Randomized Spatial PCA (RASP): a computationally efficient method for dimensionality reduction of high-resolution spatial transcriptomics data

Ian K. Gingerich[1,2,*], Brittany A. Goods[2], and H. Robert Frost[1,*]

[1]Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Hanover, NH, USA
[2]Thayer School of Engineering, Dartmouth College, Hanover, NH, USA
[*]Corresponding authors: ian.gingerich.gr@dartmouth.edu, rob.frost@dartmouth.edu

Abstract

Spatial transcriptomics (ST) provides critical insights into the complex spatial organization of gene expression in tissues, enabling researchers to unravel the intricate relationship between cellular environments and biological function. Identifying spatial domains within tissues is essential for understanding tissue architecture and the mechanisms underlying various biological processes, including development and disease progression. Here, we present Randomized Spatial PCA (RASP), a novel spatially aware dimensionality reduction method for spatial transcriptomics (ST) data. RASP is designed to be orders-of-magnitude faster than existing techniques, scale to ST data with hundreds-of-thousands of locations, support the flexible integration of non-transcriptomic covariates, and enable the reconstruction of de-noised and spatially smoothed expression values for individual genes. To achieve these goals, RASP uses a randomized two-stage principal component analysis (PCA) framework that leverages sparse matrix operations and configurable spatial smoothing. We compared the performance of RASP against five alternative methods (BASS, GraphST, SEDR, spatialPCA, and STAGATE) on four publicly available ST datasets generated using diverse techniques and resolutions (10x Visium, Stereo-Seq, MERFISH, and 10x Xenium) on human and mouse tissues. Our results demonstrate that RASP achieves tissue domain detection performance comparable or superior to existing methods with a several orders-of-magnitude improvement in computational speed. The efficiency of RASP enhances the analysis of complex ST data by facilitating the exploration of increasingly high-resolution subcellular ST datasets that are being generated.

1 Introduction

Spatial transcriptomics (ST) assays measure gene expression at thousands to hundreds-of-thousands of spatial locations within intact tissue sections, enabling researchers to characterize transcript abundance *in vivo* at a resolution varying from small cell groups to subcellular structures [1, 2]. ST technology facilitates the exploration of the spatial organization of gene expression, providing valuable insights into the cellular microenvironment and functional states of individual cells [3]. Two primary types of ST technologies are commercially available: high-throughput sequencing technologies (sST) and single-molecule fluorescent in situ hybridization or imaging-based technologies (iST) [4, 5]. High-throughput sequencing technologies, such as 10x Genomics' Visium, capture mRNA transcripts at multiple sites or spots on a slide. The original 10x Visium platform offers a resolution of 55 µm, with each spot containing between $1-10$ cells [6]. Emerging technologies, such as Visium HD, promise higher resolution of 2 µm, approaching single-cell levels, while alternatives like Slide-seq V2 capture transcripts with a resolution of 10 µm, also nearing single-cell resolution [7, 8]. On the other hand, iST technologies, which reach single cell resolution, include platforms such as MERFISH/MERSCOPE (Vizgen), seqFISH, osmFISH, COsMx, and 10x Xenium [9–13]. A notable advancement in recent years has been the introduction of iST platforms that support the use of formalin-fixed, paraffin-embedded (FFPE) tissue samples [9, 12, 13]. This capability allows researchers to leverage archival tissue banks, eliminating the need for complicated sample harvesting workflows, leading to increased use of iST technologies. Additionally, the probe sets used in iST technologies have expanded significantly, with systems like 10x Xenium developing probe sets comprising up to 5,000 individual probes [14]. This increase in probe capacity helps to bridge the trade-off between sequencing depth and resolution, allowing for more comprehensive profiling of gene expression. As costs decrease and technology advances, studies utilizing ST are transitioning from the analysis of small sample sizes (2-5 samples) to more extensive investigations [2, 13, 15–19].

Many research efforts are now generating multi-slide, multi-timepoint ST datasets that encompass millions of cells across dozens of tissues [2]. This increase in dataset size necessitates the development of novel computational tools that are computationally efficient and capable of leveraging spatial and transcriptomic information contained in these datasets.

The rich data provided by ST allows for the investigation of tissue at multiple scales, and new computational approaches must be flexible to these needs. On the cellular level, cell type assignment facilitates visualization of the spatial distribution and organization of different cell types, enhancing our structural and functional understanding of cells in their native tissue context [20, 21]. At a larger scale, the identification of tissue domains aids in comprehending the complex microenvironment of tissues. Identification of both cell types and spatial domains is achievable through recently developed computational tools. For cell type identification, popular single-cell RNA sequencing (scRNA-seq) tools such as Seurat and Scanpy are frequently employed [22, 23]. These methods perform principal component analysis (PCA) and clustering on gene expression data while disregarding spatial information. However, advancements have led to the emergence of methods that incorporate spatial context, including BASS and FICT [24, 25]. For spatial domain detection, several methodologies are currently available, including BASS, SEDR, GraphST, STAGATE, and spatialPCA, among others [24, 26–29]. Each of these methods utilizes different model architectures to embed spatial and gene expression information into a low-dimensional space, from which latent variables are clustered to identify spatial domains. ST data is often sparse, and de-noising and imputation of poorly captured transcripts is a desirable feature for many researchers. To this end, many spatially aware dimensionality reduction methods provide functionality to de-noise and impute missing values [30–33].

Despite their advantages, current methods have notable limitations. Most importantly, while these methods perform well on a single ST slice with relatively few profiled locations (such as those provided by 10x Genomics' Visium platform), they tend to be computationally expensive. For instance, methods that can be executed locally on a Visium dataset often exhibit slow processing speeds, particularly when applied to iST datasets in which cell numbers frequently exceed 100,000. With an increase in iST dataset generation, there is an acute need for computationally efficient and spatially-aware dimensionality reduction algorithms. Additionally, many of the existing methods operate as inflexible, complex models that do not easily support user customization of spatial smoothing or readily accommodate the addition of non-transcriptomic covariates, such as features derived from hematoxylin and eosin (H&E) staining images, protein abundance measurements, or cellular density metrics. Addressing these limitations is essential for deriving meaningful biological insights from large-scale ST datasets.

In this paper, we present Randomized Spatial PCA (RASP), a novel computational method designed to be orders-of-magnitude faster than existing techniques, scalable to ST data with hundreds-of-thousands of locations, and able to support non-transcriptomic covariates while remaining easy to understand and implement. RASP employs randomized linear algebra techniques for dimensionality reduction, ensuring efficient handling of very large datasets [34]. Following this reduction, RASP performs spatial smoothing using the coordinates returned from the ST platform. One of the key features of RASP is its support for the optional integration of non-transcriptomic covariates, such as cellular density, cell size, or sequencing depth, providing researchers with greater analytical flexibility. We have tested RASP on four real-world datasets, including one from the 10x Genomics Visium platform and three from subcellular imaging technologies: Stereo-Seq, MERFISH, and 10x Xenium. Our results demonstrate that RASP achieves tissue domain detection performance comparable or superior to existing methods with a runtime that is several orders-of-magnitude faster.

2 Results

2.1 RASP overview

The RASP method performs dimensionality reduction and spatial smoothing on normalized ST data, with the option to incorporate location-specific covariates. The randomized two-stage PCA approach used by RASP is designed to optimize computational efficiency while enabling the flexible and efficient analysis of high-resolution ST datasets. Specifically, RASP utilizes a sparse matrix representation of the ST data, conducts dimensionality reduction via randomized PCA, and generates spatially smoothed principal components (PCs) using an inverse distance matrix that is sparsified based on a k-nearest neighbors (kNN) threshold (Fig. 1A). The integration of location-specific covariates into the latent variables is achieved through an optional second-stage randomized PCA performed on a matrix formed by appending the covariates to the smoothed PCs from the first stage. The spatially

smoothed PCs produced by RASP are useful for a variety of downstream analyses, including clustering and reduced rank reconstruction of gene-level expression. See Algorithm 1 and Section 4.1 for details. Note that by default, RASP utilizes 20 PCs, although this is a user tunable paramter. See Extended data Fig 8. for effects of variable PC number on RASP performance.

2.2 RASP evaluation on real data

We applied RASP to four publicly available ST datasets generated using diverse techniques and resolutions on human and mouse tissues. These datasets include the mouse ovary dataset (Vizgen MERFISH technology), the mouse olfactory bulb dataset (STOmics Stereo-seq technology), the human breast cancer tumor dataset (10x Genomics, Xenium technology), and the human DLPFC dataset (10x Genomics, Visium technology)(Fig. 1B) and Table 1. For details regarding the preprocessing steps, such as filtering criteria and normalization, please refer to Sections 4.4.1-4.4.3, and SI Sections 1.2-1.3.

To evaluate the RASP method, we clustered locations using the spatially smoothed principal components (PCs) output by RASP. Cluster quality was assessed through spatial continuity and compactness quantified by the CHAOS score (SI 1.3), spatial autocorrelation quantified by Moran's I (SI 1.4) and when applicable, the Adjusted Rand Index (ARI) (SI 1.2) compared to ground truth annotations. A sensitivity analysis was conducted on the key parameters in the RASP pipeline, with a specific focus on the kNN based distance threshold for constructing the sparse distance matrix and the β power parameter used to compute inverse distances (see Figure 1A). This sensitivity analysis was performed using a one-parameter-at-a-time strategy, allowing us to thoroughly investigate the influence of individual parameters on clustering outcomes.

Performance comparisons were made between RASP and five other spatially informed dimensionality reduction or spatial domain detection methods: BASS [24], GraphST [26], SEDR [27], spatialPCA [28], and STAGATE [29] as well as performing standard PCA. For each comparison, we adhered to the respective GitHub tutorials to run these methods for spatial domain and cell type detection. Clustering of the latent dimensions found by each method was performed using Mclust [35], Louvain [36], Leiden [37], and Walktrap [38] algorithms, where appropriate. The size of the human breast cancer dataset posed significant computational challenges, requiring allocations of arrays exceeding 1TB, leading to unsuccessful runs for some methods.

Table 1: Datasets used for RASP evaluation

Platform	**Tissue**	**Organism**	**n cell/n location**	**n genes**	**Ground truth annotation?**
MERFISH	Ovary	Mouse	43,038	228	True
Stereo-Seq	Brain (Olfactory bulb)	Mouse	19,109	14,367	False
10x Xenium	Breast cancer	Human	565,916	541	True
10x Visium	Brain (DLPFC)	Human	3,638	15,124	True

2.3 RASP identifies biologically relevant structures in heterogeneous mouse ovary MERFISH data

We selected this MERFISH dataset due to its complex tissue structure and cellular resolution, characterized by spatially distinct and spatially heterogeneous cell types throughout the tissue. The data was obtained from Huang et al., [15] and includes ground truth cell type annotations based on histology and differential gene expression (Fig. 2A). The annotations include luteal, luteinizing mural, granulosa, theca, and stromal cells, which show clear spatial organization, as well as endothelial, epithelial, and immune cells, which are heterogeneous in terms of their spatial distribution across the tissue [39], making annotation of this tissue with automatic detection methods challenging.

For this dataset, RASP using Louvain clustering achieved the highest ARI of 0.69 across all parameters tested, followed by standard PCA (ARI of 0.58) and SEDR (ARI of 0.57) (see Fig. 2A and 2C). We employed three different clustering methods to assign labels to cells in the dataset. The Louvain and Leiden algorithms demonstrated the best performance, while Mclust yielded the worst results. It is important to note that the BASS method utilizes Mclust as a prior in its Bayesian framework but does not support a user-defined clustering algorithm like the other methods.

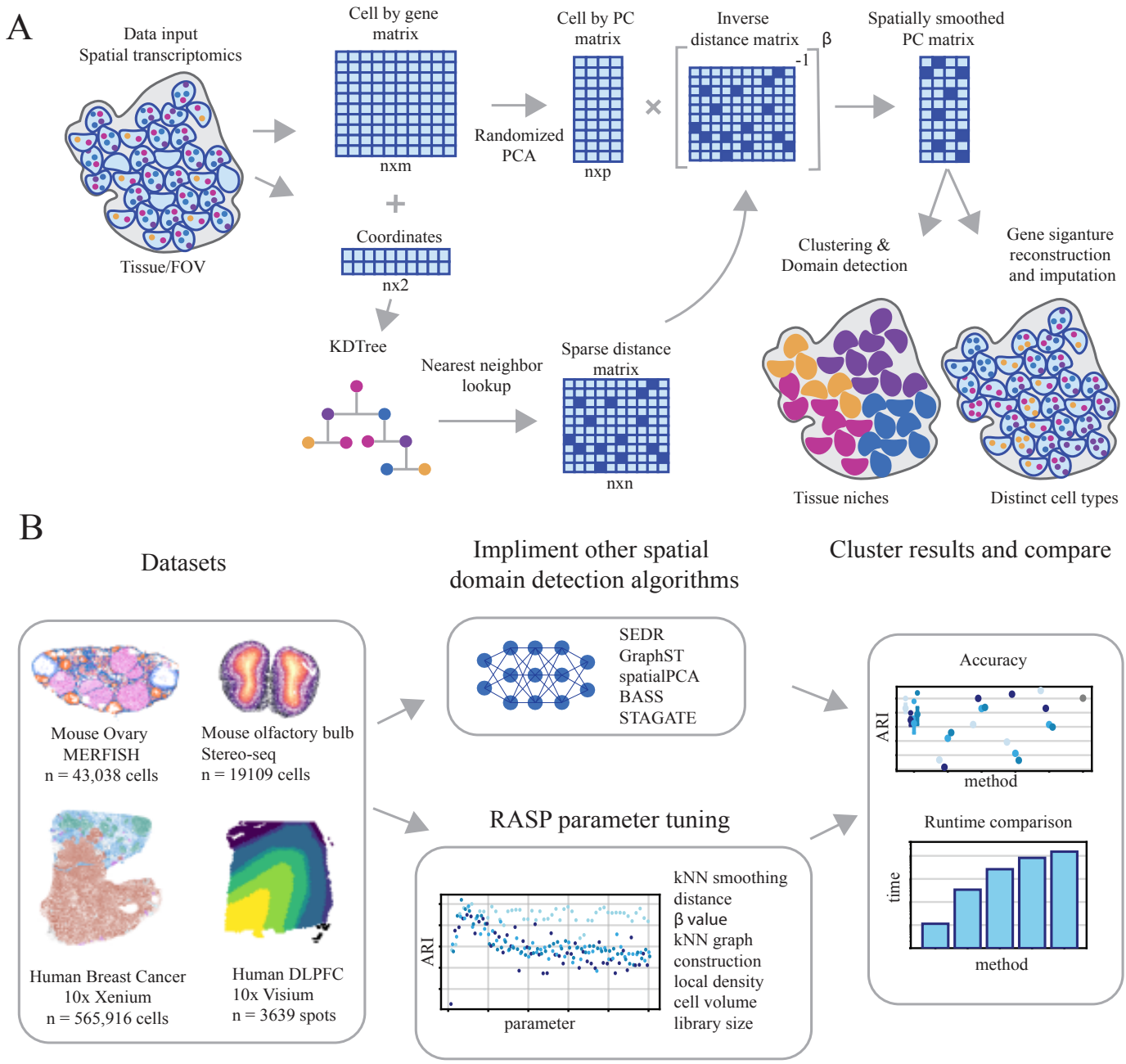

Figure 1: **RASP overview. A: Method workflow.** RASP takes as input a cell-by-gene expression matrix and spatial coordinates from a ST experiment. The expression matrix is reduced via randomized PCA and spatially smoothed by a sparse inverse distance matrix. Non-transcriptomic covariates can be added at this stage (not pictured). The output of RASP can be used for downsteam analyses including cell type annotation and spatial domain identification. **B: Evaluation pipeline.** RASP was tested on four ST datasets each generated using a different platform (MERFISH, Stereo-seq, Xenium, Visium). RASP results were compared against standard PCA and alternative spatially aware dimensionality reduction models (SEDR, GraphST, spatialPCA, BASS, and STAGATE) for both accuracy and speed.

Relative runtimes show that RASP is between one and three orders-of-magnitude faster than competing spatially aware methods on this dataset (Fig. 2B). The second best performing method, SEDR, is an order-of-magnitude slower than RASP, while BASS is almost four orders-of-magnitude slower. For achieving accurate annotations, RASP performance diminishes with increasing kNN thresholds used for smoothing the PCs; this is highlighted by decreasing ARI scores at larger kNN values (Fig. 2D). This result is not unexpected for cell type annotation, but it is noteworthy that the local smoothing employed by RASP in this task, essentially borrowing information from just a few neighboring cells, has a marked increase in performance over normal PCA for clustering-based label assignment. For this task we found that $\beta = 2$ was the best performing inverse distance power value. As the β value increased, the kNN distance threshold's impact on ARI was blunted, with comparable performance across the entire range of kNN thresholds sampled. Conversely, with $\beta = 0$, the kNN threshold dramatically affected RASP performance (Fig. 2D, Extended Data Fig.1).

To demonstrate the use of RASP for reduced rank reconstruction of individual gene expression profiles, we analyzed the spatial signature of the luteinizing hormone/chorionic gonadotropin receptor (Lhcgr) given its essential role in ovarian biology and ovulation (Fig. 2E). Lhcgr is a G-protein coupled receptor that binds luteinizing hormone and chorionic gonadotropins, and is expressed on granulosa cells, theca cells, and luteal cells [40–42]. The expression of this gene in the original dataset (Fig. 2E, left plot) was minimal and showed little spatial patterning. After reconstruction using RASP, the expression is clearly localized to the luteinizing mural and luteal cell populations (Fig. 2E, middle and right plots). For this example, we utilized the distances to only 3 kNN, which results in very similar signatures between normal PCA-based reduced rank reconstruction and RASP.

2.4 Incorporation of covariates improves RASP annotation of mouse ovary ST data

An important feature of the RASP method is support for integrating non-transcriptomic covariates into the reduced dimensions. The two-stage PCA approach used by RASP has two primary benefits relative to a single PCA performed on the merge of transcriptomic and non-transcriptomic variables. First, it avoids diluting the covariate signal within the high-dimensional transcriptomic data, and, second, it enables decoupling of the spatial smoothing applied to the expression data from any smoothing performed on the covariates. To evaluate the performance of the RASP covariate integration approach, we explored the impact of three different covariates on analysis of the mouse ovary data: local cell density, transcriptomic library size, and cell volume. For each covariate, we performed parameter sweeps over the kNN parameter using three different configurations of RASP: 1) standard, single-stage RASP without any covariates (i.e., the configuration used to generate the results in Figure 2), 2) single-stage RASP executed on the merge of the ST expression data and covariates, and 3) two-stage RASP with the covariates added to the output from the first stage (see Fig. 3A (bottom)). Note that for these experiments, the β parameter was held constant at $\beta = 2$ as this was the optimal value for this dataset (see Fig. 2D, and Extended Data Fig 1). For the local density covariate, spatial smoothing was applied to the PC matrix, but not the local density variable, as local density inherently contains spatial information. The library size and cell volume covariates where spatially smoothed by the same inverse distance weighting as the PCs matrix, prior the second round of dimensionality reduction.

Addition of all three covariates impacted the ARI value, with the employed clustering algorithm determining the magnitude and direction of the ARI change. The ARI acheived using Mclust improved regardless of the covariate added, with the single-stage version of RASP equal to or outperforming the two-stage version in most cases (see Fig. 3B (left column)). The Mclust ARI without covariates was 0.39 and adding covariates improved the score to between 0.45 and 0.5. The ARI achieved using Leiden was highest for the two-stage version of RASP in almost all cases, while performance for the single-stage version was covariate dependent, with the one-stage addition of local density decreasing ARI values, and the addition of library size and cell volume having a mixed effect (see Fig. 3B (right column)). Notably, Leiden clustering with cell volume achieved a top ARI of 0.71, the highest attained for this dataset. Louvain-based classification was hindered by the addition of local density, with ARI values dropping across all kNN values, while the affects of library size and cell volume fluctuated depending on the kNN value.

Overall, these results indicate that integration of covariates can enhance RASP performance but that careful consideration is needed regarding the selection of covariates, clustering algorithm, and form of spatial smoothing. Importantly, the two-stage approach used by RASP avoids covariate dilution and gives users more control over spatial smoothing.

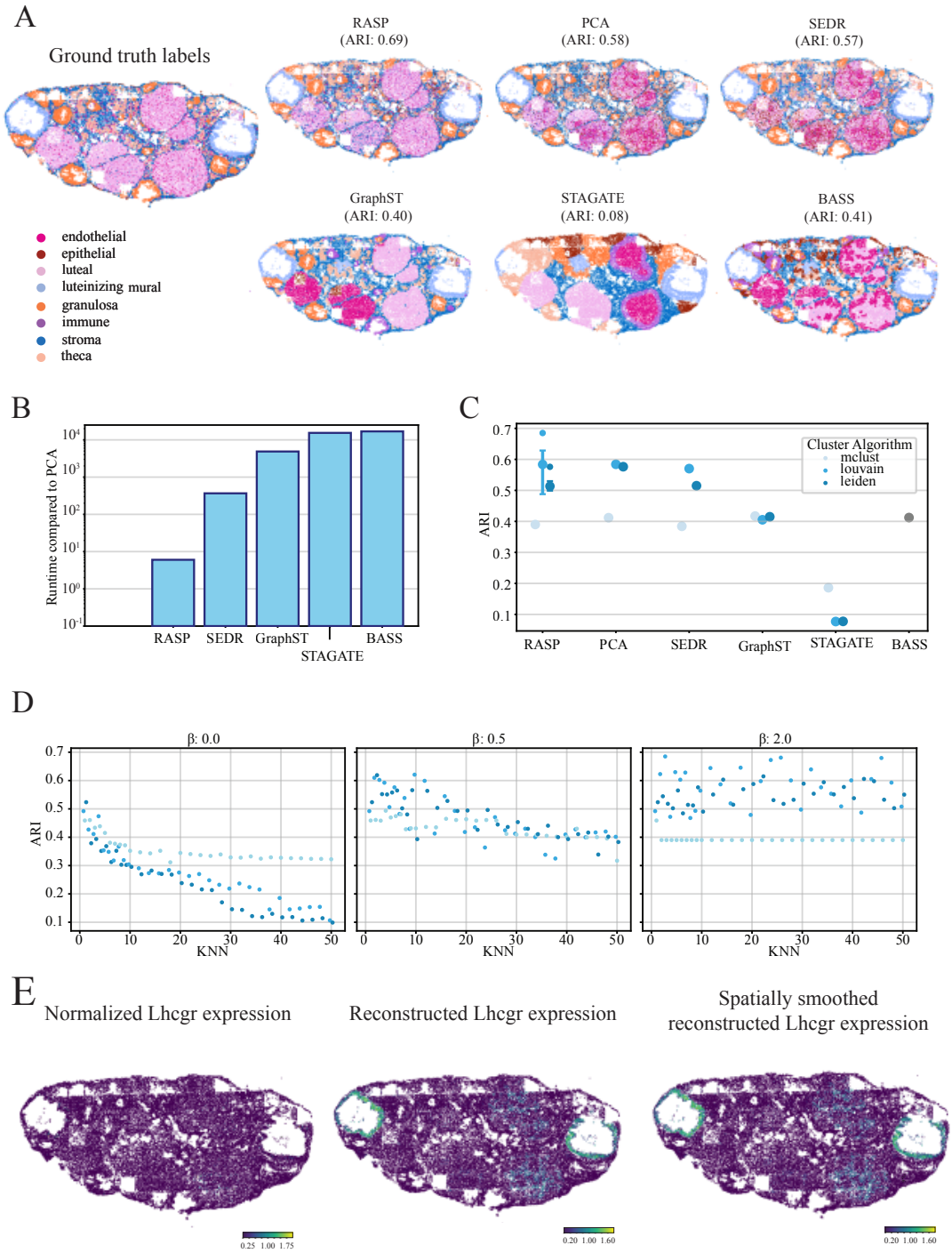

Figure 2: **Mouse ovary analysis. A:** Ground truth cell type annotations (top left) and corresponding predictions for RASP, normal PCA, and other methods **B:** Quantification of runtime for all methods compared to normal PCA. **C:** Quantification of ARI for all methods, different colors indicate the clustering algorithm used to assign labels. The interquartile range and median ARI values at default RASP parameters are shown (kNN = 3-10, $\beta = 2$), as well as the maximum ARI value achieved by RASP. **D:** ARI values for RASP with inverse distance raised to the β of 0 (left), 0.5 (middle), and 2 (right), plotted against kNN. Colors indicate clustering algorithm. **E:** Normalized Lhcgr expression (left), reduced rank reconstructed Lhcgr expression (center), and spatially smoothed reduced rank reconstructed Lhcgr expression (right).

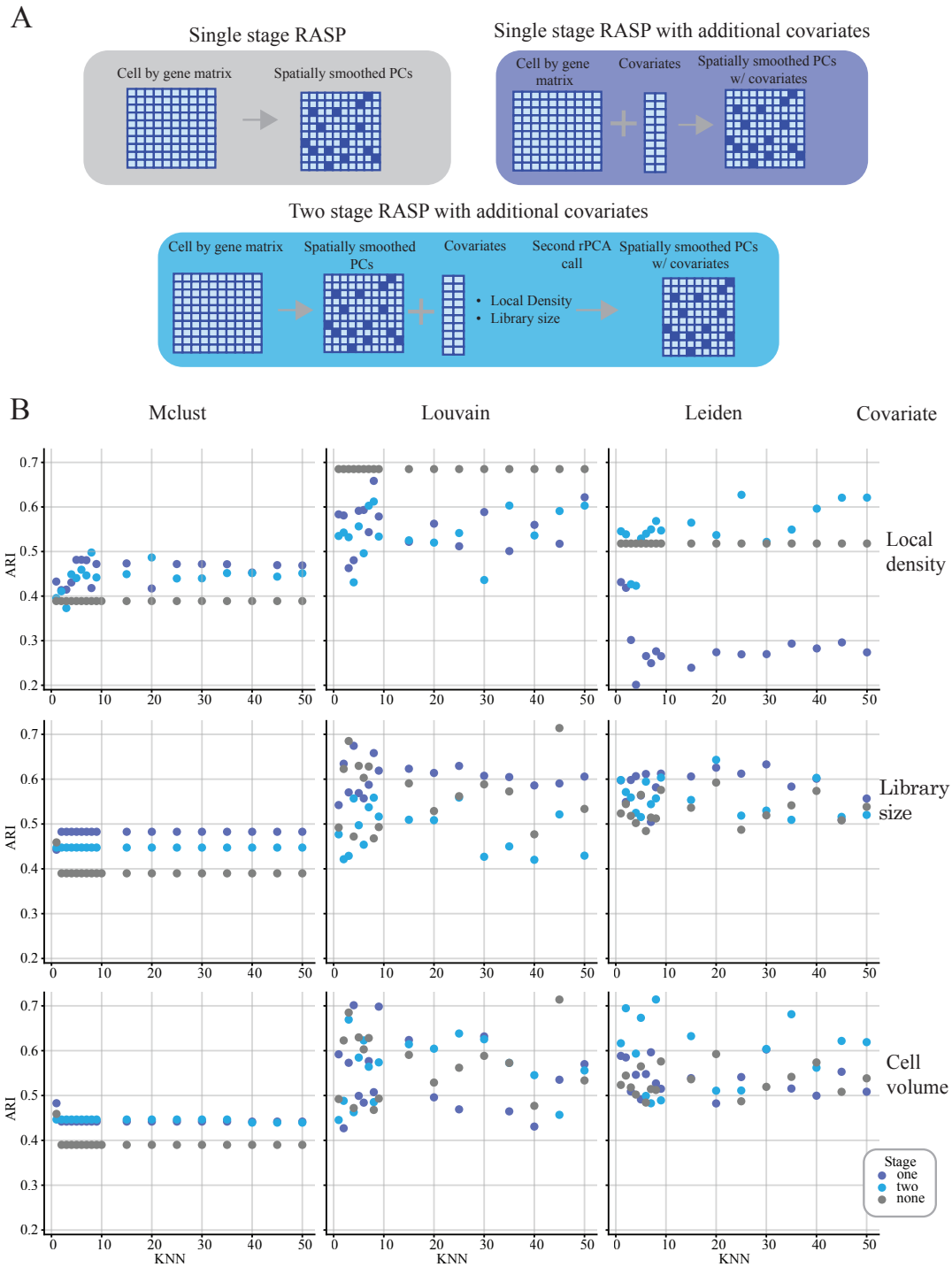

Figure 3: **Covariate analysis. A:** Schematic depicting RASP model architecture with no covariates (top left), single-stage RASP architecture with an additional covariate (top right), and two-stage RASP model architecture with an additional covariate. **B:** ARI values for no covariate, one-stage RASP, and two-stage RASP plotted against kNN. Columns show different clustering algorithms, and rows capture the type of covariate: local cell density (top), cell library size (middle), and cell volume (bottom).

2.5 RASP shows comparable performance to alternative methods for cortical layer identification using human DLPFC Visium data

The dataset utilized for this analysis was generated using the 10x Visium platform by the LIBD human Dorsolateral Prefrontal Cortex (DLPFC) project [43] and includes manual cortical layer annotations. This dataset has served as a gold standard benchmark for spatial domain detection algorithms due to its clear neuroanatomical boundaries, which provide a reliable ground truth, and a relatively small number of profiled tissue locations that facilitates the evaluation of computationally complex algorithms. For benchmarking, we selected slice #151,673 from the LIBD dataset, which consists of seven spatial domains including six cortical layers and white matter [43].

For this dataset, RASP and spatialPCA emerged as the best-performing methods, achieving ARIs of 0.62. SEDR and BASS followed closely, with both having an ARI of 0.60, while GraphST obtained slightly lower scores of 0.57 (Fig. 4A and 4D). Importantly, RASP exhibited a computational cost comparable to standard PCA but with a substantially higher ARI. RASP was found to be between one and three orders-of-magnitude faster than other spatially smoothed methods, highlighting its efficiency and accuracy for processing ST data (Fig. 4B). For optimal RASP performance, a kNN threshold between 3 and 10 was the most effective for this task; larger kNN thresholds resulted in declining scores, which is expected as larger values result in smoothing across cortical layer boundaries, resulting is overly large domain identification (Fig. 4D). Interestingly, we found that all evaluated β values generated similar ARI vs. kNN threshold trends. The Mclust algorithm was relatively invariant to both the β and kNN threshold parameters, showing consistent performance across the entire parameter space. Both Louvain and Leiden clustering algorithms showed a marked reduction in performance as the kNN threshold increased, regardless of the β value, with $\beta = 2$ showing marginally better performance than the other β values (Fig. 4, Extended Data Fig. 3A).

To demonstrate the use of RASP for reduced rank reconstruction, the TMSB10 (thymosin beta 10) gene was reconstructed [40] (Fig. 4E). Importantly, the spatially smoothed reconstruction of TMSB10 is isolated to layer 4 with substantially reduced noise relative to normalized expression and a non-smoothed reconstruction highlighting the utility of RASP for spatially informed gene-level analyses.

2.6 RASP effectively characterizes tissue structures in mouse olfactory bulb Stereo-seq data

The mouse olfactory bulb analysis utilized Stereo-seq technology, which captures subcellular resolution of the tissue, making it a valuable dataset for method benchmarking. The olfactory bulb consists of distinct cortical layers, including the meninges, olfactory nerve layer (ONL), glomerular layer (GL), external plexifrom layer (EPL), mitral cell layer (MCL), internal plexiform layer (IPL), granule cell layer (GCL), and the rostral migratory stream (RMS) [44, 45]. RASP, SEDR, and GraphST produced distinct clusters that effectively align with the layers in the olfactory bulb (Fig. 2A). In contrast, STAGATE, BASS, standard PCA, and spatialPCA were unable to effectively characterize the tissue structures in the core of the olfactory bulb.

In terms of runtime performance, RASP demonstrated computational costs comparable to standard PCA and between two and three orders-of-magnitude faster than the other spatially aware methods (Fig. 5B). Since ground truth annotations are not available for this dataset, cluster quality was assessed using Moran's I and CHAOS score (Fig. 5C,D, SI Algorithm 7,8). Notably, RASP exhibited the highest Moran's I and lowest CHAOS scores among the methods evaluated, with Louvain clustering producing the most favorable outcomes. Analysis of different kNN thresholds revealed that increased kNN values resulted in higher Moran's I values and lower CHAOS scores. However, beyond a kNN of 50, the resulting clusters begin to deviate from the anatomical layers, producing large homogeneous structures that, while rewarded by the CHAOS score, lack biological plausibility (Fig. 5D). RASP achieved the highest Moran's I value and lowest CHAOS score with low β values of between $0 - 0.5$. As the β parameter increased past this point, Moran's I decreased and the CHAOS score increased, indicating less optimal clustering solutions (Fig. 5, Extended Data Fig.5).

Reduced rank reconstruction performance was explored using the Doc2g gene, which acts upstream of neurotransmitter secretion and is located in the membrane and presynapses of neurons [46, 47]. Within the olfactory bulb, Doc2g is localized to the external plexiform layer (EPL) and internal plexiform layer (IPL). While the normalized expression of Doc2g in the original dataset appears sparse and noisy, the spatially smoothed expression generated by RASP is clearly localized to the EPL and IPL, showing greater spatial organization than the standard PCA reconstruction (Fig. 5E).

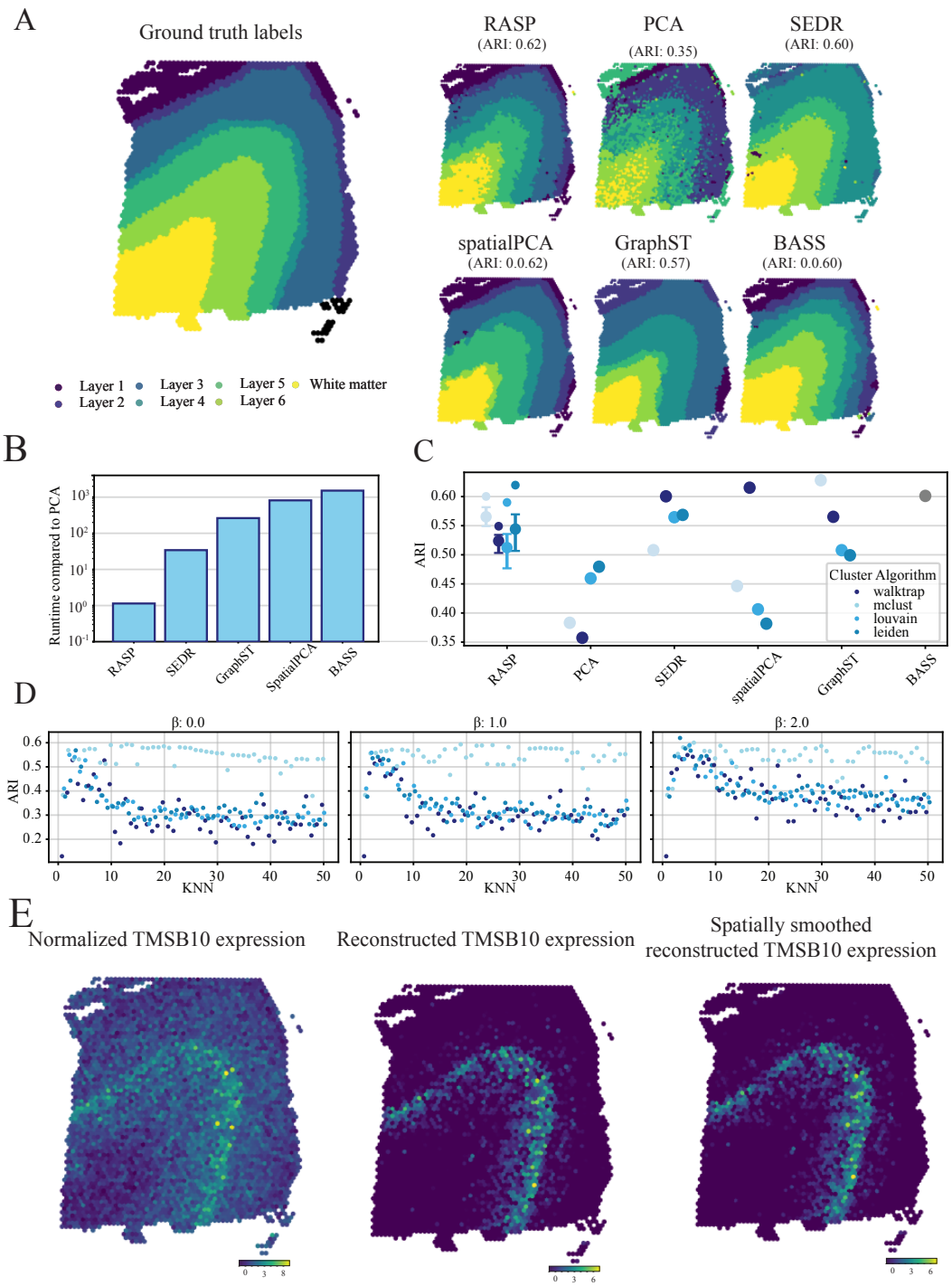

Figure 4: **Human DLPFC analysis. A:** Ground truth cortical layer annotations (left) and corresponding spatial domains identified by RASP, PCA, and competing methods. **B:** Quantification of runtime for all methods compared to normal PCA. **C:** Quantification of ARI for all methods, different colors indicate the clustering algorithm used to assign labels. Interquartile range and median ARI values at default RASP parameters (kNN =3-10, $\beta = 2$) is shown, along with maximum ARI values achieved by RASP. **D:** ARI values for RASP with inverse distance raised to the β of 0 (left), 1 (middle), and 2 (right), plotted against kNN. Colors indicate clustering algorithm. **E:** Normalize TMSB10 expression (left), reduced rank reconstructed TMSB10 expression (center), and spatially smoothed reduced rank reconstructed TMSB10 expression (right) plotted on tissue.

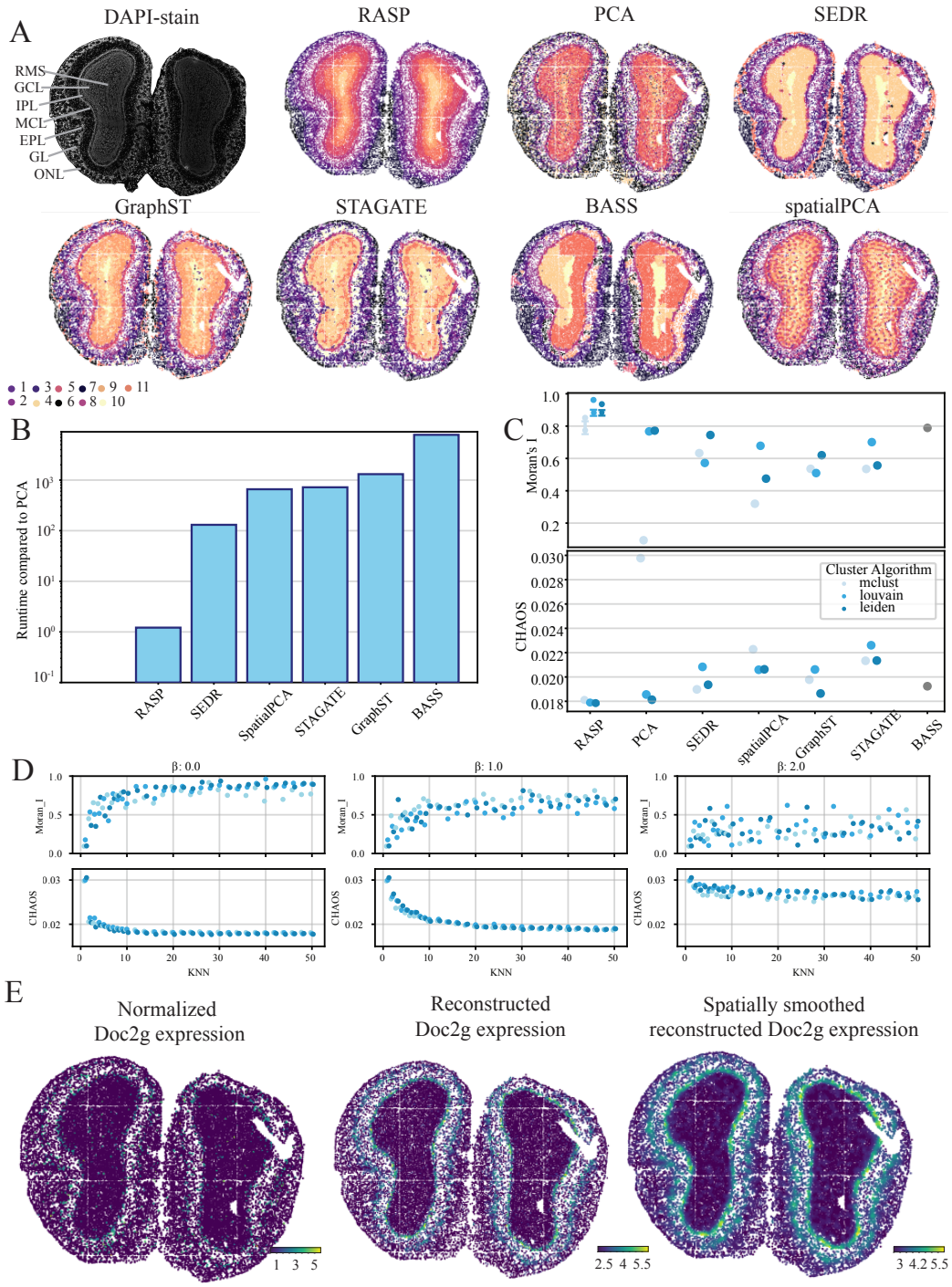

Figure 5: **Mouse olfactory bulb analysis A:** Cortical laminar structure (top left) and corresponding spatial domains identified by RASP, PCA, and other methods. **B:** Quantification of runtime for all methods compared to normal PCA. **C:** Quantification of Moran's I score (top), and CHAOS score (bottom) for all methods, different colors indicate the clustering algorithm used to assign labels. Interquartile range and median values shown for kNN between 50-100, and $\beta = 0$, along with maximum and minimum scores achived by RASP. **D:** Moran's I score (top) and CHAOS score (bottom) with inverse distance raised to the β of 0 (left), 1 (middle), and 2 (right), plotted against kNN. Colors indicate clustering algorithm. **E:** Normalize Doc2g expression (left), reduced rank reconstructed Doc2g expression (center), and spatially smoothed reduced rank reconstructed Doc2g expression (right).

2.7 RASP efficiently scales to large human breast cancer Xenium data while maintaining accurate tissue domain detection

To assess RAPS's effectiveness on large, high-resolution data, we downloaded the 10x Xenium human breast cancer dataset along with associated annotations from the R package SubcellularSpatialData [48]. This dataset originates from a publicly available 10x Genomics dataset of re-sectioned FFPE human breast tissue infiltrating ductal carcinoma *in situ*. The annotations are based on H&E stained tissue that accompany the Xenium data. Due to the dataset's size, which includes over half a million cells, execution of competing domain detection methods was not feasible due to out-of-memory errors. Successfully running these methods on this dataset requires computational resources exceeding 1 TB of RAM, which is unrealistic for many researchers, even with access to institutional high-performance computing environments. Given these limitations, we compared the performance of RASP against standard PCA. RASP substantially outscored (ARI of 0.63) standard PCA (ARI of 0.25) (Fig. 6A,B).

Analysis of the optimal kNN smoothing thresholds for RASP revealed that a larger kNN threshold range between $50 - 100$ is needed to correctly identify the tissue-spanning domains in this dataset (Fig.6D). For this dataset, Leiden and Louvain clustering were found to be optimal, followed by Mclust. It is important to consider the computational burden of clustering algorithms, especially given the lengthy runtimes associated with large datasets. Mclust can be three to four times as expensive computationally as Louvain or Leiden, however, if the user is interested in testing multiple resolutions to achieve a desired number of clusters, the three algorithms are comparable, with Mclust often running considerably faster. Similar to the olfactory bulb dataset, smaller β values between $0 - 0.5$ were found to produce better results, with ARI values dropping considerably for all clustering algorithms and all kNN thresholds at higher β values (Fig. 6D, Extended Data Fig. 6A).

To explore the reduced rank reconstruction feature of RASP, we compared normalized and reconstructed expression signatures for two genes: SFRP1 and LUM. SFRP1 encodes a protein that is part of the secreted frizzled-related protein family and modulates Wnt signaling, which, when dysregulated, occurs in several human tumor types [40, 49–51]. The spatially smoothed reduced rank reconstruction of SFRP1 shows expression in normal ducts and stromal cells, highlighting areas where the normalized expression was very sparse (Fig. 6C). LUM encodes a protein that is part of a small proteoglycan family and plays a role in collagen binding, with dysregulation in this pathway occuring in many cancer types [40, 52, 53]. Reduced rank reconstruction of LUM indicates high expression in both stromal and tumorous regions of the tissue.

2.8 PC number impacts RASP performance

For the real ST analyses detailed in Sections 2.3-2.7, we computed 20 spatially-smoothed PCs using RASP and performed location-based clustering on the components. Determining the optimal rank for dimensionality reduction is a long-standing problem in statistical analysis and ongoing topic of debate within the bioinformatics community for scRNA-seq and ST data. In practice, scRNA-seq and ST data analysis is usually performed without attempting to identify an optimal rank and instead uses a default of between 20 and 50 PCs. Notably, methods like **Scanpy** and **Seurat** default to retaining 30 and 50 PCs respectively, which serves as a useful benchmark for analysis.

To evaluate the robustness of our findings, we executed RASP using the best-performing parameters across the DLPFC, mouse ovary, and breast cancer datasets while retaining 100 PCs. We iteratively clustered each dataset using different numbers of retained PCs, ranging from 5 to 60, to better understand the correlation between PC count and cluster label accuracy (Extened Data Fig.8A). For the DLPFC dataset, we observed stable ARI values with an increasing number of PCs for the Louvain, Leiden, and Mclust clustering methods, and an increase for the Walktrap method. This invariance to PC number is likely due to the fact that the models underlying these clustering methods are able to describe the underlying distribution of the data well regardless of the number of dimensions. Conversely, in the mouse ovary dataset, we noted an upward trend in ARI as the number of PCs retained increased for both the Louvain and Leiden clustering methods. This behavior is consistent with the ground truth annotations of cell types present in this tissue, as higher-order PCs are adept at capturing the nuanced details required for accurate cell typing (Extended Data Fig.8A, middle). Interestingly, the Mclust method exhibited the opposite trend in this dataset, demonstrating improved predictive power at the lowest number of PCs. This finding is likely due to the fact that the Mclust algorithm assumes clusters have consistent size and shape, which is not the case in this dataset. In the breast cancer dataset, the Louvain and Leiden algorithms showed an increased ARI up to 30 PCs, and then ARI values decreased, while the Mclust algorithm showed decreased performance as the number of PCs increased. This reduction in classification accuracy aligns with the nature of the classification task

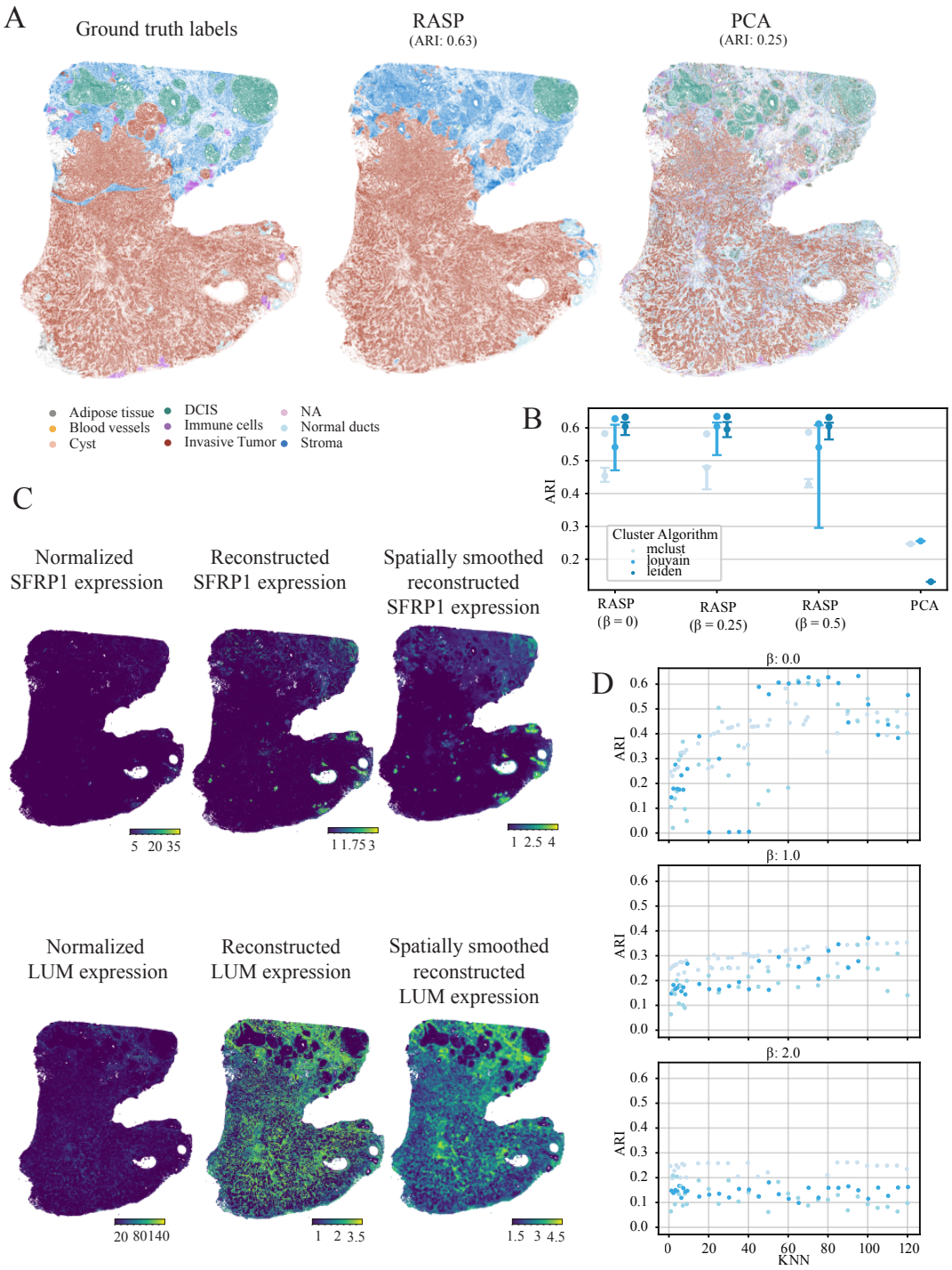

Figure 6: **Human breast cancer 10x Xenium analysis. A:** Ground truth spatial domain annotations (top left) and corresponding predictions for RASP, and normal PCA. **B:** Quantification of ARI for all methods, different colors indicate the clustering algorithm used to assign labels. Interquartile range and median ARI scores at default RASP parameters (kNN = 50-100, $\beta = 0.0.25, 0.5$), and maximum ARI values achieve by RASP. **C:** Normalize SFRP1 and LUM expression (left column), reduced rank reconstructed SFRP1 and LUM expression (center column), and spatially smoothed reduced rank reconstructed SFRP1 and LUM expression (right column) plotted on tissue. **D:** ARI values for RASP plotted against smoothing distance (top) and kNN (right). Colors indicate clustering algorithm.

in this dataset, which deals with more extensive spatial domains rather than finer cellular granularity, thereby indicating that broader broader spatial domains might be more appropriately captured with fewer PCs (Extended Data Fig.8A, right).

2.9 Influence of β on inverse distance weighting

To examine the effects of varying the parameter β on inverse distance weighting, we computed a sparse distance matrix representing the spatial relationships among data points. Subsequently, we applied an inverse weighting transformation to this distance matrix, raising it to different powers of β. The relationship between distance (um) and the corresponding weights, calculated at β values between 0 and 2, is visualized (Extended Data Fig. 9,10). A β value of 0 results in a relatively uniform weighting across distances, effectively creating a rectangular kernel. In contrast, $\beta = 2$ increases the importance of small distances, and downweighing the importance of larger distances. The inverse distance weightings for $\beta = 0, 1, 2$ were applied to the first six PCs computed on the Mouse Ovary dataset and visualized spatially (Extended Data Fig. 9B). Setting $\beta = 0$ results in large-scale smoothing across the structures present within the tissue. In contrast, $\beta = 2$ facilitates a localized smoothing effect, resulting in finer granularity of the observed structures. These results underscore the importance of β in modulating the influence of neighboring points on the distance-weighted computations.

2.10 Differential distance metrics and weighting analysis

For the RASP results presented in Sections 2.3-2.7, spatial smoothing was performed using a sparse matrix whose elements represent the inverse Euclidean distances raised to the power of β between ST locations and diagonal values set to the minimum distance for that column (or half the minimum distance in the case of Visium technology) . To assess alternative spatial smoothing approaches, we compared Euclidean and Manhattan distances using inverse distance weighting and kernel-based weightings (gaussian, quadratic) for the human breast cancer dataset (Extended Data Fig.11A). Following the application of these alternative smoothing methods, we performed clustering on the resulting smoothed PCs and calculated the ARI to evaluate the performance of each approach (Extended Data Fig.11 B,D). Our results demonstrate that the inverse Euclidean distance method consistently outperforms the other smoothing techniques in terms of clustering accuracy. This finding reinforces the effectiveness of our chosen approach in preserving the biological signals within the ST data. To provide further insight into the impact of the various smoothing strategies, we visualized the first four PCs generated by each method as kernel density plots (Extended data Fig.11C). This visualization highlights the subtle differences in how each method transforms the data, revealing distinct distributions that may influence the subsequent clustering results.

3 Discussion

Spatially informed dimensionality reduction of ST data is crucial for downstream tasks such as spatial domain identification, cell type annotation, and reduced rank gene signature reconstruction. In this study, we described RASP, a computationally efficient spatially informed dimensionality reduction method that seamlessly integrates into the standard analysis workflow provided by the Scanpy package. RASP utilizes randomized PCA followed by spatial smoothing to generate latent variables that can be leveraged for a wide range of downstream analyses including clustering to predict cell types or spatial domains. RASP also facilitates the reduced rank reconstruction of spatially smoothed gene signatures, enhancing the interpretability of transcriptomic data in spatial contexts. An important feature of RASP is its support for the incorporation of non-transcriptomic covariates via an additional round of randomized PCA. Our findings demonstrate that RASP achieves similar, or higher, accuracy than existing methods for cell type and spatial domain detection tasks, with runtime performance that is orders-of-magnitude faster. RASP can flexibly accommodate various tissue types and ST technologies. Specifically, RASP accurately identifies biologically relevant tissue domains in structures that are homogeneous in size and shape, such as in the DLPFC dataset, while also providing robust labels akin to cell types in the highly heterogeneous mouse ovary sample. Notably, RASP is currently the only method among the techniques we evaluated that can efficiently handle the massive datasets produced by the 10x Xenium platform. In summary, RASP is a novel ST dimensionality reduction method that provides researchers with an efficient option for detecting novel structures in their tissues or tissue collections without incurring substantial computational costs.

3.1 Parameter selection considerations

Several parameters require careful consideration when applying the RASP method to ST data. These include kNN threshold, β value, and the number of PCs to compute. To simplify parameter selection, we provide a recommended range of values for each parameter. Please see SI Table 1 for details.

Note that we chose to use inverse distance weighting instead of the more commonly employed Gaussian kernel weighting given the more direct biological interpretation (e.g., inverse squared distances correspond to the concentration of a secreted molecule) and increased flexibility (i.e., users can control both the power parameter and diagonal values instead of just a bandwidth parameter). While RASP defaults to inverse distance weighting, the method can be used with alternative distance metrics and smoothing approaches, each with different characteristics and potential benefits/disadvantages.

3.2 Non-transcriptomic covariates

The RASP method allows for the incorporation of additional location-specific covariates via a two-stage randomized PCA (these are specified using the covariate matrix $\mathbf{Y}$ in Algorithm 1). Potential covariates include local cell density and cell volume (for technologies that provide this information), the library size of each location/cell, and protein abundance estimates (for technologies that support joint transcriptomic/proteomic measurements). However, it is important to note that using cell density may be somewhat redundant, as the kNN threshold inherently accounts for local density variations. However, depending on the cluster algorithm, this covariate can still impart useful information (see Fig. 3B). Therefore, users should carefully select these parameters based on the specific objectives of their analysis. RASP's fast runtime allows users to quickly screen a large range of parameter values to identify the optimal configuration. RASP contains functionality to quickly calculate Moran's I and CHAOS scores on the resulting clusters and select the appropriate parameters for the final implementation.

3.3 Limitations

While RASP has numerous advantages relative to existing spatially aware dimensionality reduction techniques, it is not without limitations. One of the primary concerns involves the specification of user-provided parameters. Although the default parameters are generally appropriate for a variety of situations, RASP's smoothing threshold parameter can profoundly influence clustering results. Some degree of trial and error is often necessary to optimize this parameter. However, unlike competing tools that may require upwards of 30 minutes to run, RASP executes quickly, facilitating a straightforward optimization process for users. To assist with parameter selection, we provide guidelines that can help streamline this aspect of the workflow (SI Table1). Additionally, it is important to note that for certain datasets, RASP may not be the best performing method. For instance, in the task of cortical layer assignment in the DLPFC, models such as SEDR outperform RASP.

3.4 Future directions

Multi-slice alignment and integration of sections from the same organism over different time points are two areas that have witnessed numerous computational innovations over the last five years. Methods such as MOFA, MEFESTO and moscot utilize matrix factorization-based models to simultaneously model both temporal and spatial data [54–56]. PALMO employs variance decomposition analysis and incorporates some spatial modeling capabilities [57]. In contrast, methods such as PASTE2 and DeST-OT utilize optimal transport principles to align sections from the same tissue sample, enabling 3D models [58, 59]. Despite their promising capabilities, these models are computationally intensive, do not scale well to large datasets, and, importantly, lack streamlined approaches for joint dimensionality reduction and clustering. In the current study, we evaluated RASP's performance on single sections of ST data. Future directions include scaling RASP to support joint spatially informed dimensionality reduction of multiple ST slices, encompassing various tissues or multiple time points. While this enhancement can be simply realized using RASP by analyzing the concatenation of the cell-by-gene and coordinate matrices, how best to handle potential batch effects is a non-trivial problem. Given its computational efficiency, RASP stands out as the sole method capable of jointly reducing massive subcellular ST datasets. Another promising avenue of investigation involves integrating multi-modal information into the spatial reduction process. For instance, data related to protein abundance from methodologies like CITE-seq or chromatin accessibility assessed via spatial

ATAC-seq could significantly enhance spatial domain detection [60, 61]. RASP already includes functionality to support the incorporation of non-transciptomic covariates, and future research will focus on understanding the impact of these diverse data types in improving spatial domain detection and characterization.

4 Methods

4.1 RASP method

One of the primary motivations for spatially informed dimensionality reduction of ST data is the identification of spatial domains via unsupervised clustering. Clustering on standard PCs is often sufficient to identify cell types, but identification of larger tissue structures benefits from spatial information. The RASP method performs dimensionality reduction and spatial smoothing on normalized ST data and an optional set of location-specific covariates. The two-stage approach used by RASP was chosen to optimize computational efficiency and support the specification of covariates with an overall goal of enabling the flexible and efficient analysis of high-resolution ST datasets. Specifically, RASP uses a sparse matrix representation of the ST data, performs dimensionality reduction using randomized PCA, and generates spatially smoothed PCs using inverse distance weights sparsified according to a k-nearest neighbors (kNN) threshold (Fig. 1A). To support the integration of location-specific covariates into the latent variables, a second stage randomized PCA is optionally performed on the matrix created by appending the smoothed covariates to the smoothed PCs generated in the first stage. The spatially smoothed PCs generated by the RASP method have utility in a wide range of downstream analyses with clustering and reduced rank reconstruction of gene-level expression important use cases. A detailed description of the method is provided by Algorithm 1. For readability, this version of the algorithm automatically applies the same spatial smoothing to both the first stage PCs and the covariates, however, flexible smoothing of the covariates is supported so that users have control over whether to apply smoothing to each covariate and, if applied, the type of threshold that is employed.

4.2 Clustering on spatially smoothed principal components

For evaluation of the RASP method, clustering was performed on the spatially smoothed PCs ($\mathbf{P_s}$ or $\mathbf{P_c}$) generated by RASP 1. Specifically, clustering was performed by first computing a nearest neighbors distance matrix and a neighborhood graph of spatial locations as described in Mclnnes et al., 2018 [62]. See SI for programming language specific implementation. Note that the distances used for constructing this kNN graph are measured on the spatially smoothed PCs and capture transcriptomic differences between spatial locations, which is distinct from the spatial distances between coordinates used in RASP 1. Construction of the neighborhood graph requires specification of the number of nearest neighbors, with the recommended value being between 5 and 50. In general, the number of nearest neighbors should be set based on the size of the target spatial features. For example, users interested in cell type annotation are advised to set the threshold to a smaller value between 5 and 10, while those seeking to annotate spatial domains may consider increasing the threshold above 20. In most cases, a default of 10 is sufficient.

Distinct clusters representing spatial domains or cell types are then computed on the kNN graph using a user-defined clustering algorithm. The RASP method currently supports the use **Mclust**, **Louvain**, **Leiden**, and **Walktrap** methods. Note that for large datasets, users may want to avoid the use of the **Walktrap** clustering method as the computational complexity of the algorithm will increase runtime to an unreasonable level on a local machine.

4.3 Reduced rank reconstruction of gene-level expression

In addition to domain identification, an important application of spatially smoothed PCs is the reconstruction of gene-level expression values. Reduced rank reconstruction is motivated by the fact that the biological signal in transcriptomic data typically has a much lower rank than the rank of the observed measurements. Projecting ST data onto a lower dimensional space borrows the signal from correlated genes and can effectively mitigate both sparsity and noise [30]. For the spatially smoothed PCs generated by RASP 1, the reconstruction process follows Algorithm 2. Important features of this method relative to standard PCA-based reduced rank reconstruction include the use of spatially smoothed PCs to incorporate information from adjacent locations, the optional reconstruction

Algorithm 1 RASP method

Inputs:

- $\mathbf{X}$: $n \times m$ location-by-gene matrix that holds normalized expression values for n spatial locations and m genes.
- $\mathbf{C}$: $n \times 2$ matrix of spatial coordinates.
- p: Number of spatially smoothed PCs to compute.
- $threshold$: Threshold value. This is the number of nearest neighbors to retain.
- α: The diagonal elements of the distance matrix are set to this value to avoid division by zero when computing inverse distances. Default value is the minimum off-diagonal distance for the row.
- β: Power used for inverse distance weighting. Default is 2.
- $\mathbf{Y}$: Optional $n \times b$ matrix of b location-specific covariates, e.g., local cell density. While spatial smoothing is automatically applied to these variables in the algorithm, flexible control over covariate smoothing is possible in practice.

Outputs:

- $\mathbf{P_s}$: $n \times p$ matrix that holds the spatially smoothed projection of $\mathbf{X}$ onto the top p PCs.
- $\mathbf{Y_s}$: $n \times b$ matrix that holds the spatially smoothed version of $\mathbf{Y}$.
- $\mathbf{W}$: $m \times p$ matrix of loadings for the top p PCs of $\mathbf{X}$.
- $\mathbf{P_c}$: If $\mathbf{Y}$ is specified, the $n \times p$ matrix that holds the projection of $[\mathbf{P_s}, \mathbf{Y_s}]$ onto the top p PCs where $[\mathbf{P_s}, \mathbf{Y_s}]$ is the $(p+b) \times n$ matrix formed by appending the columns of $\mathbf{Y_s}$ to $\mathbf{P_s}$.
- $\mathbf{W_c}$: if $\mathbf{Y}$ is specified, the $(p+b) \times p$ matrix of loadings for the top p PCs computed on $[\mathbf{P_s}, \mathbf{Y_s}]$.

Notation:

- Let $\mathbf{W}[]$ represent a subsetting of the matrix $\mathbf{W}$ with $\mathbf{W}[i,j]$ the element in the ith row and j^{th} column and with $\mathbf{W}[.,.]$ the iteration over all i,j index combinations.

$\mathbf{W} \leftarrow randomizedPCA(\mathbf{X}, p)$ ▷ Perform randomized PCA on $\mathbf{X}$ for rank p. By default, $\mathbf{X}$ is centered but not scaled. See Algorithm S1 for details.
$\mathbf{P} \leftarrow \mathbf{XW}$ ▷ Project $\mathbf{X}$ onto the top p PCs.
$\mathbf{NN} \leftarrow fitNearestNeighbors(\mathbf{C}, threshold)$ ▷ Fit a nearest neighbors data structure on the spatial coordinates for $threshold$ neighbors.
$\mathbf{D} \leftarrow sparseMatrix(\mathbf{NN})$ ▷ Compute a sparse distance matrix using $\mathbf{NN}$. See Algorithm S4 for details.
$diag(\mathbf{D}) \leftarrow \alpha$ ▷ Set diagonal elements of $\mathbf{D}$ to α.
$\mathbf{D}_I[.,.] \leftarrow 1/\mathbf{D}[.,.]^{\beta}$ ▷ Create an inverse distance matrix using power β. Inverse distances for thresholded elements are set to 0.
$\mathbf{D}_I[.,.] \leftarrow \frac{\mathbf{D}_I[.,.]}{\max(\mathbf{D}_I[.,.])}$ ▷ Standardize each column of $\mathbf{D}_I$ to the maximum of that column.
$\mathbf{P_s} \leftarrow \mathbf{D}_I\mathbf{P}$ ▷ Spatially smooth the PCs using inverse distance weights.
if $\mathbf{Y}$ then ▷ If covariate matrix $\mathbf{Y}$ is specified.
    $\mathbf{Y_s} \leftarrow \mathbf{D}_I\mathbf{Y}$ ▷ Spatially smooth covariates.
    $\mathbf{W_c} \leftarrow randomizedPCA([\mathbf{P_s}, \mathbf{Y_s}], p)$ ▷ Perform randomized PCA on the merge of $\mathbf{P_s}$ and $\mathbf{Y_s}$.
    $\mathbf{P_c} \leftarrow [\mathbf{P_s}, \mathbf{Y_s}]\mathbf{W_c}$ ▷ Project $[\mathbf{P_s}, \mathbf{Y_s}]$ onto the top p PCs.
    return $\mathbf{P_s}, \mathbf{W}, \mathbf{Y_s}, \mathbf{P_c}, \mathbf{W_c}$
else
    return $\mathbf{P_s}, \mathbf{W}$
end

of the smoothed PCs to account for location-specific covariates, and the thresholding and optional scaling of the reconstructed values following the approach used by the ALRA method [30].

Algorithm 2 Reduced rank reconstruction

Inputs:

- $\mathbf{X}$: $n \times m$ location-by-gene matrix that holds normalized expression values for n spatial locations and m genes. This is the same matrix provided as input to Algorithm 1.
- $\mathbf{P_s}, \mathbf{W}, \mathbf{P_c}, \mathbf{W_c}$: Outputs from Algorithm 1. $\mathbf{P_c}$ and $\mathbf{W_c}$ are optional and should only be included if location-specific covariates were specified as inputs to Algorithm 1.
- q: Quantile probability used for thresholding reconstructed expression values.
- $scale$: True-to-scale reconstructed values to match variance of original expression.

Outputs:

- $\mathbf{X_r}$: Reduced rank reconstruction of $\mathbf{X}$.

Notation:

- Let $\mathbf{1}^{x \times y}$ be an $x \times y$ matrix of ones.
- Let $\mathbf{W}[]$ represent a subsetting of the matrix $\mathbf{W}$ with $\mathbf{W}[i,j]$ the element in the ith row and j^{th} column, $\mathbf{W}[i,]$ the ith row, $\mathbf{W}[,j]$ the jth column, and $\mathbf{W}[\mathbf{r},\mathbf{c}]$ the sub-matrix containing rows with indices in $\mathbf{r}$ and columns with indices in $\mathbf{c}$.

if $\mathbf{P_c}, \mathbf{W_c}$ then ▷ If $\mathbf{P_c}$ and $\mathbf{W_c}$ were specified
  $[\mathbf{P_{s,r}}, \mathbf{C_r}] \leftarrow \mathbf{P_c} \cdot \mathbf{W_c}^T$ ▷ Reconstruct smoothed PCs and covariates
  $\bar{\mathbf{p}}_s \leftarrow (\mathbf{P_s}^T \mathbf{1}^{n \times 1})/n$ ▷ Compute column means of $\mathbf{P_s}$
  $\mathbf{P_{s,r}} \leftarrow \mathbf{P_{s,r}} + \mathbf{1}^{n \times p} \bar{\mathbf{p}}_s^T$ ▷ Add back column means
  $\mathbf{X_r} \leftarrow \mathbf{P_{s,r}} \mathbf{W}^T$ ▷ Reconstruct ST expression matrix using reconstructed PCs
else
  $\mathbf{X_r} \leftarrow \mathbf{P_s} \mathbf{W}^T$ ▷ Reconstruct ST expression matrix using spatially smoothed PCs
end
$\bar{\mathbf{x}} \leftarrow (\mathbf{X}^T \mathbf{1}^{n \times 1})/n$ ▷ Compute column means of $\mathbf{X}$
$\mathbf{X_r} \leftarrow \mathbf{X_r} + \mathbf{1}^{n \times m} \bar{\mathbf{x}}^T$ ▷ Add back column means
for $i \in \{1, ..., m\}$ do ▷ Threshold and optionally scale each reconstructed gene
  $x_q \leftarrow quantile(\mathbf{X_r}[,i], q)$ ▷ Find quantile of reconstructed values for probability q
  for $j \in \{1, ..., n\}$ do ▷ If absolute reconstructed value is greater than absolute quantile, keep, otherwise, replace with original value if that was positive or set to zero.

$$\mathbf{X_r}[j,i] \leftarrow \begin{cases} \mathbf{X_r}[j,i] & \text{if } |\mathbf{X_r}[j,i]| \geq |x_q| \\ \mathbf{X}[j,i] & \text{if } \mathbf{X}[j,i] > 0 \text{ and } |\mathbf{X_r}[j,i]| < |x_q| \\ 0 & \text{if } \mathbf{X}[j,i] = 0 \text{ and } |\mathbf{X_r}[j,i]| < |x_q| \end{cases}$$

  end
  if scale then
    $\sigma_i \leftarrow stddev(\mathbf{X}[,i] | \mathbf{X}[,i] > 0)$ ▷ Compute standard deviation of non-zero $\mathbf{X}[,i]$
    $\sigma_{r,i} \leftarrow stddev(\mathbf{X_r}[,i] | \mathbf{X_r}[,i] > 0)$ ▷ Compute standard deviation of non-zero $\mathbf{X_r}[,i]$
    $s \leftarrow \sigma_i / max(\sigma_{r,i}, 1e-10)$ ▷ Compute scaling factor (avoiding division by zero)
    $\mathbf{X_r}[,i] \leftarrow s \mathbf{X_r}[,i]$ ▷ Apply scaling
  end
end
return $\mathbf{X_r}$

4.4 Application of RASP to existing ST datasets

4.4.1 Dataset descriptions

We applied RASP on four publicly available ST datasets that were generated using different techniques and resolutions on human and mouse tissues. Details of data preprocessing can be found in Supplementary Information. In brief, raw count matrices where filtered to remove cells/locations with mitochondrial and ribosomal RNA $> 10\%$, cells/locations with < 100 reads and > 5 times the mean absolute deviation (MAD). Resulting data were scaled, and log-normalized prior to running RASP (SI 1.5). The mouse ovary dataset (Vizgen MERFISH technology) consisted of 43,038 cells and 228 genes. The tissue section used in our analysis comes from a series of ST experiments

profiling mouse ovaries at 0, 4hr, and 12hr after induced ovulation with human choronic gonadotropin (hCG). The tissue investigated in this method was from the 12hr group and has subcellular resolution. Ground truth cell type annotations for eight cell types were taken from the original publication [15]. The mouse olfactory bulb dataset (STOmics Stereo-seq technology) consisted of 19,109 and 14,367 genes. This coronal section of the olfactory bulb is also profiled at sub-cellular resolution however the original publication did not provide ground truth regional annotations [44]. The human breast cancer tumor dataset (10x Genomics, Xenium technology) consisted of 565,916 cells and 541 genes, making it the largest dataset tested. This is a 5um thin resected infiltrating ductal carcinoma it situ with ground truth annotations provided in the **SubcellularSpatialData** R package [48]. Original annotations were determined from the accompanying histology image. The human DLPFC dataset (10x Genomics, Visium technology) consisted of 3,639 spots and 15,124 genes. The data here is a portion of the DLPFC that spans six neuronal layers plus white matter. The ground truth layer annotations come from the **spatialLIBD** R package [43].

4.4.2 RASP evaluation

To evaluate the RASP method, we applied the algorithm to the processed gene expression matrices of the four datasets described above. The locations where then clustered according to the spatially smoothed PCs contained in either the $\mathbf{P_s}$or $\mathbf{P_c}$ matrices output by RASP. Cluster quality was assessed using spatial continuity and compactness analysis quantified by the Moran's I value and CHAOS score (see SI 1.3-1.4) for all data sets and, if ground truth annotations were available, the adjusted Rand index (ARI) (see SI 1.2). The evaluation process required optimizing key parameters in the RASP pipeline, particularly the distance threshold used for constructing the sparse distance matrix and the β value used to exponentiate the inverse distance matrix.

Given the large parameter space, a systematic parameter sweep was performed. To efficiently explore this space, we adopted a one-parameter-at-a-time optimization strategy. All other variables were held constant while one parameter incremented upon. Specifically, the number of kNN for spatial smoothing was swept from 1 to 50 for the MERFISH, Visium and Stereo-seq datasets, and from 1 to 120 for the Xenium dataset. Each kNN sweep was performed a total of nine times, for $\beta = 0, 0.25, 0.5, , 0.75, 1, 1.25, 1.5, 1.75, 2$. This approach allowed us to comprehensively explore the influence of each parameter individually. For each combination of parameters, the ARI (SI 1.2) was computed to quantify the similarity between the resulting clusters and the ground truth labels, where available, enabling the identification of the optimal parameter set for each dataset. For the olfactory bulb dataset ground truth annotations were unavailable and spatial continuity and compactness analysis was quantified by the CHAOS score, and the spatial spatial autocorrelation of each cluster was calculated using Moran's I score in place of an accuracy measurement.

4.4.3 Comparison methods for spatial domain detection

We compared the performance of RASP against five other spatially informed dimensionality reduction or spatial domain detection algorithms: BASS (R, version 1.3.1), GraphST (Python, version 1.1.1), SEDR (Python, version 1.0.0), spatialPCA (R, version 1.3.0), and STAGATE pyG (Python, version 1.0.0)[24, 26–29]. These methods where selected as they demonstrated the best performance across a variety of tissues and employ diverse model architecture. BASS models ST data in a hierarchical Bayesian framework and treats the cell type or spatial domain label for each spot or cell as a latent variable and infers them through an efficient inference algorithm. GraphST utilizes a GNN encoder to learn latent representations of the variables and further refines the representations using self-supervised contrastive learning. A decoder is employed to reconstruct gene expression, where clustering is employed. SEDR learns a low-dimensional representation of the data embedded with spatial information via a masked self-supervised deep auto-encoder and a variational graph convolutional auto-encoder. This low dimensional embedding is then clustered to identify spatial domains. SpatialPCA models gene expression as a function of latent factors through a factor analysis model. It builds a kernel matrix using the coordinates from the experiment to model the spatial correlation structure of the latent factors across the tissue. The inferred low-dimensional components are then clustered to identify spatial domains. Finally, STAGATE learns a low-dimensional latent representation of spatial information and gene expression via a graph attention auto-encoder trained on a spatial neighbor network (SNN). The output is then clustered to identify spatial domains. For each method, we followed the associated GitHub tutorials to run the stated methods for spatial domain and cell type detection, unless otherwise specified. Details are provided in SI 2.3-2.6. All methods where run on a Linux server with 15 CPUs for

fair comparison of runtime. The latent dimensions found by each method were clustered using three clustering algorithms, Mclust, Louvain, and Leiden. The Walktrap clustering method was also used on the DLPFC dataset. Accuracy of the clusters generated by each method was determined by calculating ARI and CHAOS scores (SI 1.2-1.3). Note that these methods required allocations of arrays > 1TB for the human breast cancer dataset and as such where unable to run successfully.

Acknowledgments

Special thanks to Ruixu Huang for providing access to the mouse ovary dataset.

Funding

NIH grants R35GM146586 (IKG, HRF), P20GM130454 (IKG, BAG, HRF) and P30CA023108 (HRF).

Data availability

Mouse ovary: Please contact the authors for access the the Mouse ovary dataset.

DLPFC: We downloaded the raw DLPFC data (slide #151673) from the spatialLIBD website (`http://research.libd.org/spatialLIBD/index.html`).

Mouse Olfactory bulb: The dataset was downloaded from the SEDR publication GitHub repository [63] (`https://github.com/JinmiaoChenLab/SEDR_analyses/tree/master/data`).

Breast Cancer: We downloaded the Breast cancer Xenium dataset from the **SubcellularSpatialData R** package via **ExperimentHub**, dataset # EH8567, sample ID 'IDC' (`https://www.bioconductor.org/packages/release/data/experiment/html/SubcellularSpatialData.html`).

References

1. Ortiz, C. *et al.* Molecular atlas of the adult mouse brain. *Science Advances* **6.** Publisher: American Association for the Advancement of Science, eabb3446. `https://www.science.org/doi/10.1126/sciadv.abb3446` (2024) (June 26, 2020).
2. Wang, M. *et al.* High-resolution 3D spatiotemporal transcriptomic maps of developing Drosophila embryos and larvae. *Developmental Cell* **57,** 1271–1283.e4. ISSN: 15345807. `https://linkinghub.elsevier.com/retrieve/pii/S1534580722002465` (2024) (May 2022).
3. Walker, B. L., Cang, Z., Ren, H., Bourgain-Chang, E. & Nie, Q. Deciphering tissue structure and function using spatial transcriptomics. *Communications Biology* **5.** Publisher: Nature Publishing Group, 1–10. ISSN: 2399-3642. `https://www.nature.com/articles/s42003-022-03175-5` (2024) (Mar. 10, 2022).
4. Williams, C. G., Lee, H. J., Asatsuma, T., Vento-Tormo, R. & Haque, A. An introduction to spatial transcriptomics for biomedical research. *Genome Medicine* **14,** 68. ISSN: 1756-994X. `https://doi.org/10.1186/s13073-022-01075-1` (2024) (June 27, 2022).
5. Tian, L., Chen, F. & Macosko, E. Z. The expanding vistas of spatial transcriptomics. *Nature Biotechnology* **41.** Publisher: Nature Publishing Group, 773–782. ISSN: 1546-1696. `https://www.nature.com/articles/s41587-022-01448-2` (2024) (June 2023).
6. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353.** Publisher: American Association for the Advancement of Science, 78–82. `https://www.science.org/doi/full/10.1126/science.aaf2403` (2024) (July 2016).
7. Polański, K. *et al.* Bin2cell reconstructs cells from high resolution Visium HD data. *Bioinformatics* **40,** btae546. ISSN: 1367-4811. `https://doi.org/10.1093/bioinformatics/btae546` (2024) (Sept. 1, 2024).
8. Stickels, R. R. *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nature Biotechnology* **39.** Publisher: Nature Publishing Group, 313–319. ISSN: 1546-1696. `https://www.nature.com/articles/s41587-020-0739-1` (2024) (Mar. 2021).

9. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348.** Publisher: American Association for the Advancement of Science, aaa6090. https://www.science.org/doi/10.1126/science.aaa6090 (2024) (Apr. 24, 2015).

10. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nature Methods* **11.** Publisher: Nature Publishing Group, 360–361. ISSN: 1548-7105. https://www.nature.com/articles/nmeth.2892 (2024) (Apr. 2014).

11. Codeluppi, S. *et al.* Spatial organization of the somatosensory cortex revealed by osmFISH. *Nature Methods* **15.** Publisher: Nature Publishing Group, 932–935. ISSN: 1548-7105. https://www.nature.com/articles/s41592-018-0175-z (2024) (Nov. 2018).

12. He, S. *et al. High-plex Multiomic Analysis in FFPE at Subcellular Level by Spatial Molecular Imaging* Pages: 2021.11.03.467020 Section: New Results. July 19, 2022. https://www.biorxiv.org/content/10.1101/2021.11.03.467020v3 (2024).

13. Janesick, A. *et al.* High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature Communications* **14.** Publisher: Nature Publishing Group, 8353. ISSN: 2041-1723. https://www.nature.com/articles/s41467-023-43458-x (2024) (Dec. 19, 2023).

14. Henley, R. *et al.* In situ detection and subcellular localization of 5,000 genes using Xenium Analyzer.

15. Huang, R. *et al. Single-cell and spatiotemporal profile of ovulation in the mouse ovary* Pages: 2024.05.20.594719 Section: New Results. June 5, 2024. https://www.biorxiv.org/content/10.1101/2024.05.20.594719v4 (2024).

16. Shi, Y. *et al.* A spatiotemporal gene expression and cell atlases of the developing rat ovary. *Cell Proliferation* **56.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cpr.13516, e13516. ISSN: 1365-2184. https://onlinelibrary.wiley.com/doi/abs/10.1111/cpr.13516 (2024) (2023).

17. Mantri, M. *et al.* Spatiotemporal transcriptomics reveals pathogenesis of viral myocarditis. *Nature Cardiovascular Research* **1.** Publisher: Nature Publishing Group, 946–960. ISSN: 2731-0590. https://www.nature.com/articles/s44161-022-00138-1 (2024) (Oct. 2022).

18. Garcia-Alonso, L. *et al.* Mapping the temporal and spatial dynamics of the human endometrium in vivo and in vitro. *Nature Genetics* **53.** Publisher: Nature Publishing Group, 1698–1711. ISSN: 1546-1718. https://www.nature.com/articles/s41588-021-00972-2 (2024) (Dec. 2021).

19. Mantri, M., Zhang, H. H., Spanos, E., Ren, Y. A. & De Vlaminck, I. A spatiotemporal molecular atlas of the ovulating mouse ovary. *Proceedings of the National Academy of Sciences* **121.** Publisher: Proceedings of the National Academy of Sciences, e2317418121. https://www.pnas.org/doi/full/10.1073/pnas.2317418121 (2024) (Jan. 30, 2024).

20. Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362.** Publisher: American Association for the Advancement of Science, eaau5324. https://www.science.org/doi/10.1126/science.aau5324 (2024) (Nov. 16, 2018).

21. Cang, Z. *et al.* Screening cell–cell communication in spatial transcriptomics via collective optimal transport. *Nature Methods* **20.** Publisher: Nature Publishing Group, 218–228. ISSN: 1548-7105. https://www.nature.com/articles/s41592-022-01728-4 (2024) (Feb. 2023).

22. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177.** Publisher: Elsevier, 1888–1902.e21. ISSN: 0092-8674, 1097-4172. https://www.cell.com/cell/abstract/S0092-8674(19)30559-8 (2024) (June 13, 2019).

23. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19,** 15. ISSN: 1474-760X. https://doi.org/10.1186/s13059-017-1382-0 (2024) (Feb. 6, 2018).

24. Li, Z. & Zhou, X. BASS: multi-scale and multi-sample analysis enables accurate cell type clustering and spatial domain detection in spatial transcriptomic studies. *Genome Biology* **23,** 168. ISSN: 1474-760X. https://doi.org/10.1186/s13059-022-02734-7 (2024) (Aug. 4, 2022).

25. Teng, H., Yuan, Y. & Bar-Joseph, Z. Clustering spatial transcriptomics data. *Bioinformatics* **38,** 997–1004. ISSN: 1367-4803. https://doi.org/10.1093/bioinformatics/btab704 (2024) (Jan. 27, 2022).

26. Long, Y. *et al.* Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nature Communications* **14.** Publisher: Nature Publishing Group, 1155. ISSN: 2041-1723. https://www.nature.com/articles/s41467-023-36796-3 (2024) (Mar. 1, 2023).

27. Xu, H. *et al.* Unsupervised spatially embedded deep representation of spatial transcriptomics. *Genome Medicine* **16,** 12. ISSN: 1756-994X. https://doi.org/10.1186/s13073-024-01283-x (2024) (Jan. 12, 2024).

28. Shang, L. & Zhou, X. Spatially aware dimension reduction for spatial transcriptomics. *Nature Communications* **13.** Publisher: Nature Publishing Group, 7203. ISSN: 2041-1723. https://www.nature.com/articles/s41467-022-34879-1 (2024) (Nov. 23, 2022).

29. Dong, K. & Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature Communications* **13.** Publisher: Nature Publishing Group, 1739. ISSN: 2041-1723. https://www.nature.com/articles/s41467-022-29439-6 (2024) (Apr. 1, 2022).

30. Linderman, G. C. *et al.* Zero-preserving imputation of single-cell RNA-seq data. *Nature Communications* **13.** Publisher: Nature Publishing Group, 192. ISSN: 2041-1723. https://www.nature.com/articles/s41467-021-27729-z (2024) (Jan. 11, 2022).

31. Zeng, Z., Li, Y., Li, Y. & Luo, Y. Statistical and machine learning methods for spatially resolved transcriptomics data analysis. *Genome Biology* **23,** 83. ISSN: 1474-760X. https://doi.org/10.1186/s13059-022-02653-7 (2024) (Mar. 25, 2022).

32. Wang, L., Maletic-Savatic, M. & Liu, Z. Region-specific denoising identifies spatial co-expression patterns and intra-tissue heterogeneity in spatially resolved transcriptomics data. *Nature Communications* **13.** Publisher: Nature Publishing Group, 6912. ISSN: 2041-1723. https://www.nature.com/articles/s41467-022-34567-0 (2024) (Nov. 14, 2022).

33. Zhao, E. *et al.* Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology* **39.** Number: 11 Publisher: Nature Publishing Group, 1375–1384. ISSN: 1546-1696. https://www.nature.com/articles/s41587-021-00935-2 (2023) (Nov. 2021).

34. Halko, N., Martinsson, P.-G. & Tropp, J. A. *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions* Dec. 14, 2010. arXiv: 0909.4061[math]. http://arxiv.org/abs/0909.4061 (2024).

35. Scrucca, L., Fraley, C., Murphy, T. B. & Raftery, A. *Model-Based Clustering, Classification, and Density Estimation Using mclust in R* https://mclust-org.github.io/book/ (Chapamn and Hall/CRC, 2023).

36. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008,** P10008. ISSN: 1742-5468. arXiv: 0803.0476[physics]. http://arxiv.org/abs/0803.0476 (2024) (Oct. 9, 2008).

37. Traag, V. A., Waltman, L. & Eck, N. J. v. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9,** 5233. https://pmc.ncbi.nlm.nih.gov/articles/PMC6435756/ (2024) (Mar. 26, 2019).

38. Pons, P. & Latapy, M. *Computing communities in large networks using random walks (long version)* Dec. 12, 2005. arXiv: physics/0512106. http://arxiv.org/abs/physics/0512106 (2024).

39. Faire, M. *et al.* Follicle dynamics and global organization in the intact mouse ovary. *Developmental biology* **403,** 69–79. ISSN: 0012-1606. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4469539/ (2024) (July 1, 2015).

40. Fagerberg, L. *et al.* Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics. *Molecular & Cellular Proteomics : MCP* **13,** 397. https://pmc.ncbi.nlm.nih.gov/articles/PMC3916642/ (2024) (Dec. 5, 2013).

41. Madhra, M., Gay, E., Fraser, H. M. & Duncan, W. Alternative splicing of the human luteal LH receptor during luteolysis and maternal recognition of pregnancy. *Molecular Human Reproduction* **10,** 599–603. ISSN: 1360-9947. https://doi.org/10.1093/molehr/gah076 (2024) (Aug. 1, 2004).

42. Jeppesen, J. V. *et al.* LH-Receptor Gene Expression in Human Granulosa and Cumulus Cells from Antral and Preovulatory Follicles. *The Journal of Clinical Endocrinology and Metabolism* **97,** E1524. https://pmc.ncbi.nlm.nih.gov/articles/PMC3410279/ (2024) (June 1, 2012).

43. Maynard, K. R. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience* **24.** Publisher: Nature Publishing Group, 425–436. ISSN: 1546-1726. https://www.nature.com/articles/s41593-020-00787-0 (2024) (Mar. 2021).

44. Chen, A. *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **185,** 1777–1792.e21. ISSN: 00928674. https://linkinghub.elsevier.com/retrieve/pii/S0092867422003993 (2024) (May 2022).

45. Nagayama, S., Homma, R. & Imamura, F. Neuronal organization of olfactory bulb circuits. *Frontiers in Neural Circuits* **8,** 98. https://pmc.ncbi.nlm.nih.gov/articles/PMC4153298/ (2024) (Sept. 3, 2014).

46. Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515,** 355. https://pmc.ncbi.nlm.nih.gov/articles/PMC4266106/ (2024) (Nov. 19, 2014).

47. Pang, Z. P. *et al.* Doc2 Supports Spontaneous Synaptic Transmission by a Ca2+-Independent Mechanism. *Neuron* **70,** 244–251. ISSN: 0896-6273. https://www.sciencedirect.com/science/article/pii/S0896627311002078 (2024) (Apr. 28, 2011).

48. Bhuva, D. D. *et al.* Library size confounds biology in spatial transcriptomics data. *Genome Biology* **25,** 99. ISSN: 1474-760X. https://doi.org/10.1186/s13059-024-03241-7 (2024) (Apr. 18, 2024).

49. Bernemann, C. *et al.* Influence of secreted frizzled receptor protein 1 (SFRP1) on neoadjuvant chemotherapy in triple negative breast cancer does not rely on WNT signaling. *Molecular Cancer* **13,** 174. ISSN: 1476-4598. https://doi.org/10.1186/1476-4598-13-174 (2024) (July 17, 2014).

50. Ren, J., Wang, R., Song, H., Huang, G. & Chen, L. Secreted Frizzled Related Protein 1 Modulates Taxane Resistance of Human Lung Adenocarcinoma. *Molecular Medicine* **20.** Number: 1 Publisher: BioMed Central, 164–178. ISSN: 1528-3658. https://molmed.biomedcentral.com/articles/10.2119/molmed.2013.00149 (2024) (Jan. 2014).

51. Davaadorj, M. *et al.* Loss of SFRP1 Expression Is Associated with Poor Prognosis in Hepatocellular Carcinoma. *Anticancer Research* **36.** Publisher: International Institute of Anticancer Research Section: Experimental Studies, 659–664. ISSN: 0250-7005, 1791-7530. https://ar.iiarjournals.org/content/36/2/659 (2024) (Feb. 1, 2016).

52. Matsuda, Y. *et al.* Expression and roles of lumican in lung adenocarcinoma and squamous cell carcinoma. *International Journal of Oncology* **33.** Publisher: Spandidos Publications, 1177–1185. ISSN: 1019-6439. https://www.spandidos-publications.com/10.3892/ijo_00000107 (2024) (Dec. 1, 2008).

53. Yang, C.-T., Hsu, P.-C. & Chow, S.-E. Downregulation of lumican enhanced mitotic defects and aneuploidy in lung cancer cells. *Cell Cycle* **19.** Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/15384101.2019.1693189, 97–108. ISSN: 1538-4101. https://doi.org/10.1080/15384101.2019.1693189 (2024) (Jan. 2, 2020).

54. Argelaguet, R. *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology* **21,** 111. ISSN: 1474-760X. https://doi.org/10.1186/s13059-020-02015-1 (2024) (May 11, 2020).

55. Velten, B. *et al.* Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. *Nature Methods* **19.** Publisher: Nature Publishing Group, 179–186. ISSN: 1548-7105. https://www.nature.com/articles/s41592-021-01343-9 (2024) (Feb. 2022).

56. Klein, D. *et al. Mapping cells through time and space with moscot* Pages: 2023.05.11.540374 Section: New Results. May 12, 2023. https://www.biorxiv.org/content/10.1101/2023.05.11.540374v2 (2024).

57. Vasaikar, S. V. *et al.* A comprehensive platform for analyzing longitudinal multi-omics data. *Nature Communications* **14.** Publisher: Nature Publishing Group, 1684. ISSN: 2041-1723. https://www.nature.com/articles/s41467-023-37432-w (2024) (Mar. 27, 2023).

58. Liu, X., Zeira, R. & Raphael, B. J. *PASTE2: Partial Alignment of Multi-slice Spatially Resolved Transcriptomics Data* Pages: 2023.01.08.523162 Section: New Results. Jan. 8, 2023. https://www.biorxiv.org/content/10.1101/2023.01.08.523162v1 (2024).

59. Halmos, P. *et al.* DeST-OT: Alignment of Spatiotemporal Transcriptomics Data. *bioRxiv,* 2024.03.05.583575. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10942402/ (2024) (Mar. 10, 2024).

60. Liu, Y. *et al.* High-plex protein and whole transcriptome co-mapping at cellular resolution with spatial CITE-seq. *Nature Biotechnology* **41.** Publisher: Nature Publishing Group, 1405–1409. ISSN: 1546-1696. https://www.nature.com/articles/s41587-023-01676-0 (2024) (Oct. 2023).

61. Deng, Y. *et al.* Spatial profiling of chromatin accessibility in mouse and human tissues. *Nature* **609.** Publisher: Nature Publishing Group, 375–383. ISSN: 1476-4687. https://www.nature.com/articles/s41586-022-05094-1 (2024) (Sept. 2022).

62. McInnes, L., Healy, J. & Melville, J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* Sept. 18, 2020. arXiv: 1802.03426[stat]. http://arxiv.org/abs/1802.03426 (2024).

63. Chen, S. *et al.* Integration of spatial and single-cell data across modalities with weakly linked features. *Nature Biotechnology.* Publisher: Nature Publishing Group, 1–11. ISSN: 1546-1696. https://www.nature.com/articles/s41587-023-01935-0 (2024) (Sept. 7, 2023).