ARTICLES
https://doi.org/10.1038/s41592-022-01560-w

nature methods

Check for updates

Sprod for de-noising spatially resolved transcriptomics data based on position and image information

Yunguan Wang[1,6], Bing Song[1,6], Shidan Wang[1], Mingyi Chen[2], Yang Xie[1,3], Guanghua Xiao[1,3], Li Wang[4] ✉ and Tao Wang[1,5] ✉

Spatially resolved transcriptomics (SRT) 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, SRT expression data suffer from high noise levels, due to the shallow coverage in each sequencing unit and the extra experimental steps required to preserve the locations of sequencing. Fortunately, such noise can be removed by leveraging information from the physical locations of sequencing, and the tissue organization reflected in corresponding pathology images. In this work, we developed Sprod, based on latent graph learning of matched location and imaging data, to impute accurate SRT gene expression. We validated Sprod comprehensively and demonstrated its advantages over previous methods for removing drop-outs in single-cell RNA-sequencing data. We showed that, after imputation by Sprod, differential expression analyses, pathway enrichment and cell-to-cell interaction inferences are more accurate. Overall, we envision de-noising by Sprod to become a key first step towards empowering SRT technologies for biomedical discoveries.

SRT technologies, such as 10X Visium, Slide-Seq[1,2], high-definition spatial transcriptomics (HDST)[3], Seq-scope[4] and XYZeq[5], have bloomed recently. SRTs provide gene expression data for the whole transcriptome at near- or sub-single cell resolution, in tandem with matching spatial information and, for many techniques, also matched pathology images stained by hematoxylin and eosin (H&E) or immunofluorescence (IF). The units of sequencing are called spots or beads in different techniques but are essentially a small cluster of cells (up to a few dozen), or even parts of a single cell for some newer techniques with higher spatial resolution, such as HDST. These powerful techniques have enabled researchers to localize cell types within tissues, characterize spatial expression patterns, define local cell ecosystems and resolve the spatiotemporal order of cellular development.

However, such technologies suffer from severe noise in gene expression measurements. The noise comes from random variations due to the shallow nature of sequencing for each spot/bead, similar to standard single-cell RNA-sequencing (scRNA-seq), but is also further complicated by the extra and delicate experimental steps needed to preserve the spatial locations of sequencing. Thus, preprocessing of SRT data to remove such noise is necessary before any downstream analyses. The methodologies developed for addressing scRNA-Seq drop-outs (loss of expression)[6,7] are probably insufficient for this purpose, as expressional noise in SRTs can be different from that generated just by drop-outs. Moreover, such methods rely on only the expression data themselves to correct for drop-outs, so can be limited in the extent to which drop-outs can be corrected reliably. In other words, the methodologies used ignore the spatial and imaging features of the spots/beads provided by SRTs, which can potentially guide and improve the correction of SRT noise with useful external information.

In this work, we demonstrate the existence of extensive noise in SRT data. We developed Sprod, short for spatial profiling de-noising, to impute accurate gene expression by leveraging the location information of each measurement and the corresponding imaging data, which are available for many SRT techniques. By testing on a variety of SRT datasets, we validated Sprod's accuracy and robustness systematically. We also showed its superiority to existing drop-out removal methods for scRNA-seq data. Applying Sprod to several real SRT datasets revealed interesting biological features that were not discovered before due to noise in the data. Overall, using these example applications, we show that careful handling of technical noise in SRT data is a critical first step to the unbiased discovery of new biological knowledge.

Results

Extensive noise exists in SRT data. We first demonstrated the existence of extensive noise in SRT data. We investigated a 10X Visium ovarian cancer dataset, with matched IF images of CD45/LCA (leukocyte common antigen), keratin and DNA. Figure 1a presents 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 and whole transcriptome sequencing by Visium. Strikingly, there is a very poor correlation between CD45 protein IF and *PTPRC* RNA expression. When the spots were subdivided, keeping only those with higher overall sequencing depth, which are potentially of higher quality, the correlation improved by a large extent (Extended Data Fig. 1). One of

[1]Quantitative Biomedical Research Center, Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA. [2]Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX, USA. [3]Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX, USA. [4]Department of Mathematics and Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, USA. [5]Center for the Genetics of Host Defense, University of Texas Southwestern Medical Center, Dallas, TX, USA. [6]These authors contributed equally: Yunguan Wang, Bing Song. ✉e-mail: li.wang@uta.edu; Tao.Wang@UTSouthwestern.edu

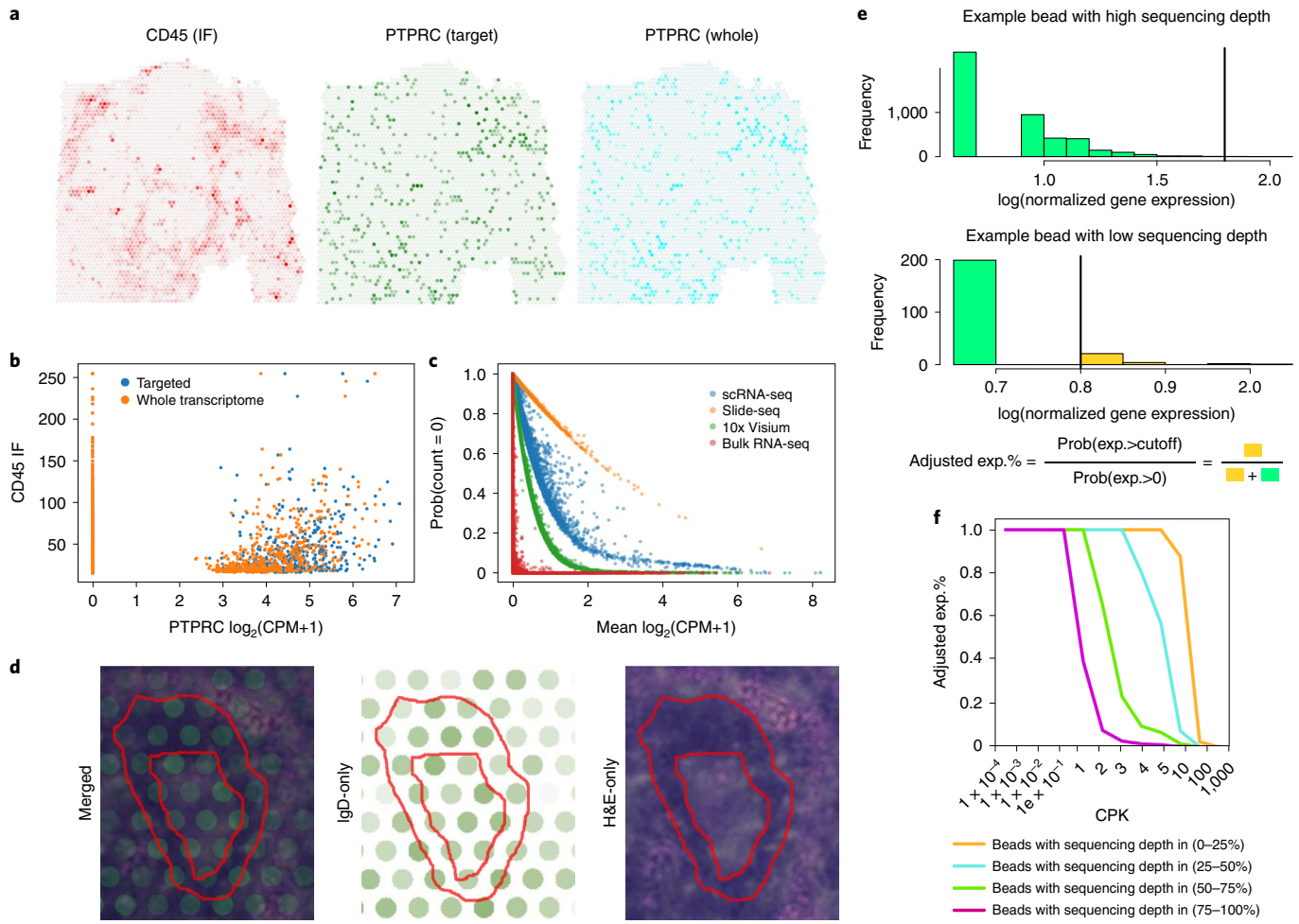


Fig. 1 | Extensive noise in spatially resolved transcriptomics data. a, Spatial expression levels of CD45 (IF), *PTPRC* (targeted panel sequencing) and *PTPRC* (whole transcriptomic sequencing) of the 10X Visium Ovarian Cancer dataset. **b**, Severe drop-outs in *PTPRC* gene expression for both targeted sequencing and whole transcriptomic sequencing. The *x* axis shows the *PTPRC* RNA expression level (unit = $\log_2$(CPM+1)). **c**, Severe drop-outs in scRNA-seq, Slide-Seq, Visium and bulk RNA-Seq. The *x* axis shows average RNA expression levels for each gene profiled by each technique/dataset, and the *y* axis shows the percentages of counts of exactly 0 for each gene. **d**, Example mantle zone structure with poor agreement with IgD expression from the 10X Visium human lymph node dataset. The red lines mark the borders of the mantle zone. **e**, Cartoon describing the calculation of the percentage of sequencing spots with expression larger than a given cutoff in the subset of spots with non-zero expression (adjusted exp.%). Prob, probability. **f**, 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.

the many possible sources of inaccuracies results from drop-outs, as in standard scRNA-seq[6–8]. In Fig. 1b, it is clear that a high level of drop-outs exists in the expression of *PTPRC* (excessive zeros at $X=0$). We created a gene signature for *PTPRC* RNA expression by including genes highly correlated with *PTPRC*, which essentially removed drop-outs by ad hoc averaging. We found that this further improved the correlation of the *PTPRC* RNA signature with CD45 protein IF intensity (Extended Data Fig. 1) compared with the single *PTPRC* gene.

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, standard scRNA-seq and standard bulk RNA-seq (from the TCGA breast cancer cohort). In all techniques, the percentages of zero counts increased with lower average gene expression levels, as expected. The drop-outs rates for Visium are lower than those of standard 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 standard scRNA-seq, possibly due to its lower per-cell sequencing depth.

Noise from SRT data can also stem from inflation of gene expression, rather than only a loss of expression (extreme cases are drop-outs). We examined another 10X Visium dataset of human benign reactive lymph nodes. In normal lymph nodes, follicles and perifollicular 'ring-like' structures exist (darker rings in H&E-stained tissue images in Fig. 1d and Extended Data Fig. 2), which represent reactive germinal centers (GCs) surrounded by well-defined mantle zones[9–11]. However, examination of the expression of IgD, a marker of mantle zone, in this Visium dataset shows only a weak correlation with the ring-like mantle zone structures on H&E-stained tissue sides. In one example, mantle zone (Fig. 1d), the green color, which stains immunoglobulin D (IgD) expression, correlates only weakly with the purple staining, which indicates the mantle zone. More importantly, drop-outs cannot completely explain the weak correlation observed, as it seems that IgD expression is strong in some places (such as the GC in the center of this mantle zone) where they should not be.

Slide-Seq data were 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. Slide-Seq has a much shallower sequencing depth compared with Visium and standard scRNA-Seq. This could result in severe noise in gene expression measurement, where the measured expression could be either artificially higher or lower than the true value. Here, we focused on the problem of inflated high expression (false positive), as the problem of drop-outs has already been investigated. We hypothesized that beads with lower sequencing quality are more likely to demonstrate overly inflated expression. As we could show (Extended Data Fig. 1) that the total sequencing depth of each bead can be used as a surrogate of bead sequencing quality, we divided all beads into four 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 nonzero expression of this gene (Fig. 1e). Since bead sequencing quality is a technical issue, we expected the beads in different bins to follow similar distributions of gene expression. However, as we show in Fig. 1f, there is a dramatic difference between beads in different quality bins in terms of their gene expression distribution. 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 distribution difference among bins of different technical qualities could confer a high level of bias in analyses.

Sprod for de-noising SRT data based on latent graph learning. To remove the substantial level of noise in SRT expression data, we developed Sprod (Fig. 2a), which corrects the noise in expression data guided by 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. However, it is imperative to consider the cell type heterogeneity of the spots/beads in this neighborhood. Therefore, borrowing information is restricted only to the cells of the same type and/or similar expression profiles. For the Visium platform, SRT data are provided with matched pathology images, from which textures and channel intensities can be derived to inform 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 (as shown in Extended Data Fig. 3), which essentially averages out noise from individual genes 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. Sprod then creates a 'pseudoimage', the 'image channels' of which are the detected cell clusters/types and their assignment probabilities. In the second stage of the model, denoised 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 Methods and Supplementary Note 1.

We performed simulation analysis to evaluate and showcase the performance of Sprod. We simulated a dataset of 5,000 spots (Supplementary Note 2). The spots were divided into three cell types: A, B and C. 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 (see example in Fig. 2b). We evaluated the de-noising 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 de-noising% = 1 – SAE(denoised)/SAE(raw). We tried combinations of all tuning parameters on this dataset (all tested parameters described in Methods). Figure 2c shows that Sprod can remove noise in the SRT data reasonably well across a wide range of parameter combinations. Our exploration of these parameter combinations shows that three tuning parameters, R (the radius defining the neighborhood of the spots), K (the dimension of the latent space) and Lambda (a scaling parameter to control clustering sparsity), make a critical contribution to the de-noising%, while the other tuning parameters (such as the t-distributed stochastic neighbor embedding (t-SNE)-type perplexity for building the similarity graph) have minimum influence. Overall, $K = 10$ works the best. In Fig. 2d, we show a surface plot of the performances of Sprod (de-noising%) with respect to 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 (Supplementary Table 1), which serves as a reference for the application of Sprod on the real datasets below.

In practice, the simulation data are still different from the real datasets, and real datasets also differ from each other, inevitably calling for tuning of parameters. We recommend that users start from the optimal parameter set in our simulation dataset above, and use the two sets of diagnostic plots that we provide with the Sprod software (Fig. 2e; 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 captured the pattern of tissue organization correctly. With a good parameter set, the pattern of the similarity graph edges should be reflective of the pattern of the pathological images. 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 uniform manifold approximation and projection (UMAP); see UMAP result in Fig. 2e). With a good parameter set, the similarity graph edges will connect the spots/beads that are close to each other in the imaging feature t-SNE/UMAP plots. We also showed in Supplementary Note 2 some example diagnostic plots of 'bad' parameters, to help users understand how to choose tuning parameters for Sprod.

Validation of Sprod shows its capability to remove noise. Sprod was first applied to the 10X Visium ovarian cancer dataset. For the matched images, we included the keratin and 4,6-diamidino-2-phenylindole (DAPI) IF channels and left out the CD45 IF channel for independent validation. Figure 3a shows the denoised *PTPRC* expression on the slide. The 'discrete' appearance of the original *PTPRC* expression (Fig. 1a) disappeared and the adjusted *PTPRC* expression demonstrates a more continuously changing pattern and, more importantly, is more consistent with the pattern of CD45 IF (Fig. 1a). A scatterplot of CD45 IF intensities and denoised *PTPRC* expression for all spots showed good overall correlation (Fig. 3b; Spearman correlation = 0.7 and Pearson correlation = 0.42), in contrast to the original expression (Fig. 1b). We also overlaid the differences between CD45 IF and *PTPRC* gene expression for each spot with respect to their physical locations. As shown in Extended Data Fig. 4, Sprod has indeed largely reduced the deviance between CD45 IF intensities and *PTPRC* gene expression for most spots. For comparison, we performed the same analyses with SAVER and scImpute[6,7], which are software tools for drop-out removal in scRNA-seq data. Figure 3c shows that SAVER and scImpute achieved only 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). This confirms that Sprod removed the noise through correctly borrowing information from external image/location information, rather than through merely smoothing of the expression data.

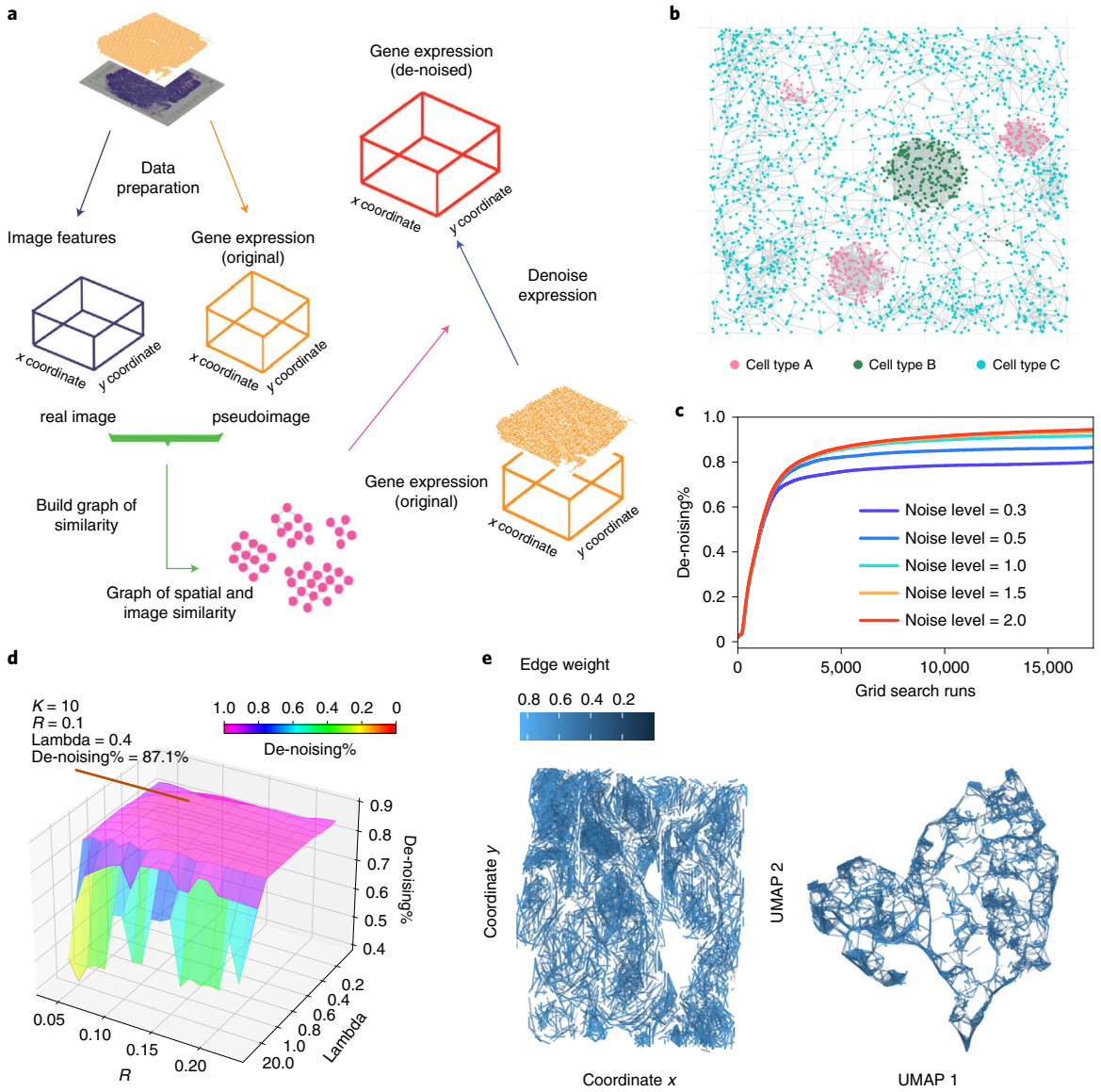


Fig. 2 | Sprod for de-noising of spatially resolved transcriptomics data. a, Cartoon describing the Sprod model, from data preparation and graph building to expressional de-noising. **b**, Simulated dataset: the figure shows the location of the simulated spots and the cell types to which they are assigned. 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. Dots have been subsampled to avoid overcluttering of the presentation. **c**, De-noising% of all parameter combinations tested. The *x* axis shows all parameter combinations, ordered from lower to higher de-noising%. **d**, Visualizing the de-noising% with respect to specific choices of parameters. $K=10$. The *x* and *y* axes show all choices tried for each *R* and Lambda. The *z* axis shows the de-noising%, which is the average of the de-noising% over all the choices of all other parameters. Noise level = 0.5. **e**, Example of the two diagnostic plots generated by Sprod from the 10X breast cancer Visium dataset. Left, spots and edges of the detected graph on the *x*–*y* coordinates. Right, spots and edges of the detected graph in the UMAP space of the image features. The coloring denotes confidence of the edges, with blue referring to high confidence and gray to low confidence.

We next investigated a Visium human lymph node dataset. After correction by Sprod, the adjusted IgD expression demonstrates a spatial pattern more concordant with the H&E-stained images (IgD counts shown in Extended Data Fig. 5 and H&E-stained image in Extended Data Fig. 2). We highlighted one mantle zone in Fig. 3d (red circle in Extended Data Fig. 5 and yellow circle in Extended Data 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 de-noising 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 correlated positively with IgD[2]. In contrast, CD3 spans the perifollicular/interfollicular T cell regions and should be correlated negatively with IgD[3]. For Sprod-corrected expression, CD1c/CD20 showed a much stronger positive correlation with IgD, and CD3 also showed a clearer negative correlation with IgD, compared with the original expression data (Fig. 3e and Extended Data Fig. 6).

Finally, we applied Sprod in the Slide-Seq mouse brain dataset used in Fig. 1. As in Fig. 1f, we calculated whether the overinflation of highly expressed genes still persists after Sprod correction from beads of lower sequencing quality. Unlike the original Slide-Seq data (Fig. 1f), the Sprod-corrected data have almost equal

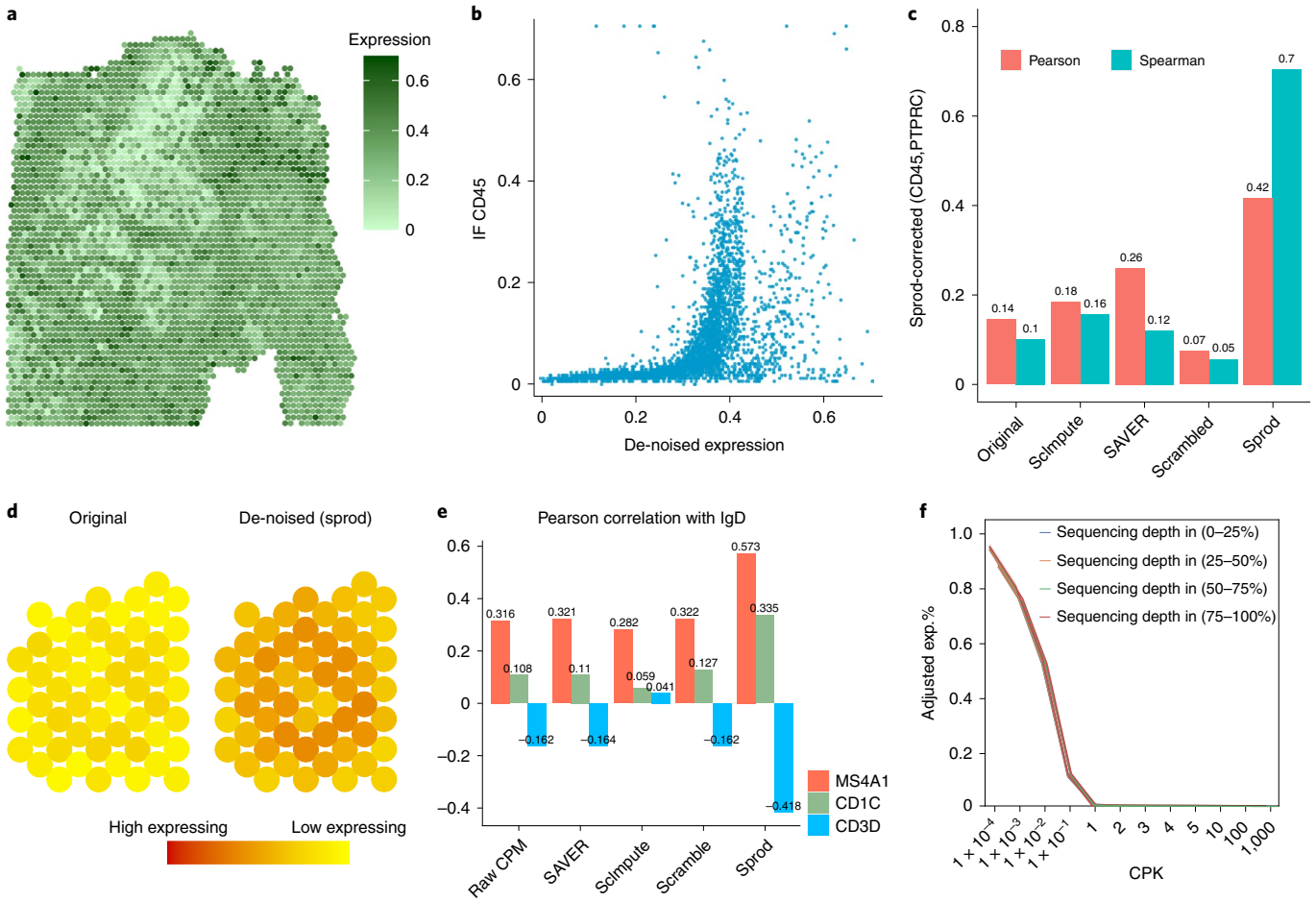


Fig. 3 | Validation of Sprod on real Visium and Slide-Seq datasets. a, Sprod-corrected *PTPRC* expression. Darker colors refer to higher expression. **b**, Scatterplots showing the correlation between CD45 IF and *PTPRC* gene expression, corrected by Sprod. **c**, Spearman and Pearson correlations between 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**, Spatial IgD expression of the mantle zone marked in Extended Data Figs. 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**, Adjusted exp.% of beads in the four bins of different sequencing qualities, in the Sprod-corrected expression data, with the cutoffs to define the adjusted exp.% shown on the *x* axis.

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 a poorer performance in terms of removing artificial differences between the four bins (Supplementary Note 2).

Spatially varying gene detection is more accurate with Sprod. The detection of spatially variable genes (that is, genes whose expression demonstrates certain spatial patterns) is one of the most prevalent analyses carried out on SRT data. Mouse hippocampal Slide-Seq data[1] were examined once again. The hippocampus comprises several regions, including Cornu Ammonis (CA) 1, CA2, CA3 neurons and dentate gyrus (DG). Stickels et al.[1] found several clusters of genes enriched in the dendritic region of CA1 by comparing beads in the proximal neuropil with the soma of neurons[1]. We tested whether the detection of spatially variable genes after Sprod correction is more meaningful. 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 the expression data from of each bead by per-bead library size, with the understanding that the normalized gene counts reflect the relative enrichment of mRNA transcripts within different neuronal regions.

In Fig. 4a, the Slide-Seq beads that correspond to the soma, proximal neuropil and basal neuropil of the hippocampus neurons are highlighted in different colors, where these regions were identified using the same methods by Stickels et al.[1]. In Fig. 4b, the spatial expression of the *Camk2a* gene is displayed in the order of basal neuropil, soma and proximal neuropil on the *x* axis. Figure 4b shows that there are severe drop-outs in the expression of *Camk2a* (dots concentrated at $y < 0.1$, 27.2% of all sequencing beads), and the beads unnaturally break into two groups by expression of *Camk2a*. *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 actively transported from the soma to the neuropils for local translation[12]. The other group of beads has *Camk2a* expression being strictly zero 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 nonzero expression of those beads with drop-outs (still concentrated close to $y < 0.1$, 27.1% of all beads), and the expression of *Camk2a* remains unnaturally dichotomized.

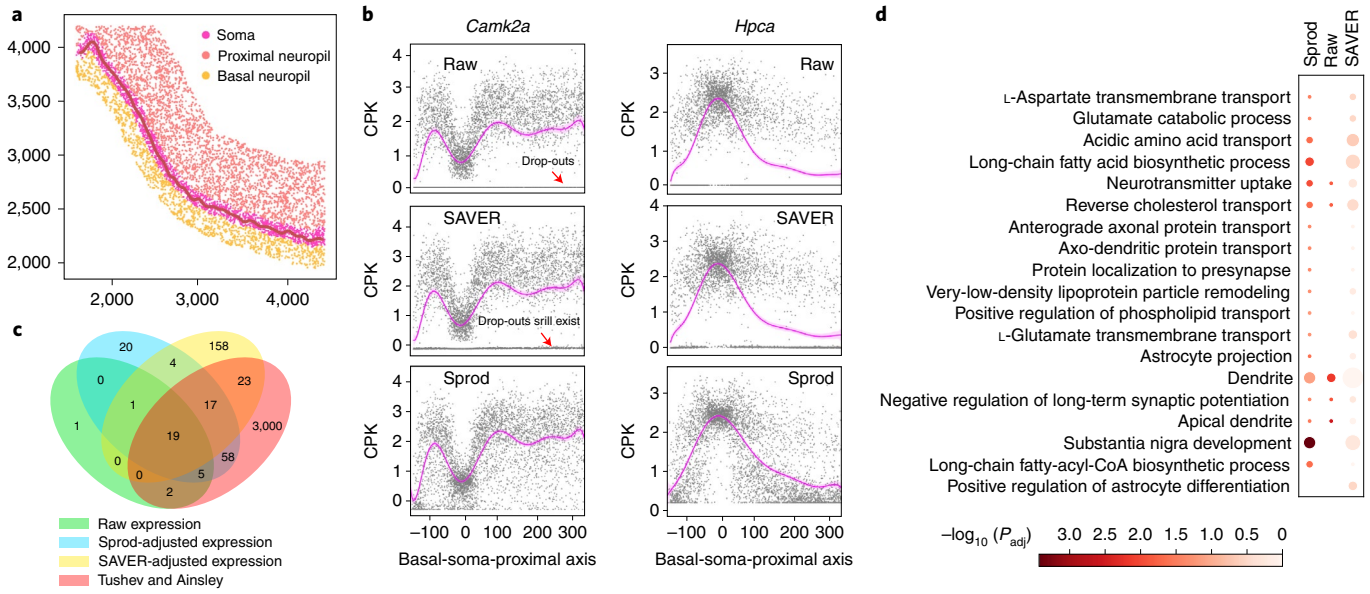


Fig. 4 | Detection of spatially differentially expressed genes is more accurate after de-noising. a, Slide-Seq beads defined as being in basal neuropils, soma and proximal neuropils of the CA1 region (following the definition of Stickels et al.[1]). **b**, Expression of *Camk2a* and *Hpca* in the CA1 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. Data are presented as mean values ± s.d. **c**, Venn diagrams showing the overlap of the genes with differential expression detected from the raw expression data, the Sprod-adjusted data or the SAVER-adjusted data, with the genes that show dendritic enrichment (from Tushev et al. and Ainsley et al.[14,15]). **d**, Enrichment of GO pathways in the genes with stronger expression in the proximal neuropil regions, detected from the raw, the SAVER-corrected and the Sprod-corrected expression data. Circle color refers to the *P* values of the GO analyses (one-sided hypergeometric test, multiple-test adjusted *P* values (P_{adj}) reported); circle size is proportional to the number of genes found in the pathway.

In contrast, Sprod-corrected *Camk2a* expression has nearly completely eliminated the drop-out effects (3.5% of beads with expression <0.1) 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 expression of another gene, *Hpca*, whose expression is higher in the soma but lower in neuropils. Expression of this gene also demonstrated a severe artificial dichotomization caused by drop-outs (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. Finally, we also found that the genes whose expression showed positive or negative correlations with *Camk2a* and *Hpca* in the raw data now have enhanced positive or negative correlations in the Sprod-corrected data (Supplementary Note 2). These results prove that Sprod has indeed addressed drop-outs and corrected noises in the SRT expression data in a biologically meaningful way, rather than merely numerically removing zero counts.

Next, we evaluated the effect of Sprod correction on the detection of spatially variable genes in a genome-wide manner. SpatialDE[13], which was developed specifically for this purpose, was used to detect the genes with stronger expression in the proximal neuropil regions than the soma (Supplementary Note 2). SpatialDE yielded 28 genes in the uncorrected data with differential expression at a false discovery rate (FDR)-adjusted *P* value cutoff of 0.05, and identified 124 genes in the Sprod-corrected data and 222 genes in the SAVER-corrected data. To validate whether these genes show consistency with those previously reported, we cross-referenced Tushev et al.[14] and Ainsley et al.[15], who identified dendritically localized transcripts via microdissection and ribosomal-RNA enrichment, respectively. In Fig. 4c, we showed that the Sprod correction has increased the sensitivity of differential expression analyses greatly, while also retaining good specificity characteristics. The overlap between the differentially expressed gene set from Sprod-corrected data and Tushev et al.[14] and Ainsley et al.[15] achieved a Hypergeometric *P* value of 6.98×10^{-57} and a log odds ratio (logOR) = 2.98). This is in comparison with the raw data (*P* value = 2.72×10^{-19}, logOR = 4.14), and SAVER-adjusted data (*P* value = 1.8×10^{-5}, logOR = 0.57), which indicates that SAVER correction has introduced too many false positives. Overall, Sprod correction achieved the best balance of sensitivity and specificity for differential expression analyses.

We further evaluated the pathways in which these spatially variable genes are enriched. Examination of enriched gene ontology (GO) pathways in the spatially variable genes from Sprod shows that the corrected data lead to the discovery of more genes/pathways (compared with the raw data) that are indicative of synapse functions (for example, 'neurotransmitter uptake') or molecular transport in neurons (for example, 'axo-dendritic protein transport') (Fig. 4d), consistent with the enrichment of these mRNA transcripts in the proximal neuropil regions. For the SAVER-corrected data, none of these pathways reached statistical significance.

Sprod facilitates inferring spatial cellular communications. We further tested whether Sprod's de-noising improved the accuracy of detection of cell-to-cell interactions. The breast cancer Visium dataset previously used (Fig. 2e) was examined. We performed expression clustering of the spots and defined four tumor cell regions (Extended Data Fig. 7), with the remaining areas filled with stromal/immune (SI) cells. Given that each Visium sequencing spot usually contains more than one cell, we removed the 'contamination' from SI cells in the tumor region spots. To do this, we calculated an SI 'contamination score' for each spot in the tumor region using the combined SI gene signature from Wang et al.[16,17]. We removed spots with high 'contamination scores' and, for the remaining spots, we also removed genes in this signature.

CellChat[18] was used to examine the interactions between tumor cells and SI cells. We further divided tumor and nontumor regions

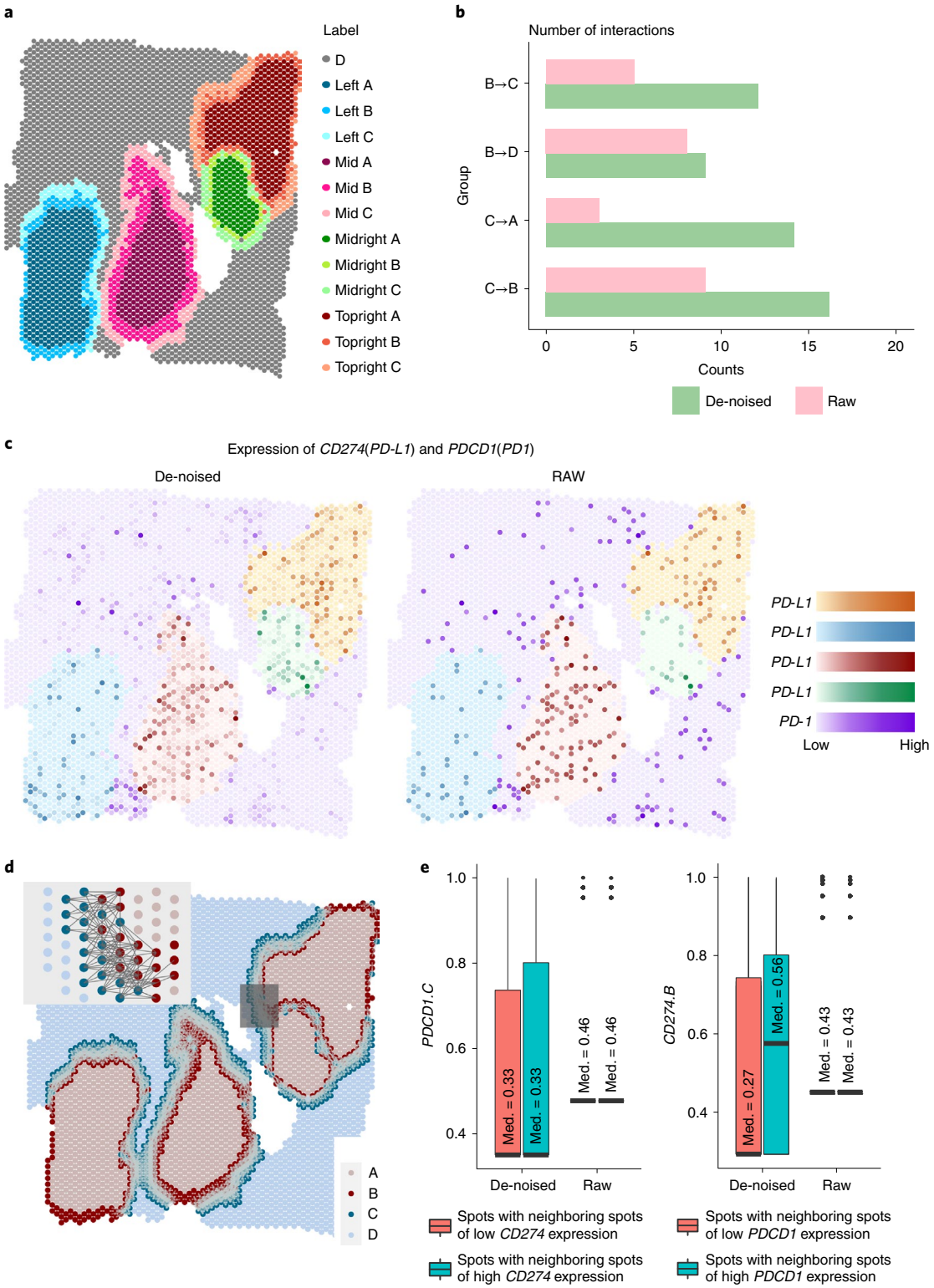


Fig. 5 | Inference of cell-to-cell communication is more accurate with Sprod-corrected expression data. a, Tumor and stroma/immune (SI) regions were both split into subregions (A, B, C and D) that are either close or not close to the tumor-stroma/immune boundaries. **b**, Numbers of CellChat-inferred significantly interacting pathways in the raw and Sprod-corrected expression matrices. **c**, Expression of *PD-1* in the SI regions and *PD-L1* in the tumor regions, for the raw and Sprod-corrected expression matrices, respectively. **d**, Close pairs of spots with one spot in the tumor region and another in the tumor/stroma region. **e**, Expression of *PD-1* in the SI-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*. Bold lines in the boxes refer to median values. Box boundaries represent interquartile ranges, whiskers extend to the most extreme data point (no more than 1.5 times the interquartile range) and the line in the middle of the box represents the median (Med.). $N = 3{,}282$.

into four sections: region A (tumor cells not adjacent to 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) (Fig. 5a). The classification procedure is described in Methods. 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 B→C communication than for B→D and higher for C→B than for C→A (Fig. 5b). In contrast, the raw expression data yields a notably reduced number of interacting pathways overall and more B→D than B→C interactions, which is less interpretable.

In particular, we noted that the pathway of *PD-L1* and *PD-1* in the significantly 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[19–21]. In the denoised data, this pathway was observed to be significant in the A→C and B→C direction, but with stronger confidence in B→C (CellChat probability score $=2.21\times10^{-8}$, P value $<1\times10^{-10}$) than in A→C (probability score $=1.72\times10^{-8}$, P value $<1\times10^{-10}$). We visualized the expression of the ligand *PD-L1*/*CD274* in the tumor cells and the receptor *PD-1*/*PDCD1* in the SI cells in Fig. 5c, for both the raw and denoised data. Overall, the denoised data demonstrated a more obvious coexpression of *PD-L1* and *PD-1* around the interfaces of the tumor–SI regions compared with the raw expression. 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 coexpression of *PD-L1*/*PD-1* 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. 5d). In the pairs from the Sprod denoised data, the expression of tumor region *PD-L1* becomes much higher when neighboring SI regions demonstrate higher PD-1 expression (Fig. 5e; t test P value $=5.6\times10^{-6}$). But when *PD-L1* becomes higher in the tumor cell regions, the expression of *PD-1* is only minimally higher in the SI regions (t test P value $=6.1\times10^{-5}$; no difference in median values). This unidirectional observation is intriguing and also very reasonable, as we anticipate that tumor cells will upregulate *PD-L1* expression in response to the cytotoxic pressure from PD-1$^+$ T cells, but not the other way around. In other words, this analysis revealed a causal relationship between the interactions of the PD-L1/PD-1 pathway. With the raw expression data, however, we do not make this observation (Fig. 5e; P value $=0.005$ and 0.007; no difference in median value in either test). Overall, our above analyses indicated that Sprod also enabled more accurate inference of cell-to-cell communications.

Higher resolution and 3-dimensional SRT datasets. Although the current study focused on 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 probably future SRT techniques as well) achieve much higher resolution than Visium and Slide-seq. Noise in the SRT data is likely to be more prevalent, necessitating even more cautious preprocessing before drawing any conclusions. We provide our results on one Seq-scope dataset in Supplementary Note 2, which proves that Sprod is applicable for such super-resolution SRT data. In addition, 3-dimensional (3D) spatial transcriptomics technologies are also emerging, such as STARmap[22]. With little modification, the graph building model employed by Sprod was easily extended to consider spatial dependency in the 3D space. We provide our results on one example 3D SRT dataset in Supplementary Note 2.

Discussion

We developed Sprod to impute accurate gene expression in SRT data. The existence of extensive noise in SRT data can impact downstream analysis severely and result in substantial bias and misleading conclusions. Sprod took an approach of leveraging the physical location and matched imaging data of SRT to remove such noise, rendering analysis and interpretation of SRT data 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[23]. 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 user-friendly and is capable of readily handling data generated from a wide variety of SRT technologies.

Drop-outs, as in standard scRNA-seq data, can be one source of the inaccuracies in the SRT data. Concerns have been raised in the field that drop-out correction methods for scRNA-seq data may introduce oversmoothing and thus erroneous signals in the scRNA-seq data[24,25]. Since standard scRNA-seq data only provide the gene expression of individual cells, these drop-out removal methods inevitably have to rely on only this information to correct the drop-outs, which may lead to the oversmoothing among cells. For SRT data, Visium, Slide-seq or similar technologies provide the spatial locations and often imaging features associated with the spots/beads. Sprod took the approach of leveraging such external information for reliable imputation, which can avoid oversmoothing. In addition, the latent graph building process in Sprod enforces a neighborhood constraint (R) in de-noising, further preventing the potential problem of oversmoothing.

Sprod is designed as a universal application that can be used with SRT technologies other than those demonstrated in the current study. However, there are differences between SRT 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 $10\,\mu m\times10\,\mu m$ region, whereas each spot in Visium profiles a $55\,\mu m\times55\,\mu m$ 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 SRT technologies that impart a weaker spatial dependency, to introduce fewer constraints on spatial dependency in the graph building process.

An important limitation of our work is that our image feature extractor is based on simple statistics of image features such as channel density or texture. We made this choice, as we want readers to focus on the de-noising of SRT expression data, and also because we already achieved decent performances with this design. Future works can expand the image feature extractor to consider those newer deep learning-based approaches to extract image features, which might lead to better performance. In fact, our Sprod software was engineered to take external image features provided by users, making this choice possible.

Despite the great advancements of SRT technologies, the field must pay close attention to the quality issues of SRT data, which are more challenging than standard scRNA-seq. We envision that Sprod, which has been developed specifically for imputation of accurate SRT gene expression, will become a key first step to empower SRT technologies for biomedical discoveries and innovations.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-022-01560-w.

Received: 23 November 2021; Accepted: 22 June 2022;
Published online: 4 August 2022

References

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**, eabg4755 (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. Lu, T. et al. Overcoming expressional drop-outs in lineage reconstruction from single-cell RNA-sequencing data. *Cell Rep.* **34**, 108589 (2021).
9. 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).
10. 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).
11. 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).
12. 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).
13. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nat. Methods* **15**, 343–346 (2018).
14. 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).
15. 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).
16. Wang, H., Wu, X. & Chen, Y. Stromal-immune score-based gene signature: a prognosis stratification tool in gastric cancer. *Front. Oncol.* **9**, 1212 (2019).
17. 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).
18. Jin, S. et al. Inference and analysis of cell-cell communication using CellChat. *Nat. Commun.* **12**, 1088 (2021).
19. 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**, 1033 (2019).
20. Yuan, C. et al. Expression of PD-1/PD-L1 in primary breast tumours and metastatic axillary lymph nodes and its correlation with clinicopathological parameters. *Sci. Rep.* **9**, 14356 (2019).
21. 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).
22. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
23. 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).
24. 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).
25. Andrews, T. S. & Hemberg, M. False signals induced by single-cell imputation. [version 2; peer review: 4 approved]. *F1000Res.* **7**, 1740 (2018).

Methods

Overview of Sprod. Sprod takes the positional and image features as input, calculates a latent graph from these features, and finally uses the latent graph to smooth out noise in the original SRT expression matrix. In our study, the SRT expression matrices were transformed to the CPM (or CPK) scales and log-transformed. For future users, preprocessing 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 is conducted before Sprod analysis, as necessary. The Sprod model will use only 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 Supplementary Note 1.

Image processing and feature extraction. The Python 'skimage' package (v.0.18) was used for all image processing and feature extraction operations. IF and H&E images were first normalized into an eight-bit format. 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 from which the features were extracted 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[26], including contrast, dissimilarity, homogeneity, ASM, energy and correlation, for each sequencing spot/block in each channel. Specifically, 'skimage.feature.greycomatrix' was used to extract the texture features with 'offset=[1]' and 'angles=[0, π/4, π/2, 3π/4]'. We enabled the option of choosing block- versus spot-level and intensity versus texture image features to the user. For the datasets used in our study, our choices were shown in Supplementary Table 1.

For datasets without matching images, such as those generated by the Slide-Seq, pseudoimages 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[27,28]. The pseudoimage 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 (ten by default) equal-sized subsets, and this process was repeated b (ten by default) times. Sprod de-noising was performed on each of the $n \times b$ 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 subsampling 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 IF matrix, in the same way as Sprod created the 'pseudoimages'.

In particular, the expression matrix, E, was a sum of three parts: (1) expressional variation of cell types, $E1$. This 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, $E2$. Here, we reasoned that spots of the same cell types that were nearby should have more similar expression. We simulated $E0$ from a multivariate normal distribution with mean of zero 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 happens only for spots of the same cell types. $E2$ was scaled so that its variation was comparable with that of $E1$. (3) White noise in expression, N. N was a matrix of white noise generated from independent normal distributions with a mean of zero and equal variance, which is controlled at several different levels (Fig. 2c). We admit that the white noise approach could be overly simplistic, which is a potential caveat of the simulation, but it is hard to obtain the real distribution of the noises. The summed expression matrix is then $E = E1 + E2 + N$. Finally, the E matrix was transformed by an exponential function and scaled such that its distribution mimics the distribution of typical SRT 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 1,000), 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 t-SNE-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 adjusted based on the diagnostic plots (Fig. 2e). The parameters used in all datasets involved in this work are listed in Supplementary Table 2.

Defining tumor and SI spots/regions that are close neighbors. In the breast cancer 10X Visium dataset, the spots were clustered based on gene expression, examined manually and merged to the tumor and SI 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 SI cell contamination. Specifically, a stromal gene module and an immune gene module were defined by the union of the gene signatures for SI cells defined by Wang et al.[16,17]. 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. The tumor region spots with the top 5% of the module scores were then filtered out. For the remaining tumor region spots, these immune- and stromal-related genes were also filtered out.

In Fig. 5a, we further split the tumor regions into two subsets: tumor regions not close to SI cells (A) or close to SI cells (B). For each spot in the tumor region, we draw a circle based on a radius of twice the size of the spacing between two neighboring spots in the Visium spot lattice. We count the number of SI region spots in this circle. This spot will be classified as 'B' if its closest neighboring spots in the SI 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 adjusted slightly to ensure a smooth appearance of the classified 'A' and 'B' tumor regions. The SI region spots were classified similarly into 'C' (close to tumor region) and 'D' (not close to tumor region).

In Fig. 5d, 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 were conducted in the R or Python programming languages. UMAP and t-SNE were conducted by the UMAP v.0.2.7.0 or Rtsne v.0.15R packages. Drop-out removals were conducted by SAVER v.1.1.2 and scimpute v.0.0.9. For all boxplots appearing in this study, box boundaries represent interquartile ranges, whiskers extend to the most extreme data point (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, cell-to-cell communications were inferred by the CellChat v.1.1.3 package.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

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 standard 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. The doi is https://doi.org/10.5281/zenodo.604775229.

References

26. Haralick, R. M., Shanmugam, K. & Dinstein, I. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **3**, 610–621 (1973).
27. Zhang, Z., Xiong, D., Wang, X., Liu, H. & Wang, T. Mapping the functional landscape of T cell receptor repertoires by single-T cell transcriptomics. *Nat. Methods* **18**, 92–99 (2021).
28. Adossa, N. A., Rytkönen, K. T. & Elo, L. L. Dirichlet process mixture models for single-cell RNA-seq clustering. *Biol. Open* **11**, bio059001 (2022).

Acknowledgements

We acknowledge the ENCODE Consortium and the ENCODE production laboratories that generated the eCLIP datasets used in our study. We acknowledge J. Johnson for providing input on the interpretation of the mouse Slide-Seq data. This study was supported by the National Institutes of Health (NIH) (5P30CA142543 to T.W., G.X. and Y.X., 1R01CA258584 to T.W., U01AI156189 to T.W. and Y.X., R01DE030656 to G.X., R01GM141519 to G.X., R01GM140012 to G.X., U01CA249245 to G.X., R35GM136375 to Y.X., 2P50CA070907 to T.W., Y.X. and G.X., R01AG075582 to L.W., 3U01AI156189-01S1 to T.W.), National Science Foundation (NSF DMS-2009689 to L.W.), and Cancer Prevention Research Institute of Texas (CPRIT RP190208 to T.W. and RP190107 to G.X.).

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.

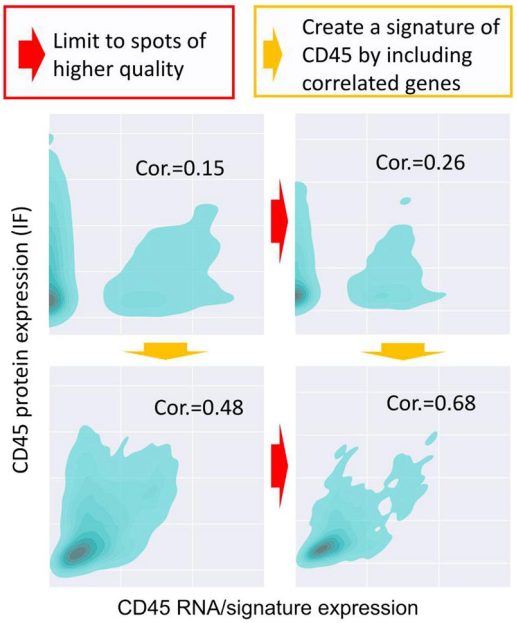
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41592-022-01560-w.

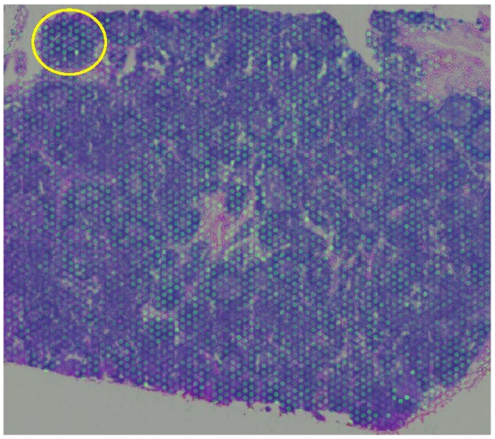
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-022-01560-w.

Correspondence and requests for materials should be addressed to Li Wang or Tao Wang.

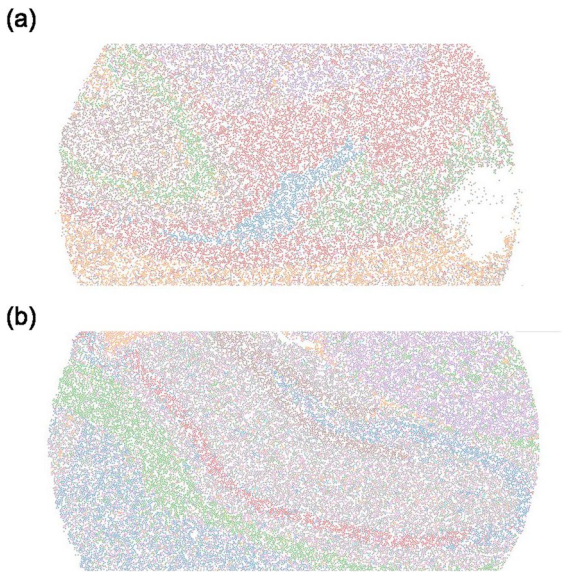
Peer review information *Nature Methods* thanks Nikos Karaiskos and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary handling editor: Lin Tang, in collaboration with the *Nature Methods* team.



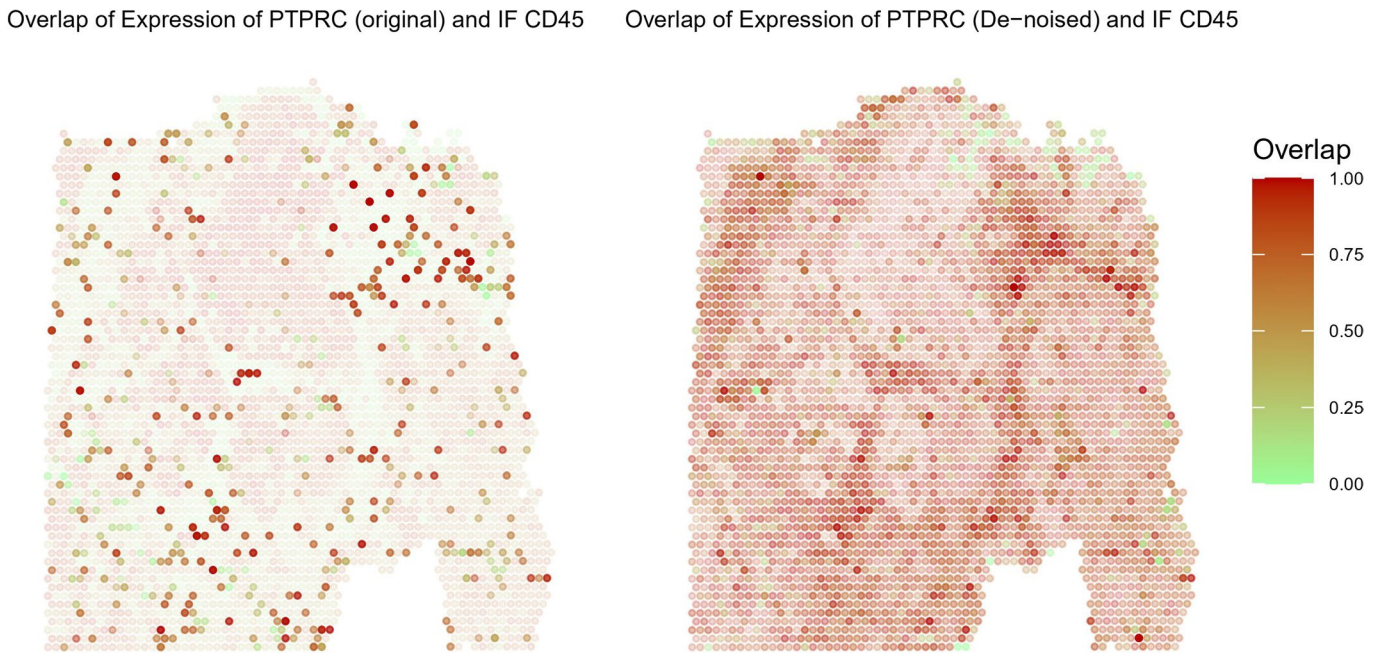
Extended Data Fig. 1 | Correlation between *PTPRC* RNA expression (targeted) and CD45 protein expression (IF). Red arrows mean that the results are limited to spots of higher quality. Yellow arrows mean that the single gene of *PTPRC* is replaced by a signature of *PTPRC* by including correlated genes.



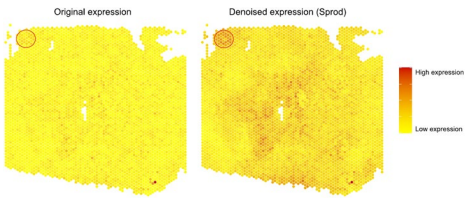
Extended Data Fig. 2 | Overlaying the un-corrected IgD gene expression on the H&E stained image in the 10X Visium human lymph node dataset. The whole slide is shown, with the three examples of Fig. 1d picked from this slide. The yellow circle marks the mantle zone to be highlighted in Fig. 3d.



Extended Data Fig. 3 | Gene expression clustering of the beads in the mouse brain Slide-Seq dataset. Gene expression clustering of the beads in the mouse brain Slide-Seq dataset reflects the multi-cellular structures of mouse brain hippocampus. **a**, Slide-seq dataset puck 200306_03 and **b** puck 200115_08 by Stickels et al.[1].

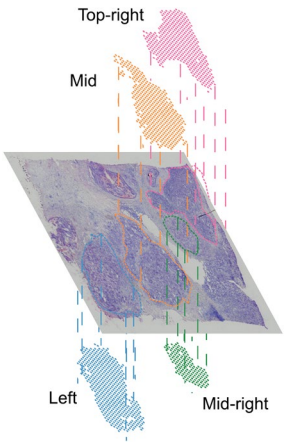


Extended Data Fig. 4 | Deviances between CD45 IF intensities and the expression levels of *PTPRC* (left: original, right: denoised). CD45 IF intensities and *PTPRC* expression values were normalized and distributionally warped to the same scale so they can be directly compared. The differences between CD45 IF and *PTPRC* on each spot are denoted by color. Red refers to small differences and green refers to larger differences.



Extended Data 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.

Extended Data 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.



Extended Data Fig. 7 | Extraction of four tumor regions. The four tumor regions (blue, green, orange and red) that were extracted, according to expressional clustering and concordance with the H&E stained slide.

nature research

Corresponding author(s): Tao Wang

Last updated by author(s): May 13, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Python (3.7), R (4.0.2), skimage (0.18)
Data analysis	Sprod(https://github.com/yunguan-wang/SPROD, v1.0, 10.5281/zenodo.6047752). umap v0.2.7.0, Rtsne v0.15, SAVER v1.1.2, scimpute v0.0.9, Slingshot v1.8.0, Monocle3, pseudotimeDE v0.9.0, clusterprofiler v3.18.1, CellChat v1.1.3

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on 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-ovariancancer-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.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation. We used all data
Data exclusions	No data were excluded
Replication	All analyses were replicated with a fixed random seed for at least once. All replicates are successful
Randomization	Not applicable. This is not a randomized clinical trial
Blinding	Not applicable. This is not a randomized clinical trial

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging