (2014).

15. **Doyle, M. & Kiebler, M. A. Mechanisms of dendritic mRNA transport and its role in synaptic tagging. *EMBO J.* 30, 3540–3552 (2011).**
16. **Dalla Costa, I. *et al.* The functional organization of axonal mRNA transport and translation. *Nat. Rev. Neurosci.* 22, 77–91 (2021).**
17. **Wang, T., Xie, Y. & Xiao, G. dCLIP: a computational approach for comparative CLIP-seq analyses. *Genome Biol.* 15, R11 (2014).**
18. **Wang, T. *et al.* Design and bioinformatics analysis of genome-wide CLIP experiments. *Nucleic Acids Res.* 43, 5263–5274 (2015).**
19. **Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* 13, 508–514 (2016).**
20. **Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* 46, D794–D801 (2018).**
21. **Chu, J.-F., Majumder, P., Chatterjee, B., Huang, S.-L. & Shen, C.-K. J. TDP-43 Regulates Coupled Dendritic mRNA Transport-Translation Processes in Co-operation with FMRP and Staufen1. *Cell Rep.* 29, 3118-3133.e6 (2019).**
22. **Thelen, M. P. & Kye, M. J. The Role of RNA Binding Proteins for Local mRNA Translation: Implications in Neurological Disorders. *Front. Mol. Biosci.* 6, 161 (2019).**
23. **Sahoo, P. K., Smith, D. S., Perrone-Bizzozero, N. & Twiss, J. L. Axonal mRNA transport and translation at a glance. *J. Cell Sci.* 131, (2018).**
24. **Wang, E. T. *et al.* Dysregulation of mRNA Localization and Translation in Genetic Disease. *J. Neurosci.* 36, 11418–11426 (2016).**

25. **Estes, P. S., O'Shea, M., Clasen, S. & Zarnescu, D. C. Fragile X protein controls the efficacy of mRNA transport in Drosophila neurons. *Mol. Cell. Neurosci.* 39, 170–179 (2008).**

26. **Huang, Y.-S., Carson, J. H., Barbarese, E. & Richter, J. D. Facilitation of dendritic mRNA transport by CPEB. *Genes Dev.* 17, 638–653 (2003).**

27. **Ravanidis, S., Kattan, F.-G. & Doxakis, E. Unraveling the Pathways to Neuronal Homeostasis and Disease: Mechanistic Insights into the Role of RNA-Binding Proteins and Associated Factors. *Int. J. Mol. Sci.* 19, (2018).**

28. **Refaeli, R. *et al.* Features of hippocampal astrocytic domains and their spatial relation to excitatory and inhibitory neurons. *Glia* 69, 2378–2390 (2021).**

29. **Xu, A.-L., Zheng, G.-Y., Ye, H.-Y., Chen, X.-D. & Jiang, Q. Characterization of astrocytes and microglial cells in the hippocampal CA1 region after transient focal cerebral ischemia in rats treated with Ilexonin A. *Neural Regen. Res.* 15, 78–85 (2020).**

30. **Chacon-De-La-Rocha, I. *et al.* Accelerated Dystrophy and Decay of Oligodendrocyte Precursor Cells in the APP/PS1 Model of Alzheimer's-Like Pathology. *Front. Cell. Neurosci.* 14, 575082 (2020).**

31. **Franzén, O., Gan, L.-M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database (Oxford)* 2019, (2019).**

32. **Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* 48, D845–D855 (2020).**

33. **Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* 19, 477 (2018).**

34. **Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 566, 496–502 (2019).**

35. **Wang, T. *et al.* An empirical approach leveraging tumorgrafts to dissect the tumor microenvironment in renal cell carcinoma identifies missing link to prognostic inflammatory factors. *Cancer Discov.* 8, 1142–1155 (2018).**

36. **Debald, M. *et al.* Specific expression of k63-linked ubiquitination of calmodulin-like protein 5 in breast cancer of premenopausal patients. *J. Cancer Res. Clin. Oncol.* 139, 2125–2132 (2013).**

37. **Gong, B. *et al.* Caprin-1 is a novel microRNA-223 target for regulating the proliferation and invasion of human breast cancer cells. *Biomed. Pharmacother.* 67, 629–636 (2013).**

38. **Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102, 15545–15550 (2005).**

39. **Roche, J. The Epithelial-to-Mesenchymal Transition in Cancer. *Cancers (Basel)* 10, (2018).**

40. **Planes-Laine, G. *et al.* PD-1/PD-L1 Targeting in Breast Cancer: The First Clinical Evidences Are Emerging. A Literature Review. *Cancers (Basel)* 11, (2019).**

41. **Yuan, C. *et al.* Expression of PD-1/PD-L1 in primary breast tumours and metastatic axillary lymph nodes and its correlation with clinicohistological parameters. *Sci. Rep.* 9, 14356 (2019).**

42. **Li, C.-J., Lin, L.-T., Hou, M.-F. & Chu, P.-Y. PD-L1/PD-1 blockade in breast cancer: The immunotherapy era (Review). *Oncol. Rep.* 45, 5–12 (2021).**

43. **Wang, L. & Li, R.-C. Learning Low-Dimensional Latent Graph Structures: A Density Estimation Approach. *IEEE Trans. Neural Netw. Learn. Syst.* 31, 1098–1112 (2020).**

44. **Zhang, R., Atwal, G. S. & Lim, W. K. Noise regularization removes correlation artifacts in single-cell RNA-seq data preprocessing. *Patterns (N Y)* 2, 100211 (2021).**

45. **Andrews, T. S. & Hemberg, M. False signals induced by single-cell imputation. [version 2; peer review: 4 approved]. *F1000Res.* 7, 1740 (2018).**

46. **Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* 361, (2018).**

47. **Haralick, R. M., Shanmugam, K. & Dinstein, I. Textural Features for Image Classification. *IEEE Trans. Syst. Man Cybern.* 3, 610–621 (1973).**

48. **Wang, H., Wu, X. & Chen, Y. Stromal-Immune Score-Based Gene Signature: A Prognosis Stratification Tool in Gastric Cancer. *Front. Oncol.* 9, 1212 (2019).**

Fig. 1

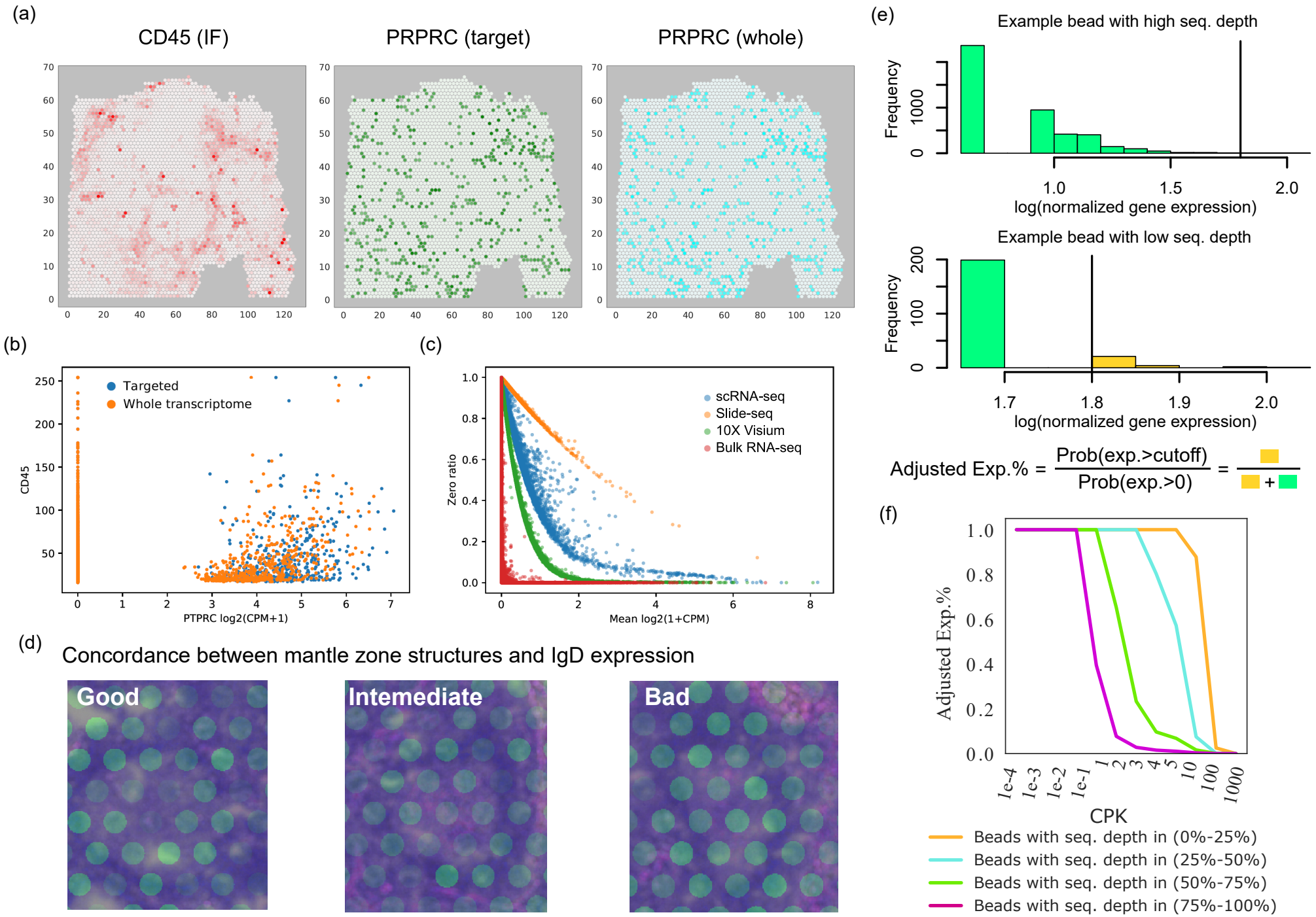

(a)
CD45 (IF)
PRPRC (target)
PRPRC (whole)
(b)
Targeted
Whole transcriptome
CD45
PTPRC log2(CPM+1)
(c)
scRNA-seq
Slide-seq
10X Visium
Bulk RNA-seq
Zero ratio
Mean log2(1+CPM)
(d)
Concordance between mantle zone structures and IgD expression
Good
Intemediate
Bad
(e)
Example bead with high seq. depth
Frequency
log(normalized gene expression)
Example bead with low seq. depth
Frequency
log(normalized gene expression)
Adjusted Exp.% = Prob(exp.>cutoff) / Prob(exp.>0) =
(f)
Adjusted Exp.%
CPK
Beads with seq. depth in (0%-25%)
Beads with seq. depth in (25%-50%)
Beads with seq. depth in (50%-75%)
Beads with seq. depth in (75%-100%)

Fig. 2

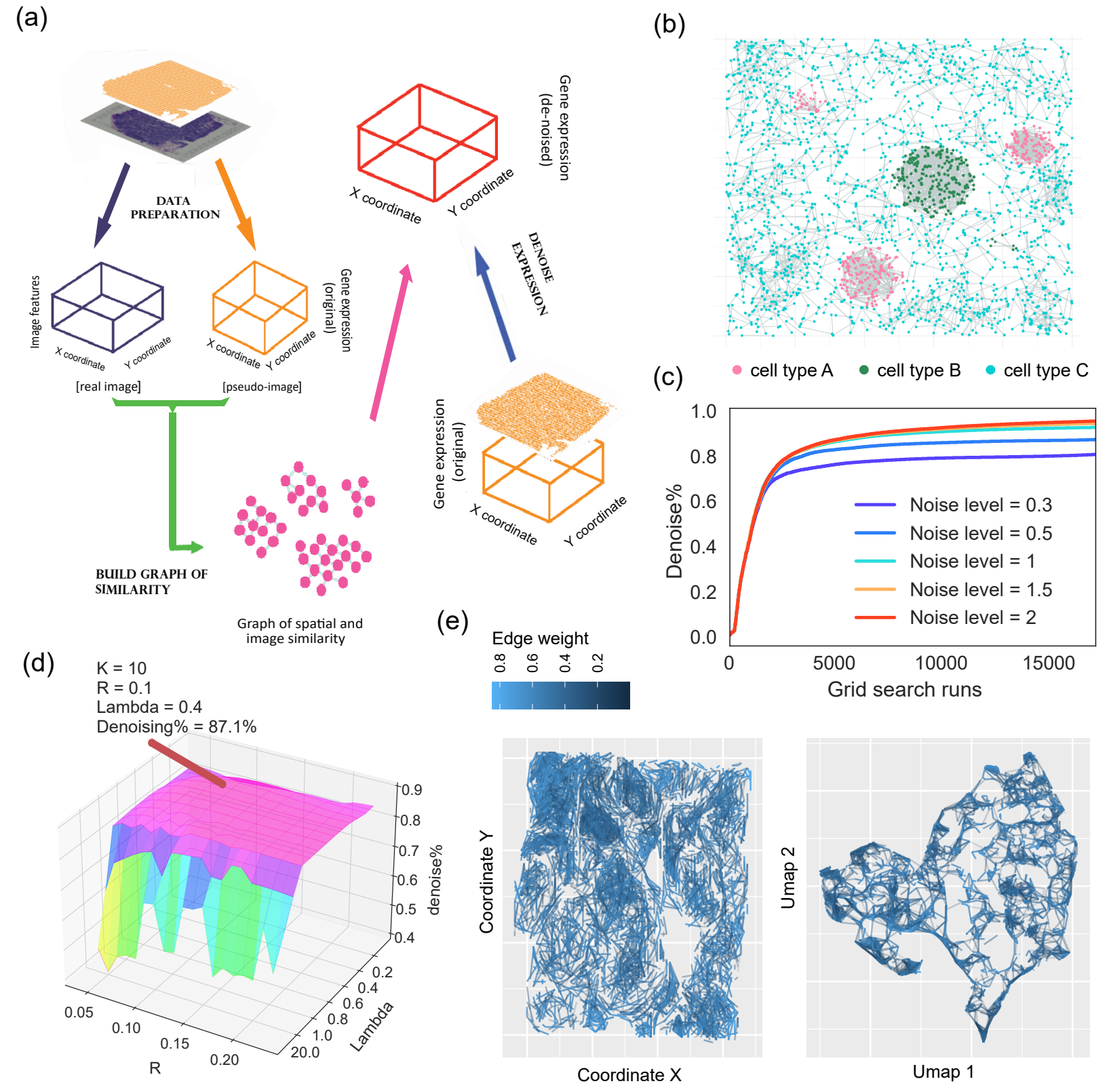

(a)
DATA PREPARATION
Image features
X coordinate
Y coordinate
[real image]
Gene expression (original)
X coordinate
Y coordinate
[pseudo-image]
Gene expression (de-noised)
X coordinate
Y coordinate
DENOISE EXPRESSION
Gene expression (original)
X coordinate
Y coordinate
BUILD GRAPH OF SIMILARITY
Graph of spatial and image similarity
(b)
cell type A
cell type B
cell type C
(c)
Denoise%
Grid search runs
Noise level = 0.3
Noise level = 0.5
Noise level = 1
Noise level = 1.5
Noise level = 2
(d)
K = 10
R = 0.1
Lambda = 0.4
Denoising% = 87.1%
denoise%
R
Lambda
(e)
Edge weight
Coordinate Y
Coordinate X
Umap 2
Umap 1

Fig. 3

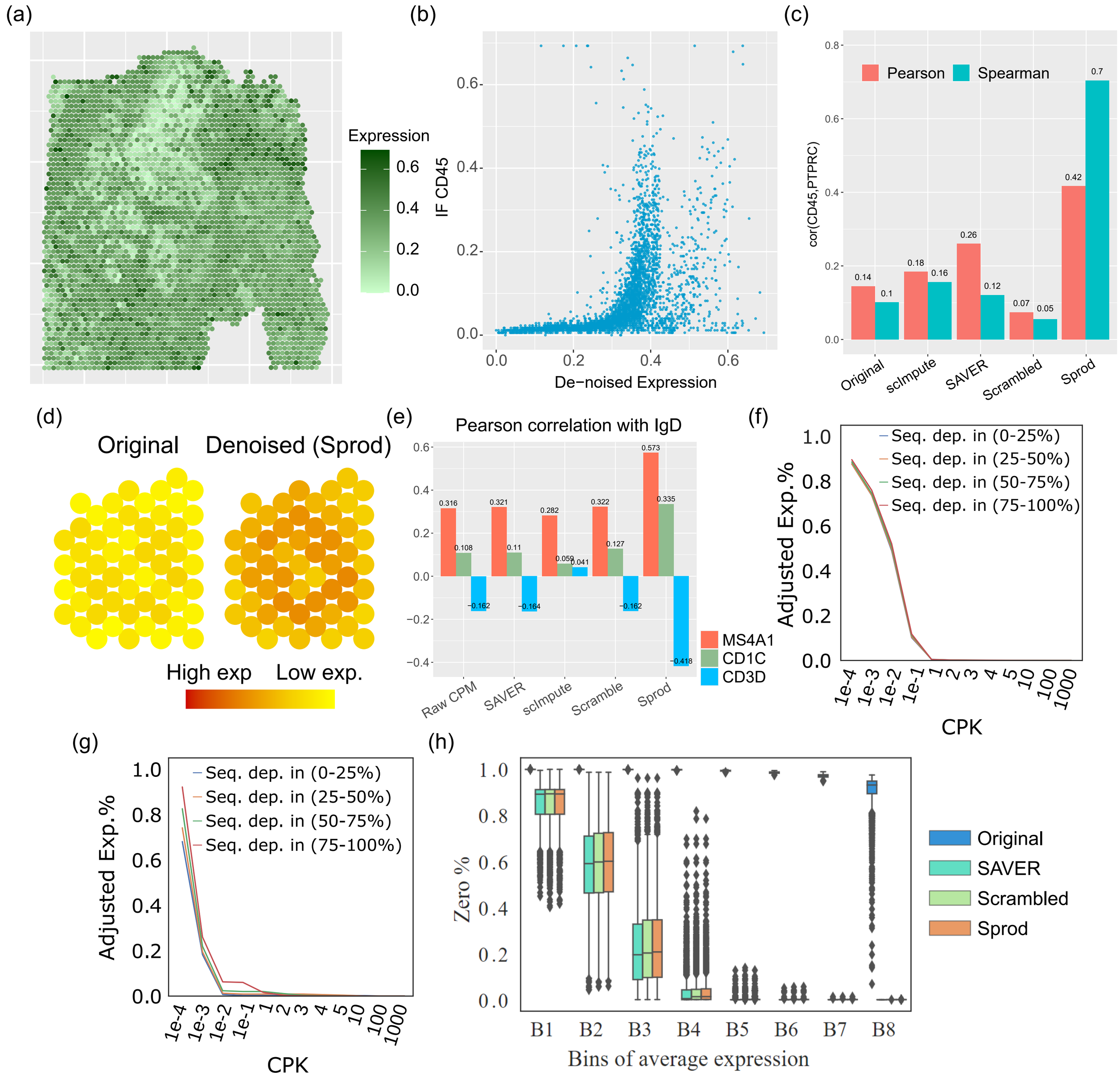

(a)
Expression
0.6
0.4
0.2
0.0
(b)
IF CD45
De−noised Expression
(c)
Pearson
Spearman
cor(CD45,PTPRC)
0.14
0.1
0.18
0.16
0.26
0.12
0.07
0.05
0.42
0.7
Original
scImpute
SAVER
Scrambled
Sprod
(d)
Original
Denoised (Sprod)
High exp
Low exp.
(e)
Pearson correlation with IgD
0.316
0.108
−0.162
0.321
0.11
−0.164
0.282
0.059
0.041
0.322
0.127
−0.162
0.573
0.335
−0.418
MS4A1
CD1C
CD3D
Raw CPM
SAVER
scImpute
Scramble
Sprod
(f)
Seq. dep. in (0-25%)
Seq. dep. in (25-50%)
Seq. dep. in (50-75%)
Seq. dep. in (75-100%)
Adjusted Exp.%
CPK
(g)
Seq. dep. in (0-25%)
Seq. dep. in (25-50%)
Seq. dep. in (50-75%)
Seq. dep. in (75-100%)
Adjusted Exp.%
CPK
(h)
Zero %
Original
SAVER
Scrambled
Sprod
B1
B2
B3
B4
B5
B6
B7
B8
Bins of average expression

Fig. 4

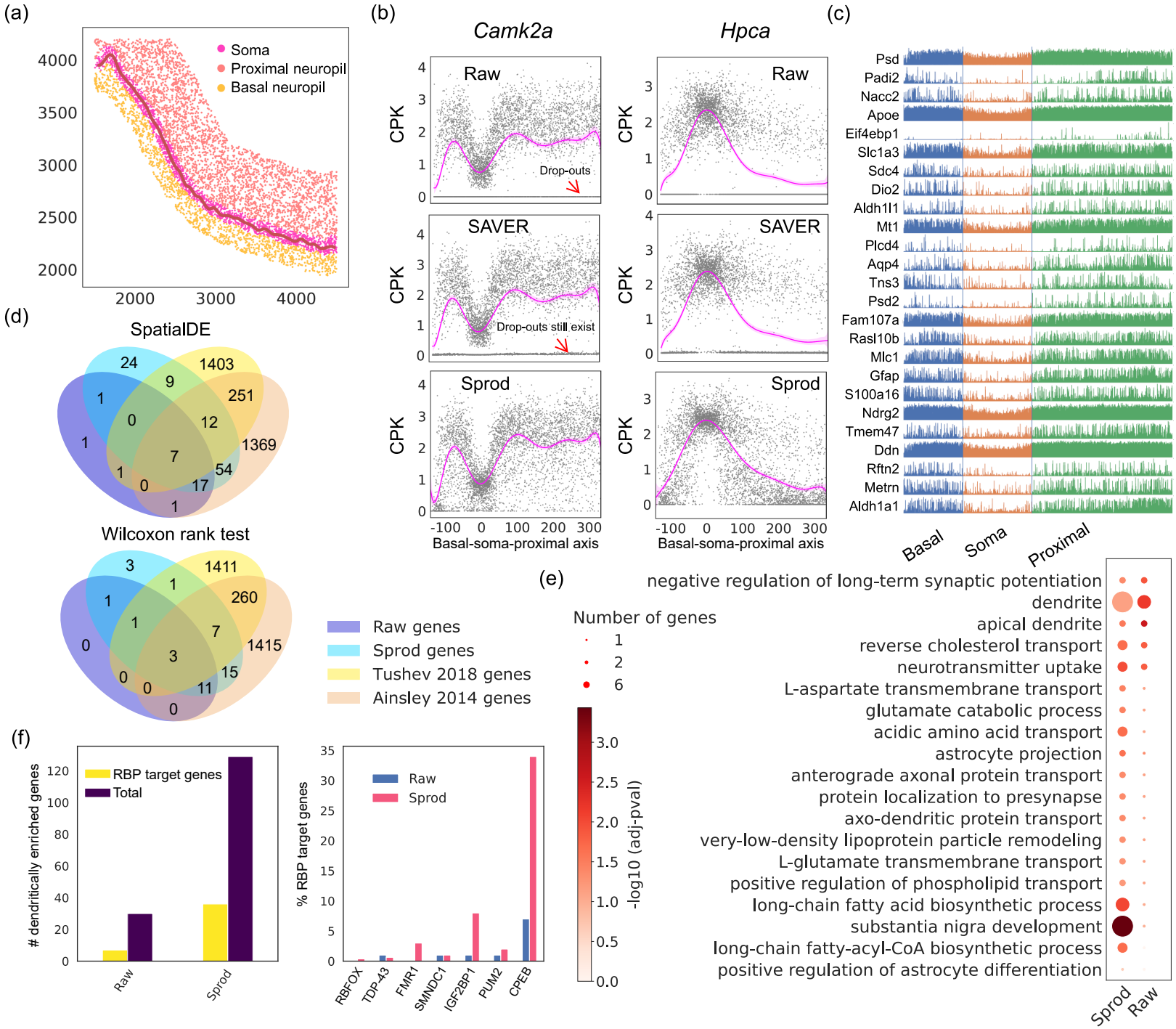

(a)
Soma
Proximal neuropil
Basal neuropil
(b)
Camk2a
Hpca
Raw
SAVER
Sprod
Drop-outs
Drop-outs still exist
CPK
Basal-soma-proximal axis
(c)
Psd
Padi2
Nacc2
Apoe
Eif4ebp1
Slc1a3
Sdc4
Dio2
Aldh1l1
Mt1
Plcd4
Aqp4
Tns3
Psd2
Fam107a
Rasl10b
Mlc1
Gfap
S100a16
Ndrg2
Tmem47
Ddn
Rftn2
Metrn
Aldh1a1
Basal
Soma
Proximal
(d)
SpatialDE
Wilcoxon rank test
Raw genes
Sprod genes
Tushev 2018 genes
Ainsley 2014 genes
(e)
Number of genes
-log10 (adj-pval)
negative regulation of long-term synaptic potentiation
dendrite
apical dendrite
reverse cholesterol transport
neurotransmitter uptake
L-aspartate transmembrane transport
glutamate catabolic process
acidic amino acid transport
astrocyte projection
anterograde axonal protein transport
protein localization to presynapse
axo-dendritic protein transport
very-low-density lipoprotein particle remodeling
L-glutamate transmembrane transport
positive regulation of phospholipid transport
long-chain fatty acid biosynthetic process
substantia nigra development
long-chain fatty-acyl-CoA biosynthetic process
positive regulation of astrocyte differentiation
Sprod
Raw
(f)
dendritically enriched genes
RBP target genes
Total
Raw
Sprod
% RBP target genes
RBFOX
TDP-43
FMR1
SMNDC1
IGF2BP1
PUM2
CPEB

Fig. 5

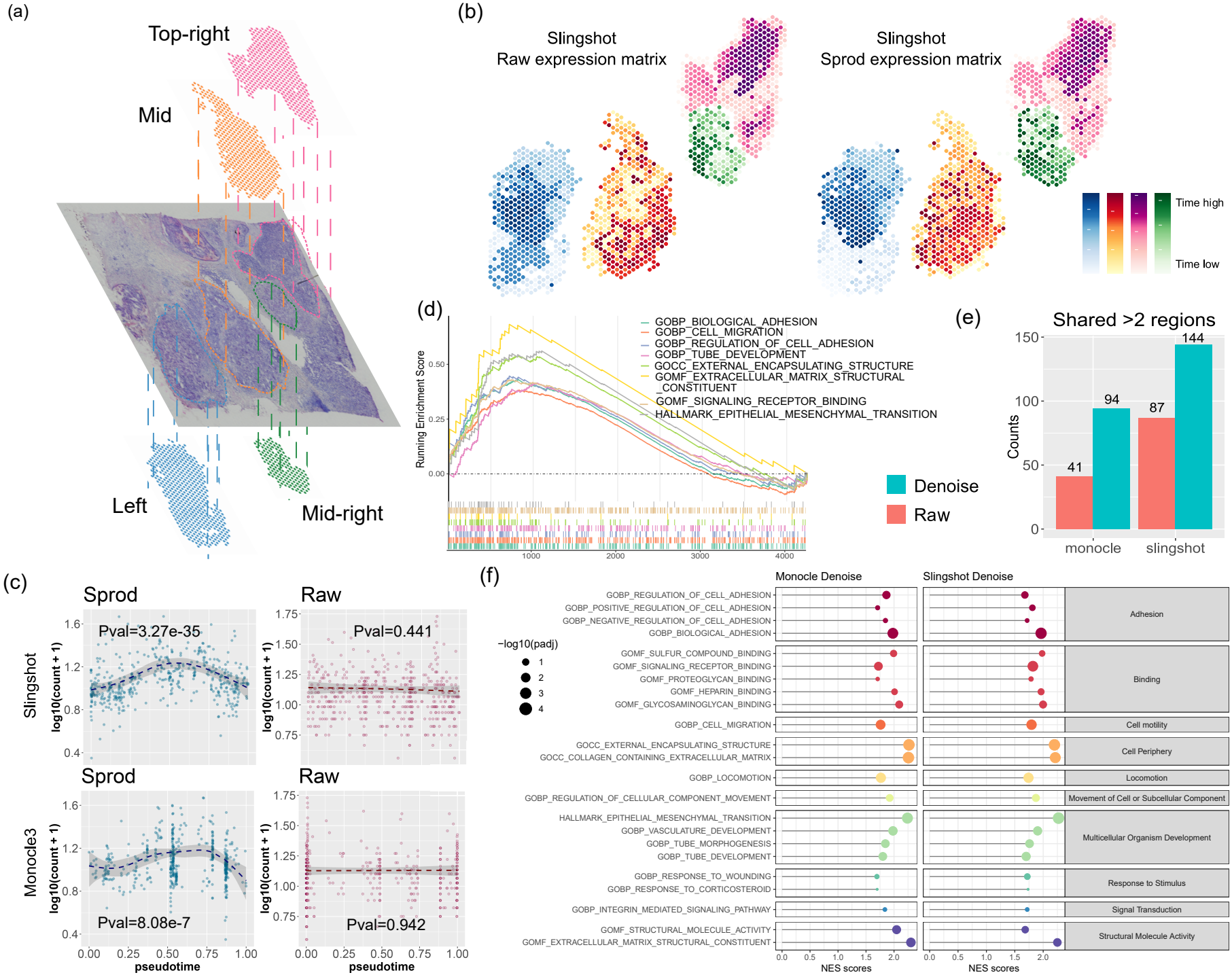
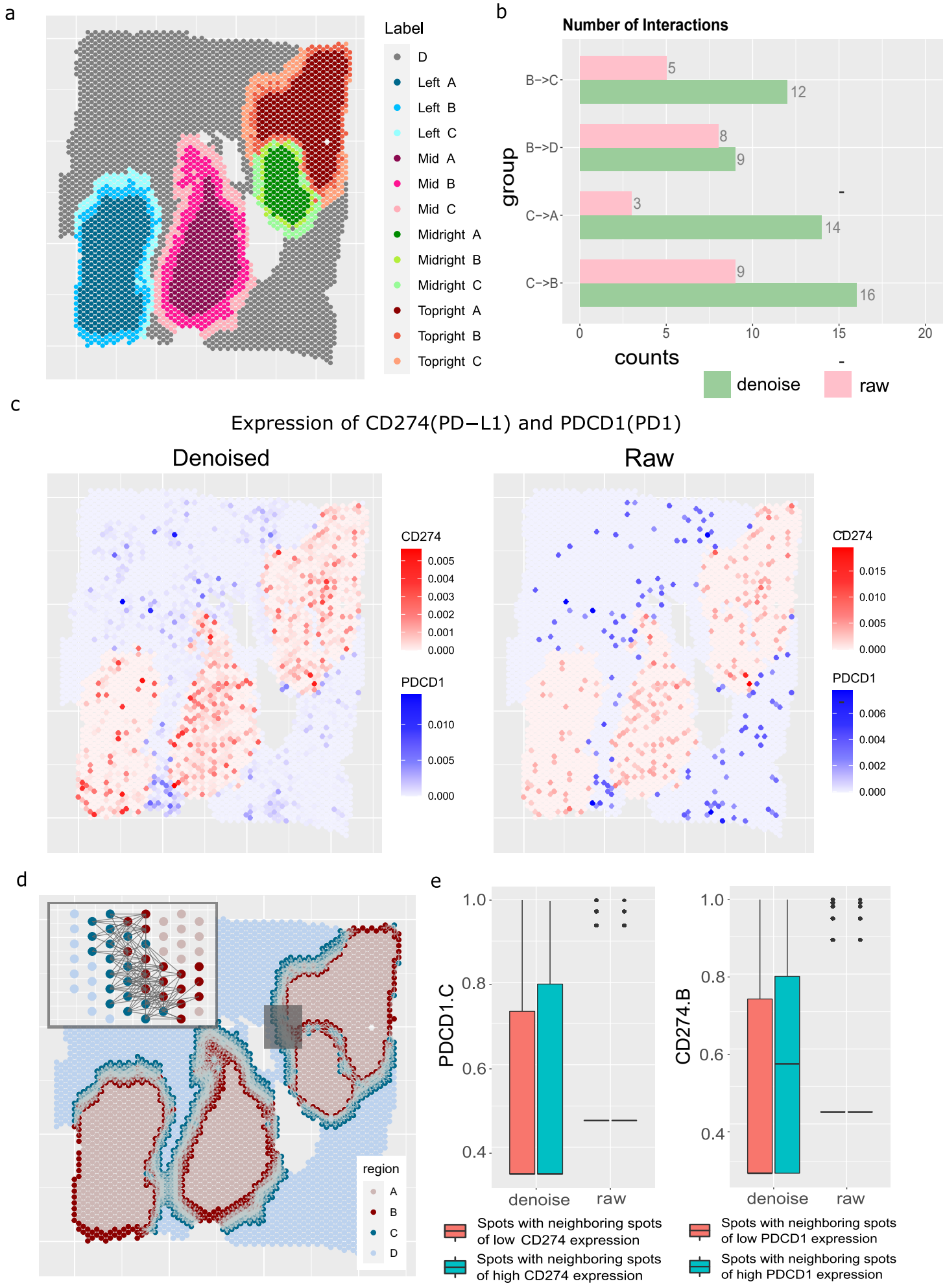
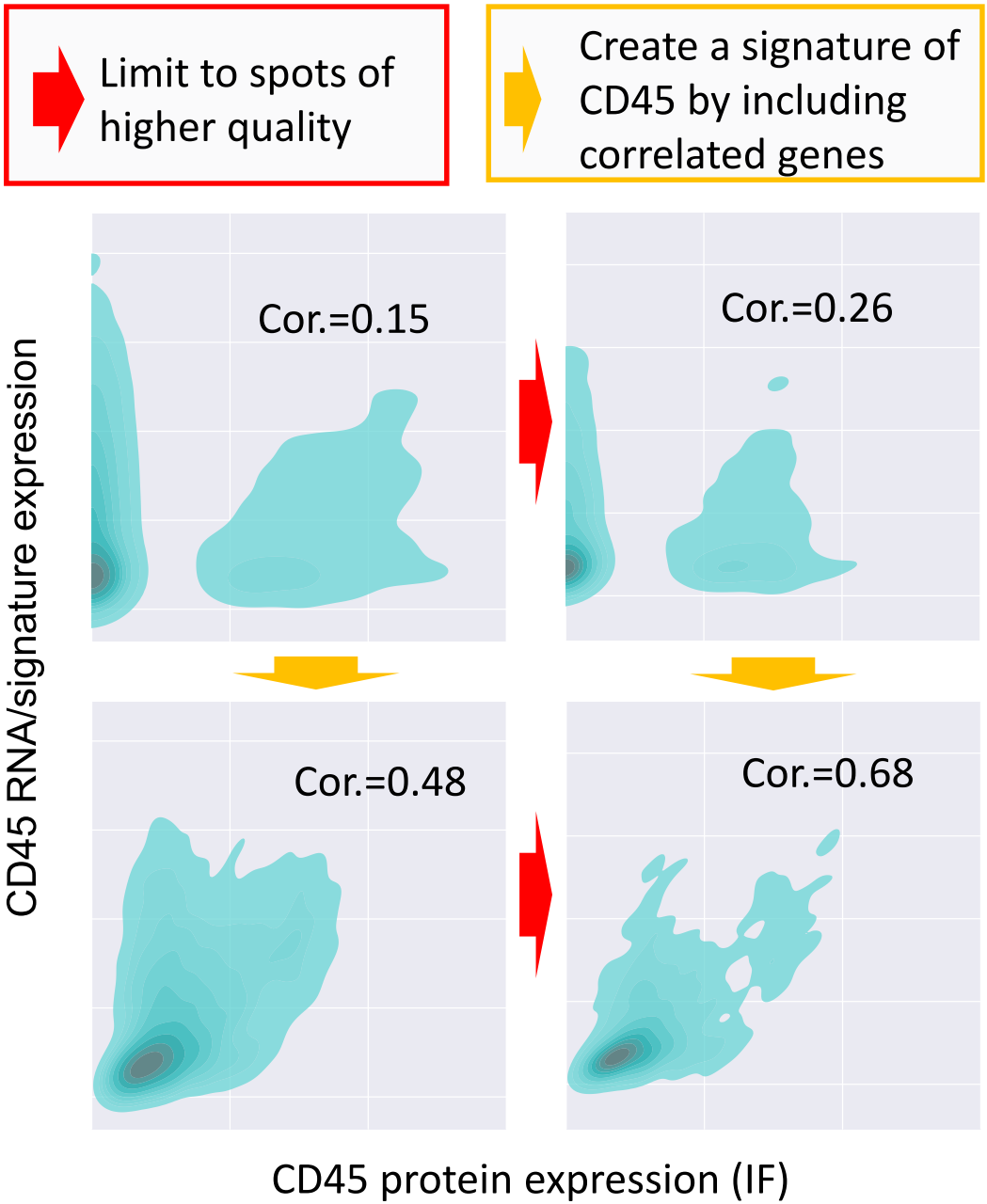

(a)
Top-right
Mid
Left
Mid-right
(b)
Slingshot
Raw expression matrix
Slingshot
Sprod expression matrix
Time high
Time low
(d)
Running Enrichment Score
GOBP_BIOLOGICAL_ADHESION
GOBP_CELL_MIGRATION
GOBP_REGULATION_OF_CELL_ADHESION
GOBP_TUBE_DEVELOPMENT
GOCC_EXTERNAL_ENCAPSULATING_STRUCTURE
GOMF_EXTRACELLULAR_MATRIX_STRUCTURAL_CONSTITUENT
GOMF_SIGNALING_RECEPTOR_BINDING
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
Denoise
Raw
(e)
Shared >2 regions
Counts
41
94
87
144
monocle
slingshot
(c)
Sprod
Raw
Slingshot
Monocle3
log10(count + 1)
Pval=3.27e-35
Pval=0.441
Pval=8.08e-7
Pval=0.942
pseudotime
(f)
Monocle Denoise
Slingshot Denoise
−log10(padj)
GOBP_REGULATION_OF_CELL_ADHESION
GOBP_POSITIVE_REGULATION_OF_CELL_ADHESION
GOBP_NEGATIVE_REGULATION_OF_CELL_ADHESION
GOBP_BIOLOGICAL_ADHESION
GOMF_SULFUR_COMPOUND_BINDING
GOMF_SIGNALING_RECEPTOR_BINDING
GOMF_PROTEOGLYCAN_BINDING
GOMF_HEPARIN_BINDING
GOMF_GLYCOSAMINOGLYCAN_BINDING
GOBP_CELL_MIGRATION
GOCC_EXTERNAL_ENCAPSULATING_STRUCTURE
GOCC_COLLAGEN_CONTAINING_EXTRACELLULAR_MATRIX
GOBP_LOCOMOTION
GOBP_REGULATION_OF_CELLULAR_COMPONENT_MOVEMENT
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
GOBP_VASCULATURE_DEVELOPMENT
GOBP_TUBE_MORPHOGENESIS
GOBP_TUBE_DEVELOPMENT
GOBP_RESPONSE_TO_WOUNDING
GOBP_RESPONSE_TO_CORTICOSTEROID
GOBP_INTEGRIN_MEDIATED_SIGNALING_PATHWAY
GOMF_STRUCTURAL_MOLECULE_ACTIVITY
GOMF_EXTRACELLULAR_MATRIX_STRUCTURAL_CONSTITUENT
Adhesion
Binding
Cell motility
Cell Periphery
Locomotion
Movement of Cell or Subcellular Component
Multicellular Organism Development
Response to Stimulus
Signal Transduction
Structural Molecule Activity
NES scores

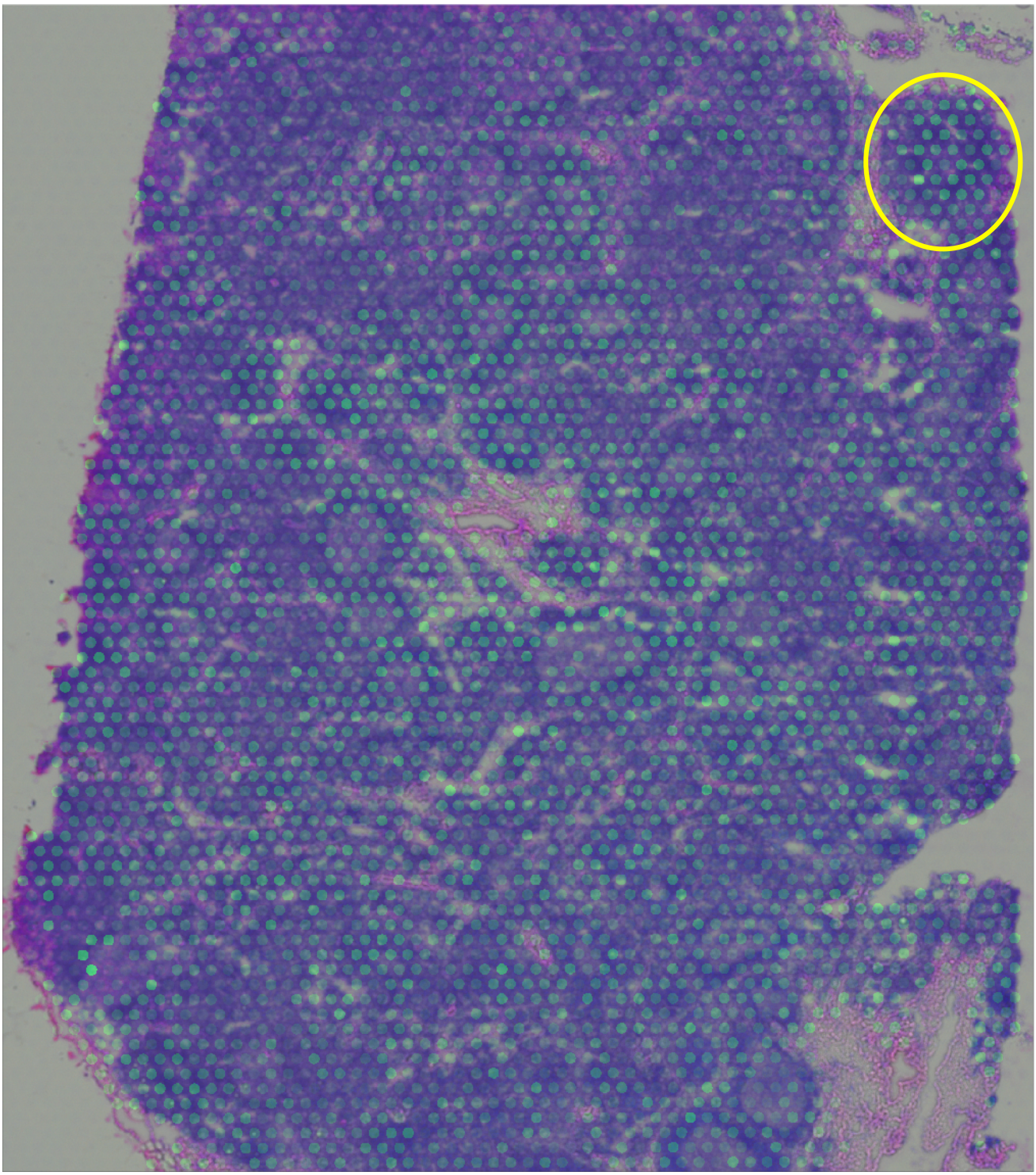
Fig. 6


a
Label
D
Left A
Left B
Left C
Mid A
Mid B
Mid C
Midright A
Midright B
Midright C
Topright A
Topright B
Topright C
b
Number of Interactions
group
B->C
B->D
C->A
C->B
5
12
8
9
3
14
9
16
0
5
10
15
20
counts
denoise
raw
c
Expression of CD274(PD−L1) and PDCD1(PD1)
Denoised
Raw
CD274
0.005
0.004
0.003
0.002
0.001
0.000
PDCD1
0.010
0.005
0.000
CD274
0.015
0.010
0.005
0.000
PDCD1
0.006
0.004
0.002
0.000
d
region
A
B
C
D
e
PDCD1.C
1.0
0.8
0.6
0.4
denoise
raw
CD274.B
1.0
0.8
0.6
0.4
denoise
raw
Spots with neighboring spots of low CD274 expression
Spots with neighboring spots of high CD274 expression
Spots with neighboring spots of low PDCD1 expression
Spots with neighboring spots of high PDCD1 expression

Sup. Fig. 1

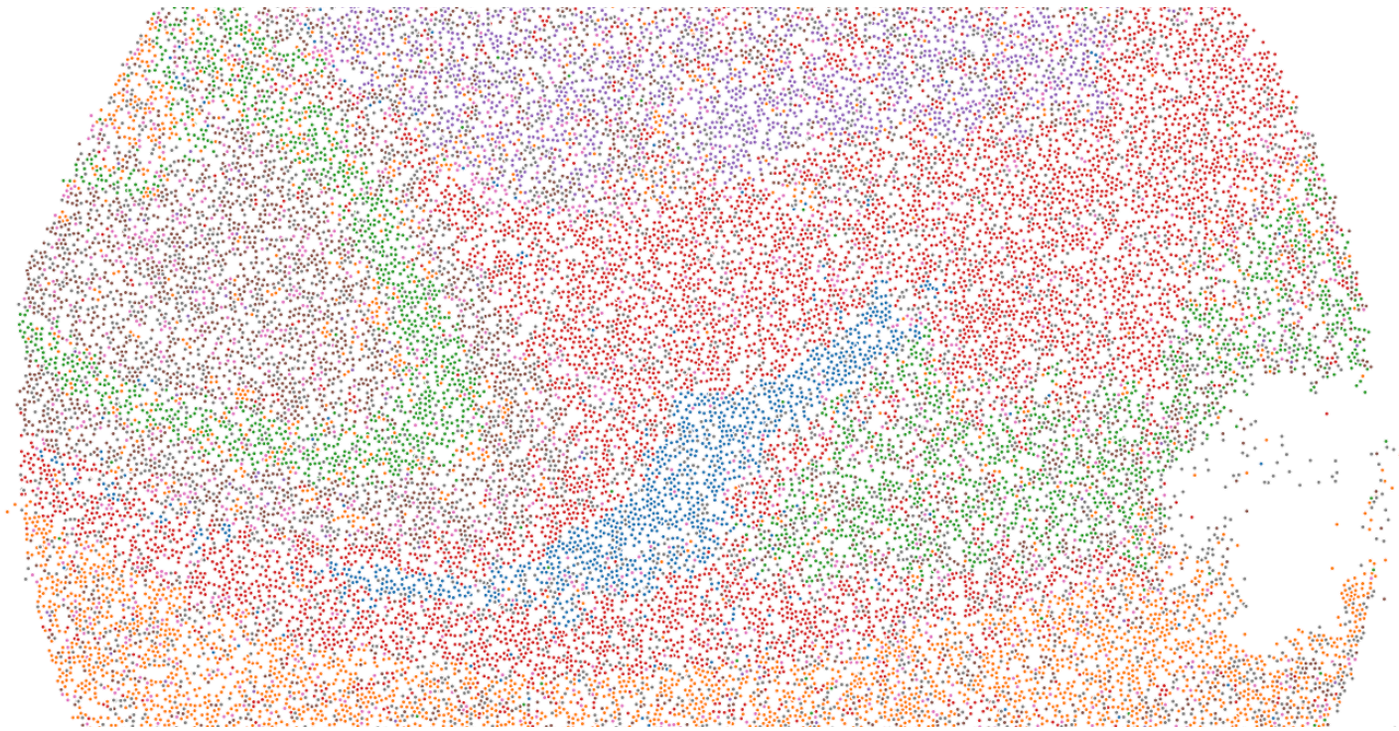

Limit to spots of higher quality
Create a signature of CD45 by including correlated genes
Cor.=0.15
Cor.=0.26
Cor.=0.48
Cor.=0.68
CD45 RNA/signature expression
CD45 protein expression (IF)

Sup. Fig. 2

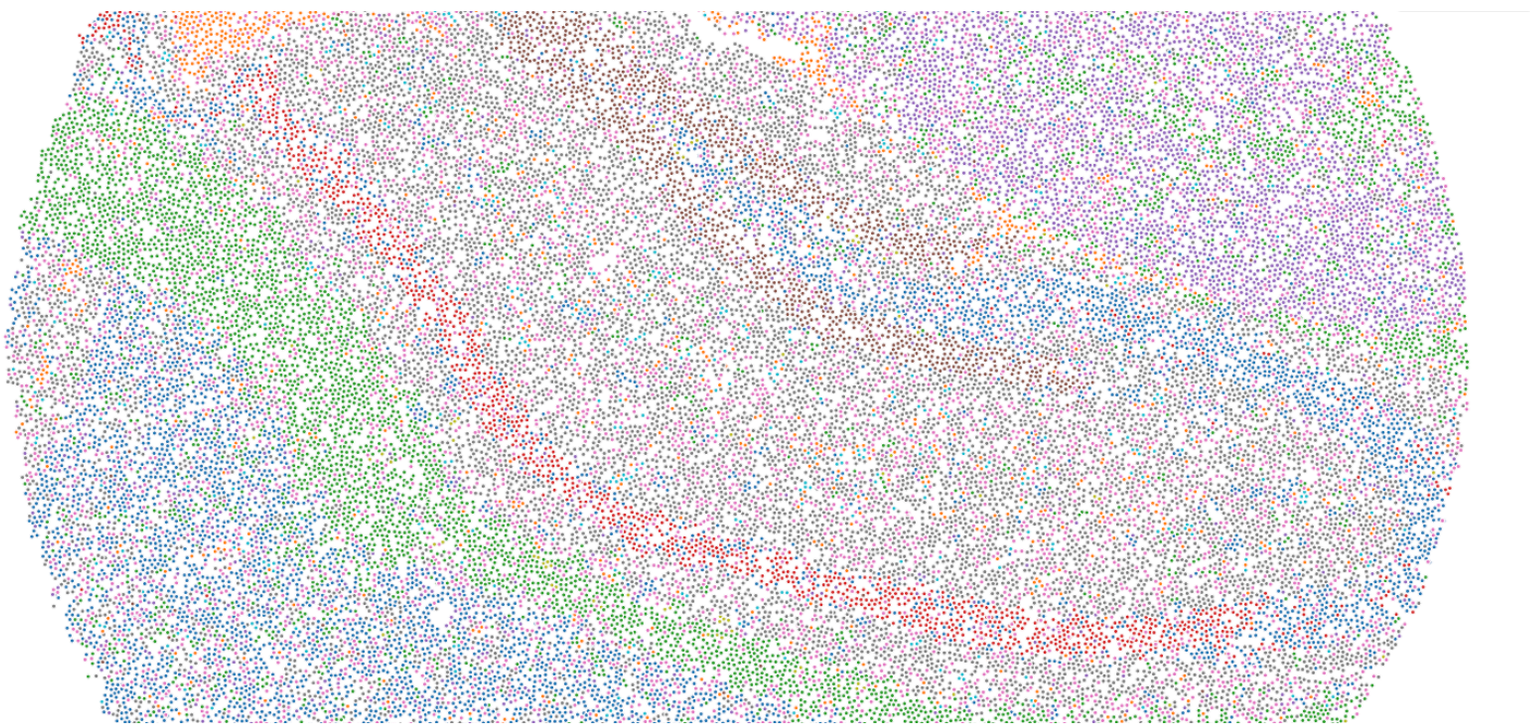


Sup. Fig. 2

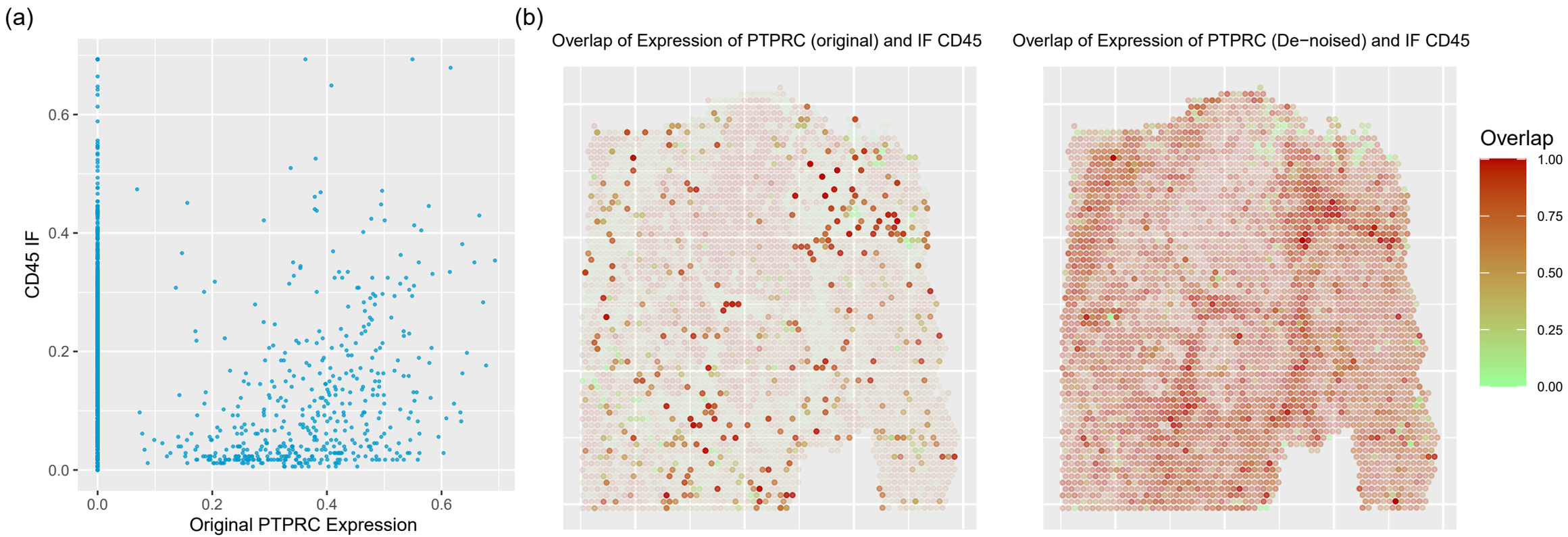
(a)



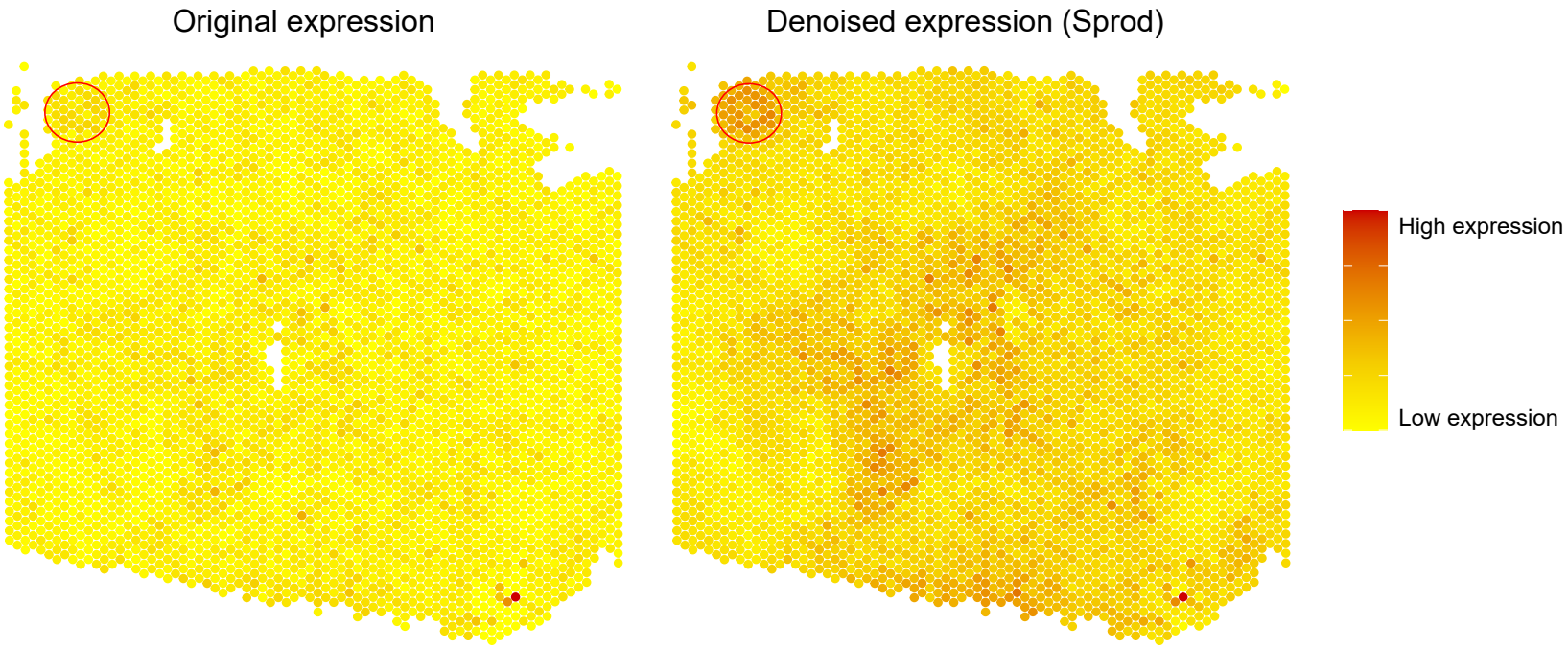
(b)



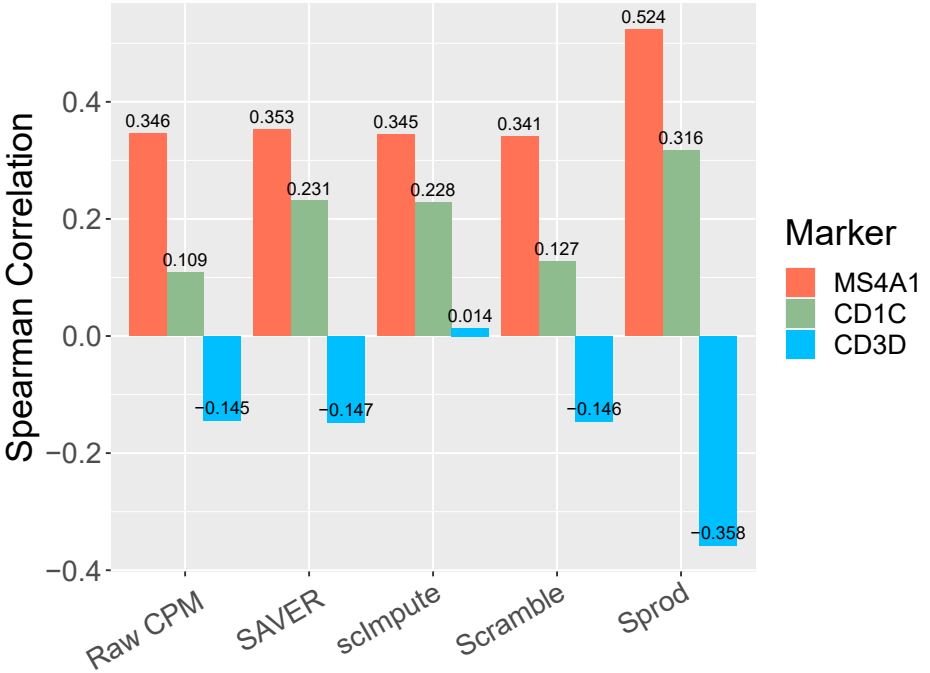
Sup. Fig. 4


(a)
CD45 IF
0.0
0.2
0.4
0.6
Original PTPRC Expression
(b)
Overlap of Expression of PTPRC (original) and IF CD45
Overlap of Expression of PTPRC (De−noised) and IF CD45
Overlap
1.00
0.75
0.50
0.25
0.00

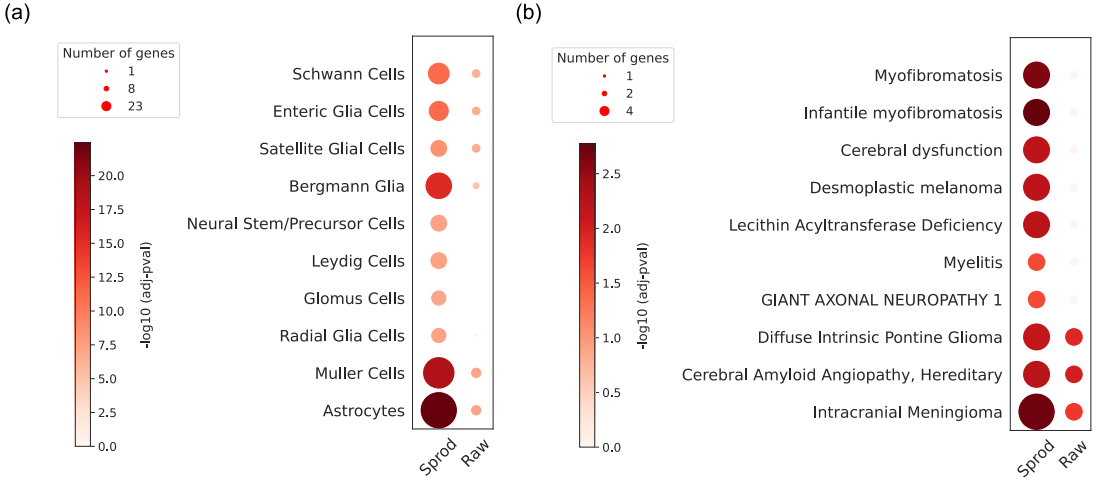
Sup. Fig. 5


Original expression
Denoised expression (Sprod)
High expression
Low expression

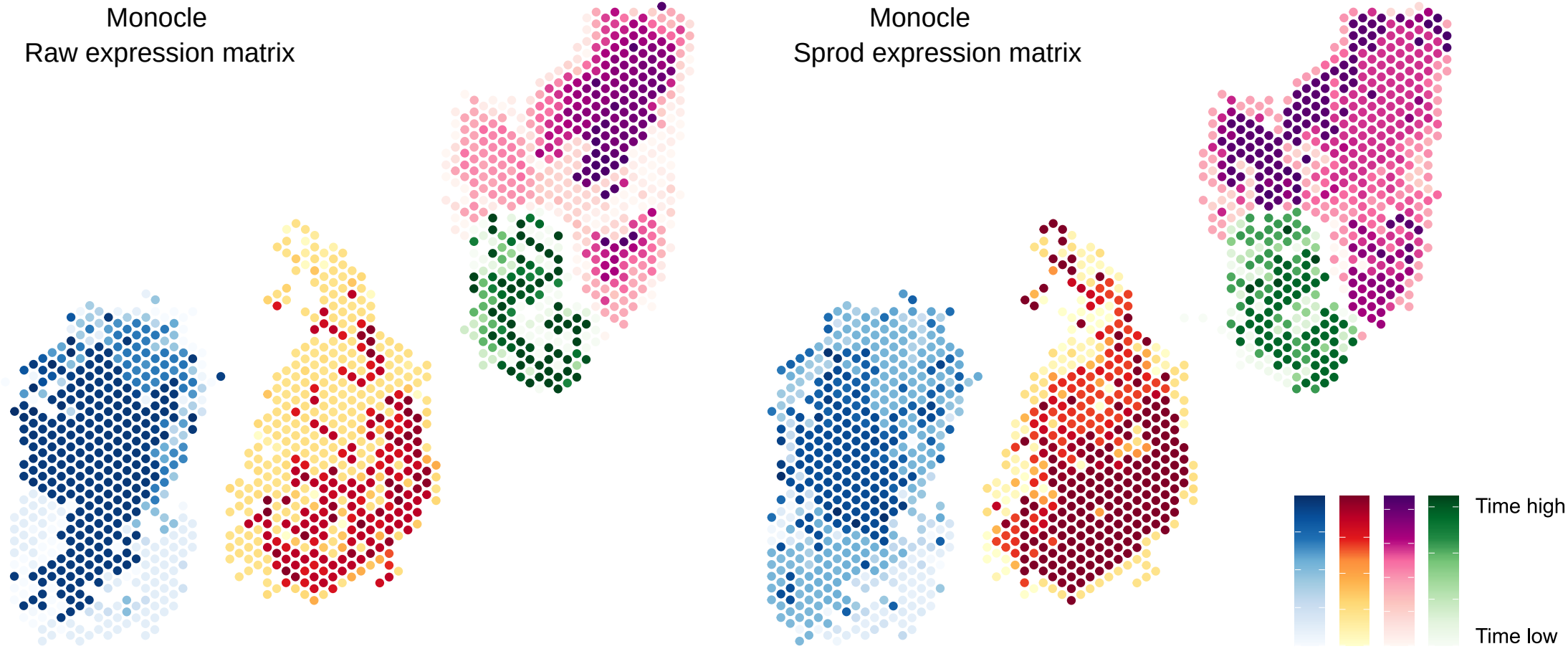
Sup. Fig. 6


Spearman Correlation
0.4
0.2
0.0
−0.2
−0.4
0.346
0.109
−0.145
0.353
0.231
−0.147
0.345
0.228
0.014
0.341
0.127
−0.146
0.524
0.316
−0.358
Raw CPM
SAVER
scImpute
Scramble
Sprod
Marker
MS4A1
CD1C
CD3D

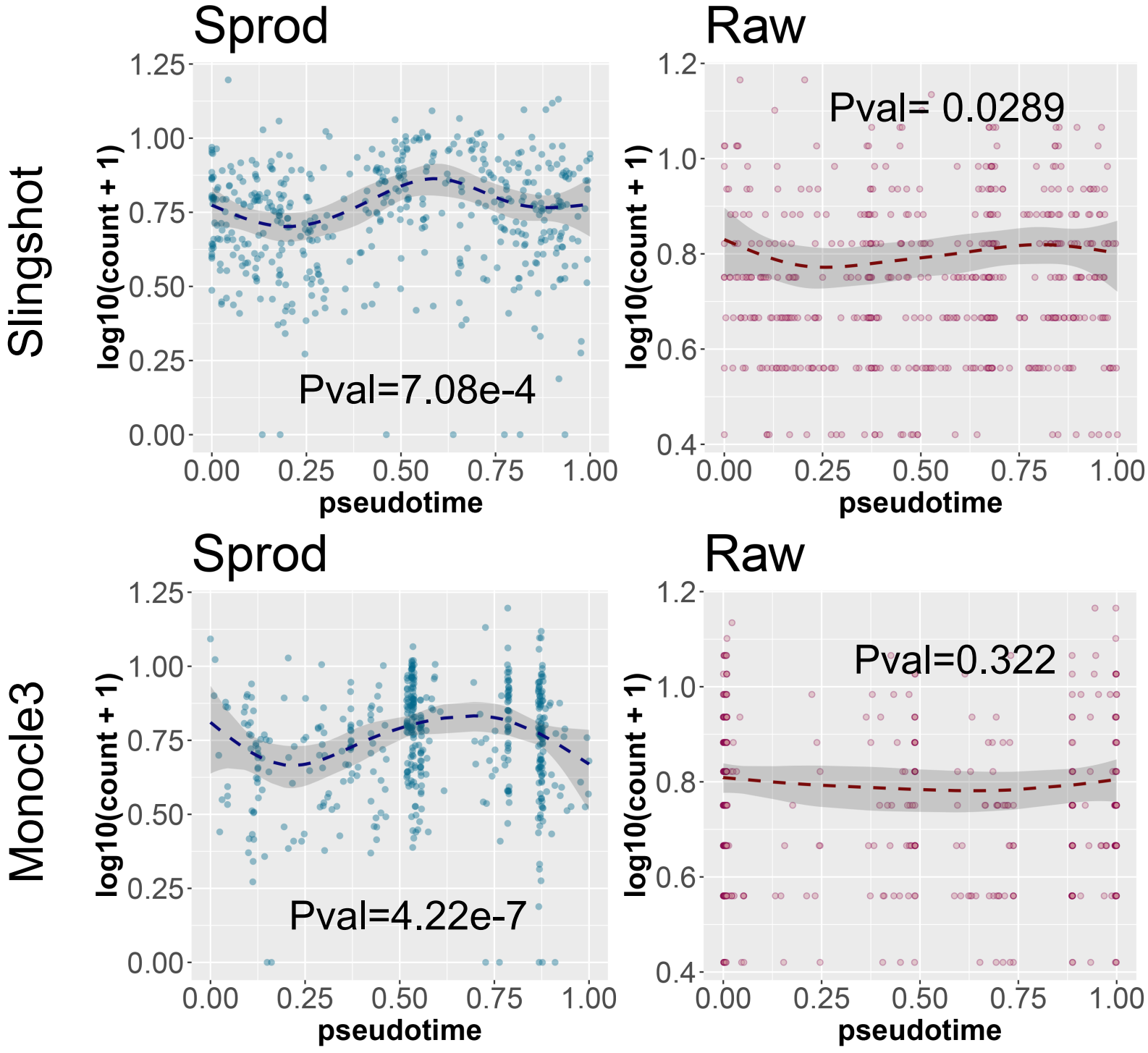
Sup. Fig. 7


(a)
Number of genes
1
8
23
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
-log10 (adj-pval)
Schwann Cells
Enteric Glia Cells
Satellite Glial Cells
Bergmann Glia
Neural Stem/Precursor Cells
Leydig Cells
Glomus Cells
Radial Glia Cells
Muller Cells
Astrocytes
Sprod
Raw
(b)
Number of genes
1
2
4
2.5
2.0
1.5
1.0
0.5
0.0
-log10 (adj-pval)
Myofibromatosis
Infantile myofibromatosis
Cerebral dysfunction
Desmoplastic melanoma
Lecithin Acyltransferase Deficiency
Myelitis
GIANT AXONAL NEUROPATHY 1
Diffuse Intrinsic Pontine Glioma
Cerebral Amyloid Angiopathy, Hereditary
Intracranial Meningioma
Sprod
Raw

Sup. Fig. 8


Monocle
Raw expression matrix
Monocle
Sprod expression matrix
Time high
Time low

Sup. Fig. 9


Sprod
Raw
Slingshot
Monocle3
log10(count + 1)
pseudotime
Pval=7.08e-4
Pval= 0.0289
Pval=4.22e-7
Pval=0.322
1.25
1.00
0.75
0.50
0.25
0.00
1.2
1.0
0.8
0.6
0.4