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 spatial organization of gene expression, enabling researchers to unravel the intricate relationship between cellular environments and biological function. Identifying spatial domains within tissues is key to understanding tissue architecture and mechanisms underlying development and disease progression. Here, we present Randomized Spatial PCA (RASP), a novel spatially-aware dimensionality reduction method for ST data. RASP is designed to be orders-of-magnitude faster than existing techniques, scale to datasets with $100,000+$ locations, support flexible integration of non-transcriptomic covariates, and reconstruct de-noised, spatially-smoothed gene expression values. It employs a randomized two-stage PCA framework and configurable spatial smoothing. RASP was compared to BASS, GraphST, SEDR, SpatialPCA, STAGATE, and CellCharter using diverse ST datasets (10x Visium, Stereo-Seq, MERFISH, 10x Xenium) on human and mouse tissues. In these benchmarks, RASP delivers comparable or superior accuracy in tissue-domain detection while achieving substantial improvements in computational speed. Its efficiency not only reduces runtime and resource requirements but also makes it practical to explore a broad range of spatial-smoothing parameters in a high-throughput fashion. By enabling rapid re-analysis under different parameter settings, RASP empowers users to fine-tune the balance between resolution and noise suppression on large, high-resolution subcellular datasets—a critical capability when investigating complex tissue architecture.

1 Introduction

Spatial transcriptomics (ST) assays measure gene expression at thousands to hundreds-of-thousands of spatial locations within tissue sections, enabling researchers to characterize transcript abundance *in vivo* at resolutions ranging from small cell groups to subcellular structures [1, 2]. ST technology facilitates exploration of gene expression's spatial organization, providing insights into cellular microenvironments and functional states of individual cells [3]. Two primary types of ST technologies exist: 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]. 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, allowing for use of archival tissue banks, increasing the use of iST technologies [9, 12, 13]. Additionally, probe sets used in iST technologies have expanded, with systems like 10x Xenium developing probe sets with up to 5,000 individual probes [14], helping to bridge the trade-off between sequencing depth and resolution, allowing for more comprehensive profiling of gene expression. As costs decrease and technology advances, ST studies will transition from analyzing small sample sizes to extensive investigations [2, 13, 15–19], generating multi-slide, multi-timepoint ST datasets covering millions of cells across tissues [2]. Larger datasets necessitate development of novel computational tools that leverage both spatial and transcriptomic data efficiently.

ST's rich data allows tissue investigation at multiple biological scales, and computational approaches must adapt. On a cellular level, cell type assignment facilitates spatial distribution visualization, enhancing understanding of cellular structure and function in native tissue [20, 21]. At larger scales, identifying tissue domains aids comprehension of complex tissue microenvironments. Identifying cell types and spatial domains is achievable with recently developed computational tools. Popular single-cell RNA sequencing (scRNA-seq) tools like Seurat and Scanpy perform principal component analysis (PCA) and clustering on gene expression data while disregarding spatial context [22, 23]. However, methods like BASS and FICT now incorporate spatial context [24, 25]. For spatial domain detection, methodologies like BASS, SEDR, GraphST, STAGATE, SpatialPCA, CellCharter, Banksy, NichePCA, and MENDER exist [24, 26–33], utilizing different model architectures to embed spatial and gene expression data into low-dimensionality spaces for clustering and domain identification. ST data is often sparse, so de-noising and imputing poorly captured transcripts is a desirable feature, and many spatially-aware methods provide this functionality [34–37].

Yet, current methods have notable limitations. While effective on a single ST slice with few profiled locations, they are computationally intensive, exhibiting slow processing locally, particularly on iST datasets exceeding $100,000$ cells. With iST dataset generation increasing, efficient, spatially-aware dimensionality reduction algorithms are necessary. Many existing methods are also inflexible and complex, hindering customization of spatial smoothing or integration of non-transcriptomic covariates, such as those derived from hematoxylin and eosin (H&E) images or protein abundance metrics. Addressing these limitations is crucial for deriving meaningful insights from large-scale ST datasets.

In this paper, we present Randomized Spatial PCA (RASP), a computational method designed to be much faster than existing techniques, scalable to ST data with hundreds-of-thousands of locations, and supportive of non-transcriptomic covariates, while remaining easy to understand and implement. RASP employs randomized linear algebra techniques for efficient dimensionality reduction of large datasets [38]. It performs spatial smoothing using the coordinates returned from the ST platform and supports optional integration of covariates like cellular density or sequencing depth, providing greater analytical flexibility. Tested on four real-world and two simulated datasets (including 10x Genomics Visium and subcellular imaging technologies like Stereo-Seq, MERFISH, and 10x Xenium), RASP demonstrates tissue domain detection performance comparable to or superior to existing methods, with orders-of-magnitude faster runtimes.

2 Results

2.1 RASP overview

RASP performs dimensionality reduction and spatial smoothing on normalized ST data, with the option to incorporate location-specific covariates. The randomized two-stage PCA optimizes computational efficiency while enabling the flexible and efficient analysis of high-resolution ST datasets. RASP utilizes a sparse matrix representation of the data, conducts dimensionality reduction via randomized PCA, and generates spatially smoothed principal components (PCs) using an inverse distance matrix sparsified by a k-nearest neighbors (kNN) threshold (Fig. 1A). Covariates integration into the latent variables is achieved through an optional second-stage randomized PCA on a matrix with covariates appended to first-stage smoothed PCs. RASP's spatially smoothed PCs 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 parameter. See Extended data Fig 15,16. for effects of variable PC number on RASP performance.

2.2 RASP evaluation on real data

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

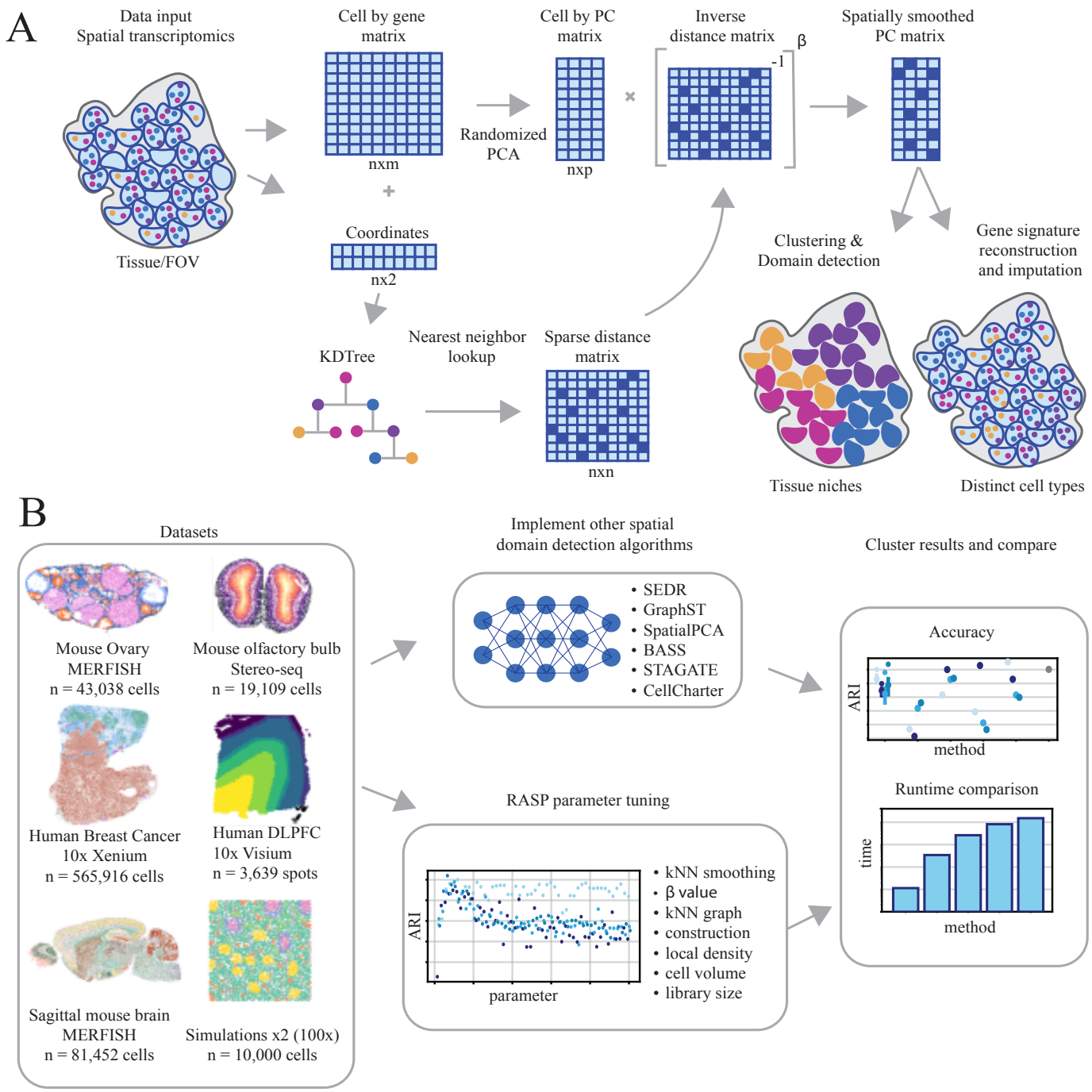

Figure 1: **RASP overview. A: Method workflow.** RASP takes as input a cell-by-gene expression matrix and spatial coordinates from a ST experiment. The expression matrix is reduced via randomized PCA and spatially smoothed by a sparse inverse distance matrix. Non-transcriptomic covariates can be added 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) and two synthetic datasets (100 replicates each). RASP results were compared against standard PCA and alternative spatially aware dimensionality reduction models (SEDR, GraphST, SpatialPCA, BASS, STAGATE, and CellCharter) for both accuracy and speed.

We evaluated RASP by clustering locations using the spatially smoothed principal components (PCs) output by the algorithm. Cluster quality was assessed through spatial continuity and compactness via CHAOS score (SI 1.3), spatial autocorrelation by Moran's I (SI 1.4) and the Adjusted Rand Index (ARI) (SI 1.2) compared to ground truth annotations. We conducted a sensitivity analysis on key RASP parameters, focusing on the kNN distance threshold for sparse matrix construction and the β power parameter (see Figure 1A). This analysis used a one-parameter-at-a-time strategy to investigate parameter impacts on clustering outcomes.

Performance comparisons were made between RASP and six other spatially informed dimensionality reduction or spatial domain detection methods: BASS [24], GraphST [26], SEDR [27], SpatialPCA [28], STAGATE [29], and CellCharter [30] 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 latent dimensions found by each method was performed using Mclust [39], Louvain [40], Leiden [41], and Walktrap [42] 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.

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

We selected this MERFISH dataset for its complex tissue structure and cellular resolution, featuring spatially distinct and heterogeneous cell types. Obtained from Huang et al., [15] it includes ground truth cell type annotations based on histology and differential gene expression (Fig. 2A). Annotations include luteal, luteinizing mural, granulosa, theca, and stromal cells, with clear spatial organization, alongside endothelial, epithelial, and immune cells, which are heterogeneously distributed across the tissue [43], making automatic annotation challenging.

For this dataset, RASP using Louvain clustering achieved the highest ARI of 0.69, 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 performed best, while Mclust the worst. Note that BASS utilizes Mclust as a prior in its Bayesian framework but does not support a user-defined clustering algorithm.

Relative runtimes show RASP is one to three orders-of-magnitude faster than competing methods on this dataset (Fig. 2B). SEDR is an order-of-magnitude slower than RASP, while BASS is almost four orders-of-magnitude slower. RASP performance diminishes with increasing kNN thresholds for PC smoothing; 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. We found $\beta = 2$ the best performing inverse distance power value. At higher β kNN threshold's impact on ARI was blunted, with comparable performance across the entire range of kNN thresholds sampled. Conversely, at $\beta = 0$, the kNN threshold dramatically affected RASP performance (Fig. 2D, Extended Data Fig.1).

Using RASP for reduced rank reconstruction of 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, is expressed on granulosa cells, theca cells, and luteal cells [44–46]. Originally, gene expression showed minimal spatial patterning (Fig. 2E, left plot). After RASP reconstruction, expression is localized to luteinizing mural and luteal cells (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

A key feature of RASP is support for integrating non-transcriptomic covariates into reduced dimensions. The two-stage PCA approach used by RASP offers benefits relative to a single PCA on merged transcriptomic and non-transcriptomic variables: It prevents covariate signal dilution within the high-dimensional transcriptomic data, and decouples spatial smoothing of expression data from covariates. To RASP covariate integration, we explored three 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 RASP configurations: 1) standard RASP without covariates (i.e., the configuration used to generate the results in Figure 2) 2), single-stage RASP executed on the merge of ST expression data and covariates, and 3) two-stage RASP with the covariates

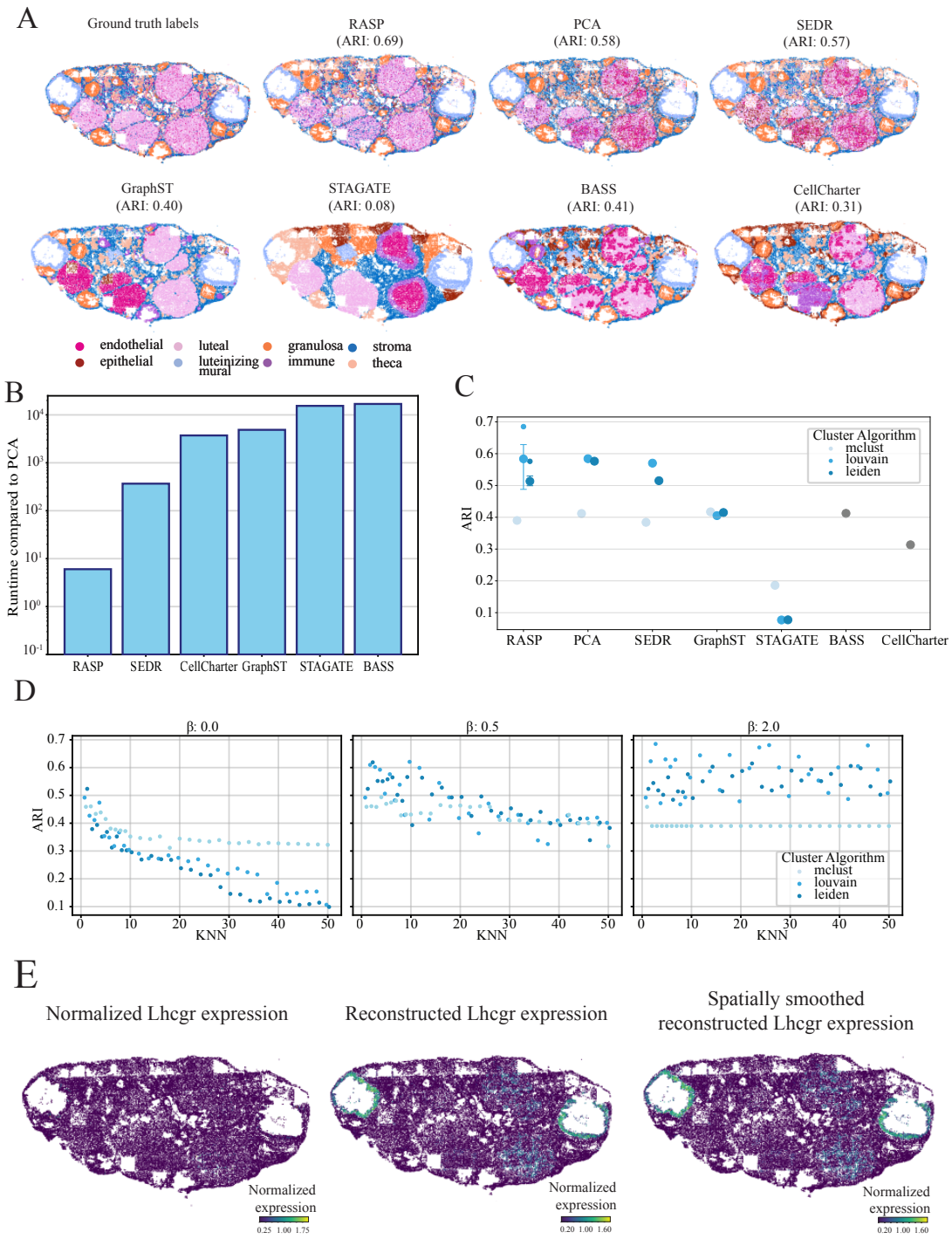

Figure 2: **Mouse ovary analysis (Vizgen MERFISH). A:** Ground truth cell type annotations (top left) and corresponding predictions for RASP, normal PCA, and other methods **B:** Quantification of runtime for all methods compared to normal PCA. **C:** Quantification of ARI for all methods, different 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).

added to the output from the first stage (see Fig. 3A (bottom)). For these experiments, the β parameter was held constant at $\beta = 2$,the optimal value for this dataset (see Fig. 2D, and Extended Data Fig 1). For local density, spatial smoothing was applied to the PC matrix, but not the local density variable, as local density inherently contains spatial information. Library size and cell volume covariates where smoothed by the same inverse distance weighting as the PCs matrix, prior the second round of dimensionality reduction. In addition, we explored the addition of local cell density for the breast cancer dataset, using the same experimental design described above, with the RASP kNN held constant at 60. The results were less striking than the ovary dataset, with both the one and two stage versions improving the clustering across some of the density values for Louvain clustering, and generally reducing performance when using Leiden or mclust algorithms (Extended data Fig 6).

All three covariates impacted ARI values, with clustering algorithm determining the magnitude and direction of the ARI change. The ARI using Mclust improved regardless of covariate added, with single-stage RASP equal to or outperforming the two-stage version in most cases (see Fig. 3B (left column)). Mclust ARI without covariates was 0.39 improving to 0.45 and 0.5 with covariates. ARI with Leiden was highest for two-stage RASP in most cases, while the single-stage version was covariate dependent, with the addition of local density decreasing ARI values, and addition of library size and cell volume having mixed effects (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-classification was hindered by local density additions, with ARI values dropping across kNN values, while affects of library size and cell volume fluctuated with kNN value. Overall, these results indicate that integration of covariates can enhance RASP performance, but covariate selection, clustering algorithm, spatial smoothing require careful consideration. Importantly, two-stage RASP avoids covariate dilution offering more control over spatial smoothing.

2.5 RASP enables the investigation of tissue domains at multiple scales

To investigate RASP's capacity to resolve tissue structure across a hierarchy of physical scales, we applied it to the Allen Institute's adult mouse brain atlas [47],[1]. This dataset provides comprehensive cell type annotations as well as multiple hierarchical levels of cortical and subcortical brain region annotations, allowing for the assessment of both fine-grained (cell type) and coarse-grained (anatomical region) clustering performance. We compared RASP followed by clustering to the ground truth cell type labels and broad brain region divisions (Figs.4A and 4D). For the cell typing task, RASP achieved an ARI of 0.77, substantially surpassing other state-of-the-art clustering methods, using parameter values optimized for cellular identity ($\beta = 2.0$, kNN=2–20; Figs. 4B, and 4C). For the higher-level regional annotation task, RASP achieved an ARI of 0.55, outperforming other approaches with parameters designed for large-scale spatial domains ($\beta = 0$, kNN=60–100; Figs. 4E and 4H). See 2.10 for further discussion on RASP's multi-scale ability.

2.6 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 [48] and includes manual cortical layer annotations. This dataset serves as a gold standard benchmark for spatial domain detection algorithms due to its clear neuroanatomical boundaries providing reliable ground truth and a small number of profiled tissue locations facilitating 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 [48].

For this dataset, RASP and SpatialPCA emerged as the best-performers, achieving ARIs of 0.62 and 0.61 respectively. SEDR and BASS followed closely, with ARIs of 0.60, while GraphST and CellCharter scored lowest (0.57, 0.55) (Fig. 5A and 5D). Importantly, RASP has comparable computational cost to standard PCA but with substantially higher ARI. RASP was one to three orders-of-magnitude faster than other methods, highlighting its efficiency and accuracy (Fig. 5B). For optimal RASP performance, a kNN ranging from $3 - 10$ was effective; larger kNN thresholds reduced scores, as they smooth across cortical layer boundaries, creating overly large domain identification (Fig. 5D). Interestingly,all 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 parameter space. Louvain and Leiden algorithms dropped in performance as the kNN threshold increased, regardless of the β value, with $\beta = 2$ performing marginally better (Fig. 5, Extended Data Fig. 3A).

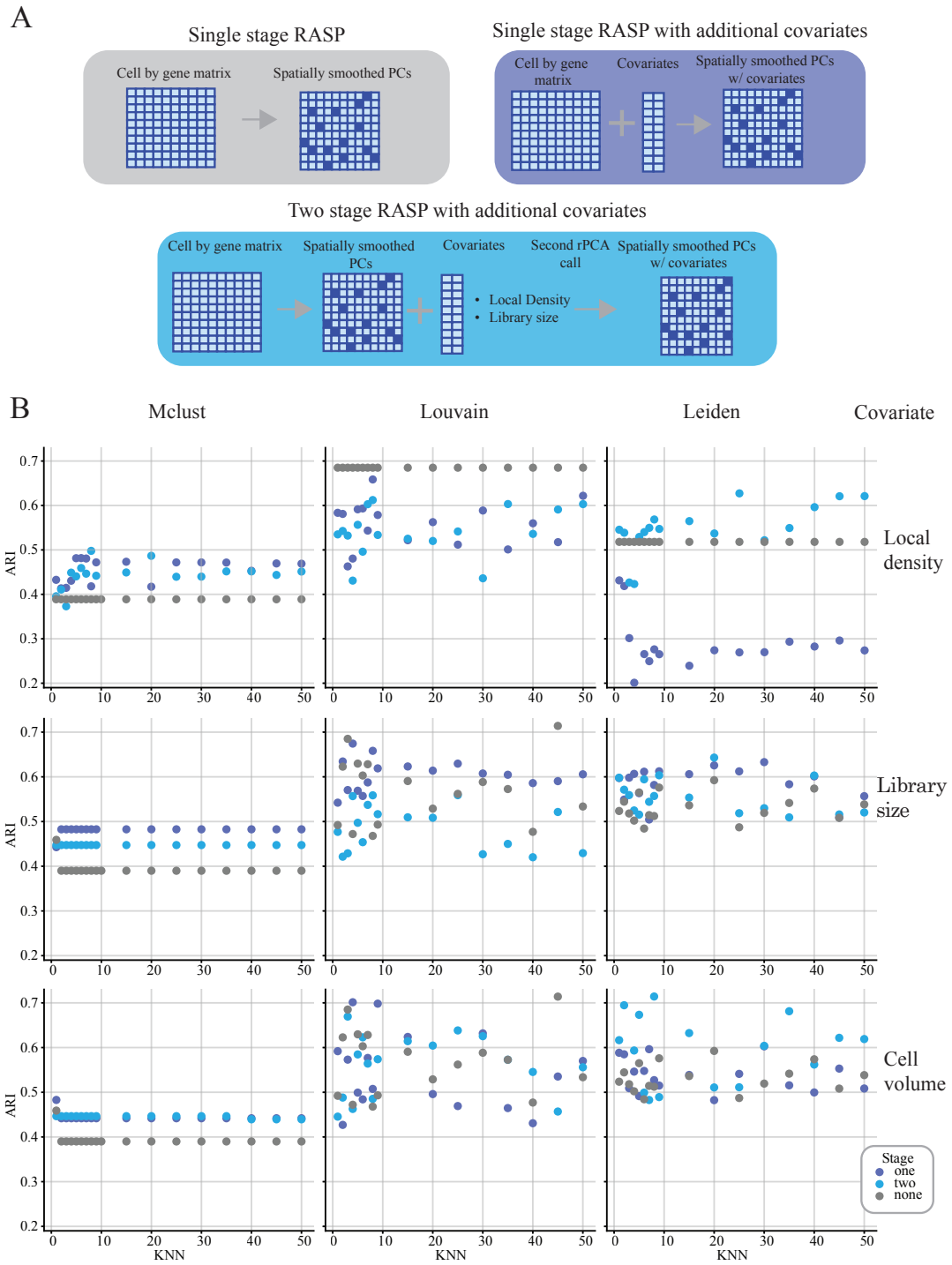

Figure 3: **Covariate analysis (Mouse ovary dataset). A:** Schematic depicting RASP model architecture with no covariates (top left), single-stage RASP architecture with an additional covariate (top right), and two-stage RASP model architecture with an additional covariate. **B:** ARI values for no covariate, 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).

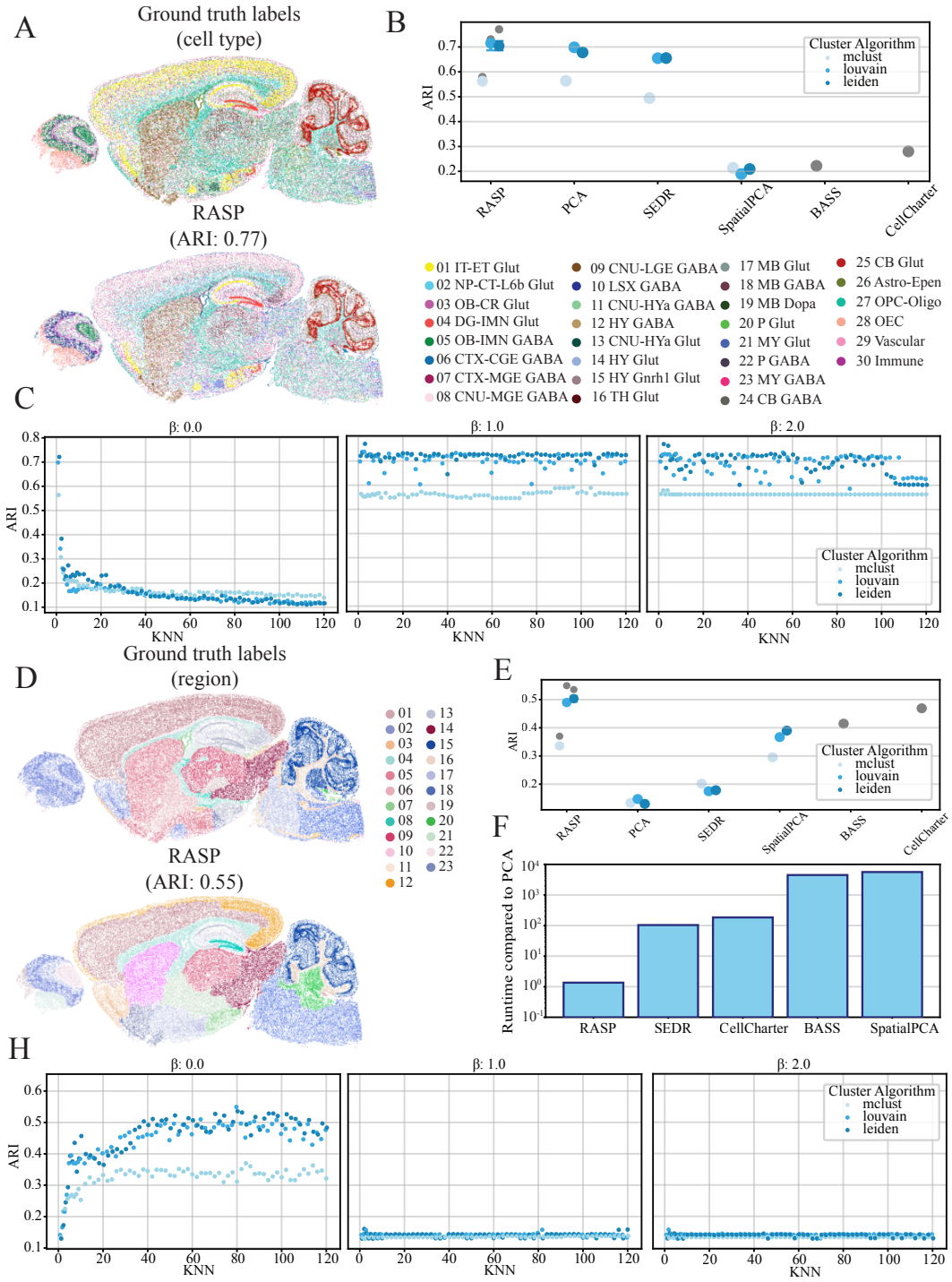

Figure 4: **RASP at performance different scales (Mouse Sagittal brain dataset). A:** Ground truth cell type annotations (top) and corresponding labels identified by RASP (bottom). **B:** ARI values for all methods computed on cell type annotations in **A** with different colors indicating the clustering algorithm used to assign labels. Interquartile range and median ARI values at default RASP parameters (kNN =2-20, $\beta = 2$) and maximum ARI values achieved by RASP are shown. **C:** 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. **D:** Ground truth region annotations and corresponding regions identified by RASP. **E:** Similar quantification to **B**, with ARI computed on the region instead of cell types annotations. Interquartile range and median ARI values at default RASP parameters (kNN =60-100, $\beta = 0$) and maximum ARI values achieved by RASP are shown. **F:** Runtime performance for all methods relative to normal PCA. **H:** ARI values for RASP computed on the region annotations pictured in **D** with inverse distance raised to the β of 0 (left), 1 (middle), and 2 (right), plotted against kNN. Colors indicate clustering algorithm.

To demonstrate the use of RASP for reduced rank reconstruction, the TMSB10 (thymosin beta 10) was reconstructed [44] (Fig. 5E). The spatially smoothed reconstruction of TMSB10 is isolated to layer 4 with noise substantially relative to normalized expression and non-smoothed reconstruction, highlighting RASP's utility for spatially informed gene-level analyses.

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

The mouse olfactory bulb analysis used a dataset generated with Stereo-seq technology, capturing subcellular tissue resolution, providing a valuable benchmark. The olfactory bulb consists of distinct cortical layers: 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) [49, 50]. RASP, SEDR, and GraphST produced clusters that effectively align with olfactory bulb layers (Fig. 2A). In contrast, STAGATE, BASS, standard PCA, and SpatialPCA were unable to characterize tissue structures in the core of the olfactory bulb.

RASP demonstrated computational costs comparable to standard PCA and two to three orders-of-magnitude faster than the other methods (Fig. 6B). Due to lack of ground truth annotations cluster quality was assessed using Moran's I and CHAOS score (Fig. 6C,D, SI Algorithm 7,8). RASP exhibited the highest Moran's I and lowest CHAOS scores among evaluated methods, with Louvain clustering producing favorable outcomes. Analysis of different kNN thresholds revealed increased kNN values resulted in higher Moran's I values and lower CHAOS scores, though beyond a kNN of 50, clusters deviated from anatomical layers, producing large homogeneous structures that, while rewarded by the CHAOS score, lack biological plausibility (Fig. 6D). RASP achieved the highest Moran's I value and lowest CHAOS score with low β values ranging from $0 - 0.5$. As the β parameter increased past this point, Moran's I decreased and the CHAOS score increased, indicating less optimal clustering solutions (Fig. 6, Extended Data Fig.5).

Reduced rank reconstruction performance was explored using the Doc2g gene, which acts upstream of neurotransmitter secretion and locates in membrane and presynapses of neurons [51, 52]. Within the olfactory bulb, Doc2g is localized to the EPL and IPL. While normalized expression of Doc2g in the original dataset appears sparse and noisy, RASP's spatially smoothed expression is clearly localized to the EPL and IPL, showing greater spatial organization than standard PCA reconstruction 6E).

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

To assess RASP on large, high-resolution data, we downloaded the 10x Xenium human breast cancer dataset and annotations from the R package SubcellularSpatialData [53]. This dataset originates from a publicly available 10x Genomics dataset of re-sectioned FFPE human breast tissue infiltrating ductal carcinoma *in situ*. Annotations are based on H&E stained tissue accompanying the Xenium data. Due to it's size, over half a million cells, competing domain detection methods encounter out-of-memory errors, requiring over 1 TB of RAM- unrealistic for many researchers, even with institutional high-performance computing environments. Given these limitations, we compared RASP against standard PCA. RASP substantially outscored PCA (ARI 0.63 vs ARI 0.25) (Fig. 7A,B).

Optimal kNN smoothing thresholds for RASP revealed a larger kNN threshold range of $50 - 100$ is needed to identify the tissue-spanning domains in this dataset (Fig.7D). For this dataset, Leiden and Louvain clustering were 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 more computationally expensive as Louvain or Leiden, but if testing multiple resolutions to achieve a desired number of clusters, the three algorithms are comparable. Similar to the olfactory bulb dataset, smaller β values $(0 - 0.5)$ produce better results, with ARI values dropping considerably for all clustering algorithms and all kNN thresholds at higher β values (Fig. 7D, Extended Data Fig. 7A).

To compare RASP against other domain detection methods, we down-sampled the breast cancer dataset to 10% of its original size and compared the ARI values against RASP (Extenedd Data Figs 6A and 6B). RASP outperformed all other methods, which generated ARI values comparable to normal PCA. We do not provide a runtime quantification against RASP for this analysis as RASP operated on the full dataset while the other methods did not.

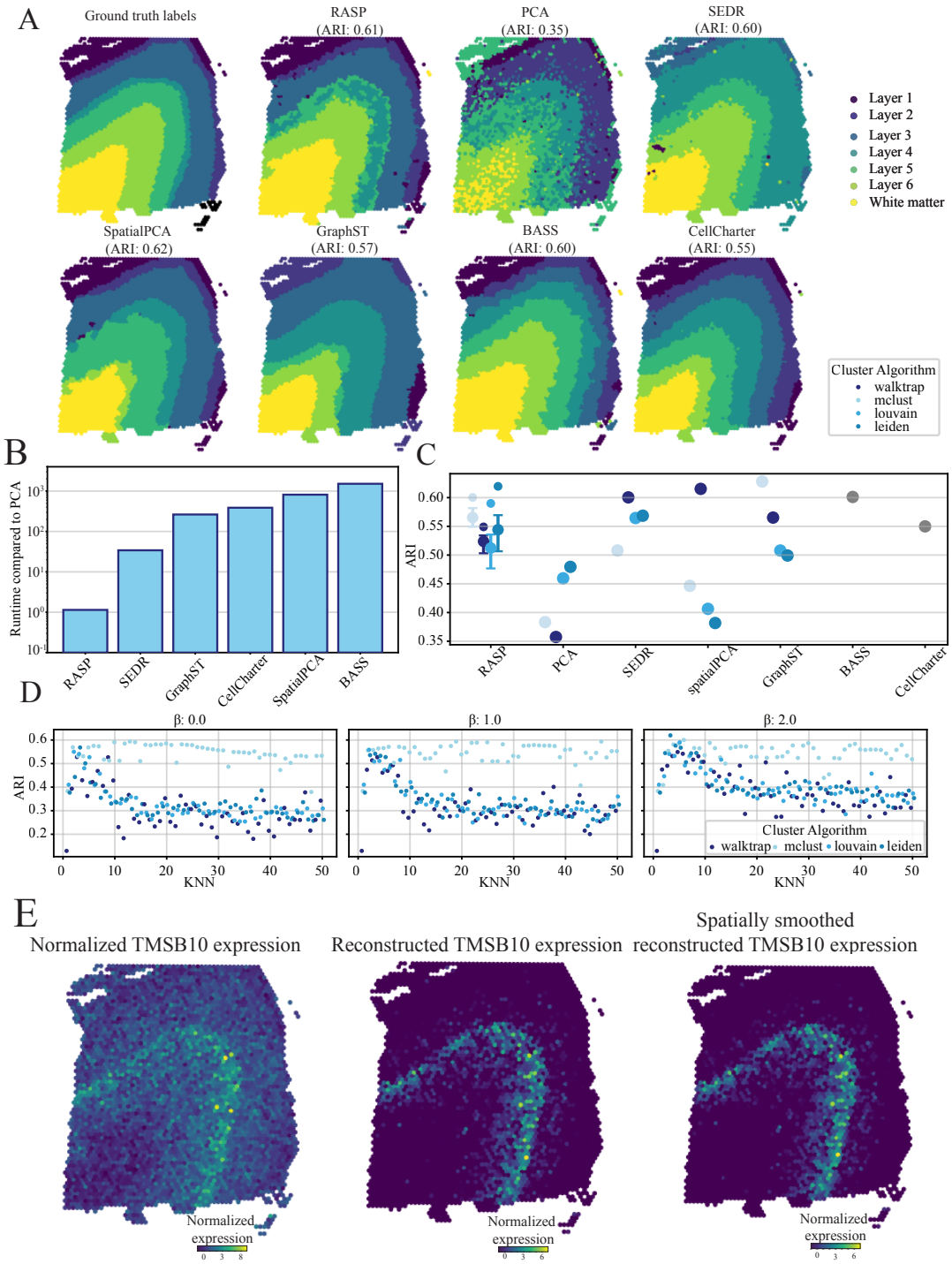

Figure 5: **Human DLPFC analysis (10X Visium). A:** Ground truth cortical layer annotations (left) and corresponding spatial domains identified by RASP, PCA, and competing methods. **B:** Quantification of runtime for all methods compared to normal PCA. **C:** Quantification of ARI for all methods, different 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.

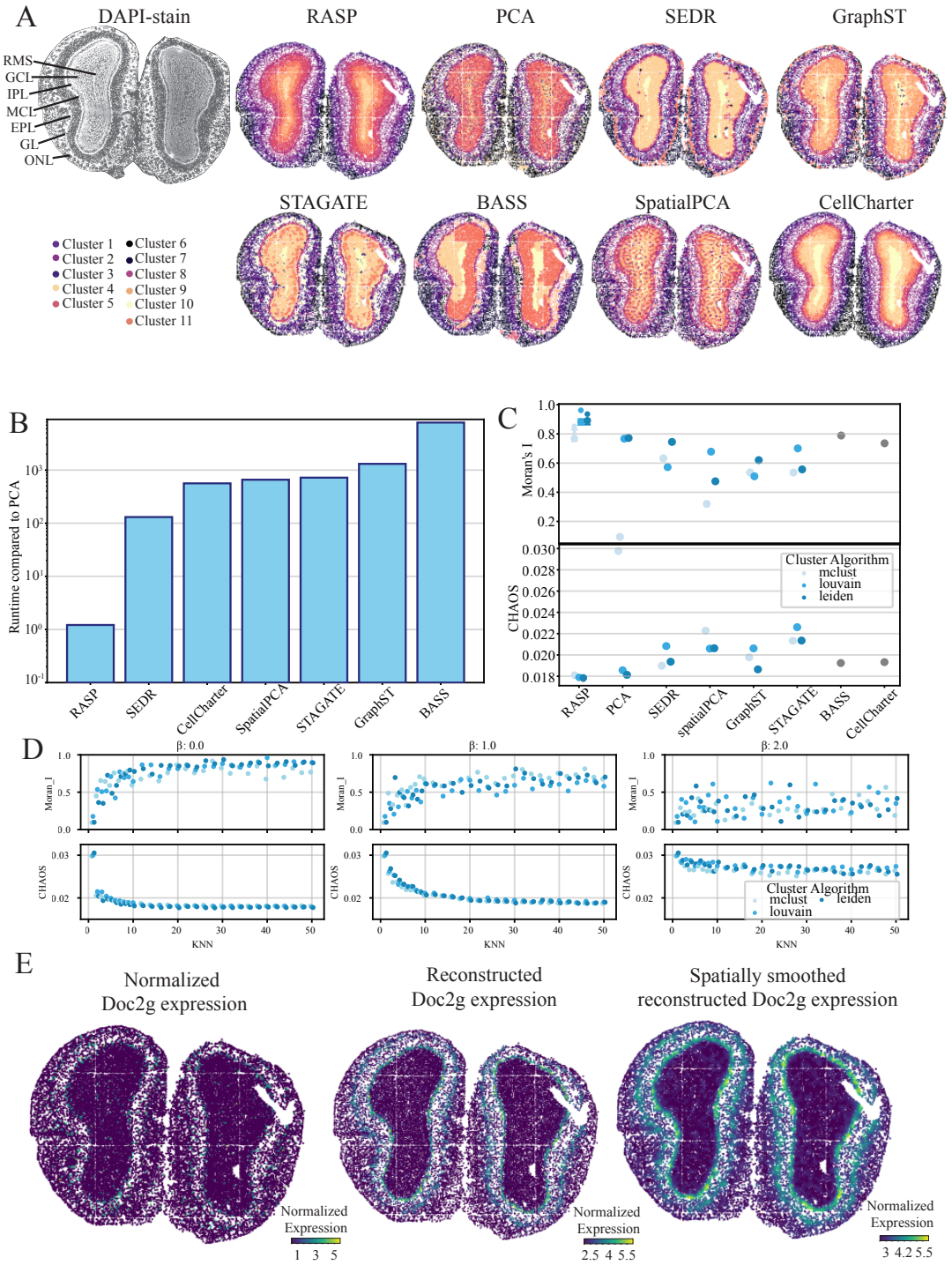

Figure 6: **Mouse olfactory bulb analysis (STOmics Stereo-seq) A:** Cortical laminar structure (top left) and corresponding spatial domains identified by RASP, PCA, and other methods. **B:** Quantification of runtime for all methods compared to normal PCA. **C:** Quantification of Moran's I score (top), 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).

To explore RASP's reduced rank reconstruction feature, we compared normalized and reconstructed expression signatures for two genes: SFRP1 and LUM. SFRP1, part of the secreted frizzled-related protein family, and modulating Wnt signaling, is dysregulated in several human tumor types [44, 54–56]. RASP's 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. 7C). LUM, a small proteoglycan family protein, binds collagen, and exhibits pathway dysregulation in many cancers [44, 57, 58]. Reduced rank reconstruction of LUM indicates high expression in both stromal and tumorous regions of the tissue.

Together, these findings highlight the unique capacity of RASP to deliver accurate spatial domain detection at scale, even for ultra-high-resolution datasets beyond the reach of competing methods. By enabling rapid parameter exploration and robust clustering on the entire Xenium dataset—without the need for sketching or specialized hardware—RASP empowers researchers to retain biological complexity and make high-confidence discoveries in challenging clinical specimens.

2.9 Simulation analysis

In addition to evaluating RASP on real data, we applied RASP and comparison techniques to simulated ST datasets generated using the **R** package **SRTsim**[59] for two different tissue models with 100 replicates each. The first model, the *Stripes* model, represents a simple tissue with a laminar structure of eight tissue domains (see Extended Data Fig. 9A). The second model, the *Dots* model, represents a complex tissue with heterogeneous circular regions scattered across the tissue, and a background cell type (Extended Data Fig. 12A).

RASP outperformed all other methods in computational speed and accuracy using default parameters. For the *Stripes* model, a larger kNN range of $15 - 50$ is needed to correctly identify the laminar tissue-spanning domains (Extended Data Fig. 9D). For clustering *Stripes* data, Leiden and Louvain were found to be optimal, followed by Walktrap and Mclust. Similar to the olfactory bulb and breast cancer datasets, smaller β values $(0 - 0.5)$ produced better results, with ARI values dropping for all clustering algorithms and kNN thresholds at higher β values (Extended Data Fig. 9D).

For the *Dots* model, a smaller kNN range of $2 - 20$ is needed to identify the regions in this dataset (Extended Data Fig. 12D). For clustering *Dots* data, Walktrap was optimal, followed by Mclust, Leiden and Louvain. Similar to the Ovary dataset, larger β values between $1.5 - 2.0$ were found to produce better results, with ARI values dropping considerably for all clustering algorithms at lower β values (Extended Data Fig. 12D). Similar to the *Stripes* model, RASP outperformed all other methods in speed and produced accurate label predictions across its default parameters.

2.10 RASP facilitates scalable, multi-resolution exploration through computational efficiency

A unique advantage of RASP is its computational efficiency, which enables the biologically meaningful exploration of ST data at multiple scales. Across a diverse collection of ST datasets, spanning multiple tissues and profiling technologies, RASP consistently outperformed or matched other state-of-the-art methods while being orders-of-magnitude faster. In the complex, cellular-resolution MERFISH ovary dataset, RASP was one-to-three orders-of-magnitude faster than competing algorithms (Fig. 2B). This speedup persisted even when annotations were spatially heterogeneous and challenging to resolve, such as immune and endothelial populations. SEDR was approximately one order-of-magnitude slower, and BASS nearly four orders-of-magnitude slower, highlighting the impact of algorithmic efficiency when analyzing large, high-resolution ST data. For datasets at anatomical or intermediate scale (e.g, the DLPFC, mouse olfactory bulb, breast cancer, and mouse brain datasets), RASP's efficiency enabled practical multi-scale analysis that would otherwise be prohibitive due to memory and runtime constraints. Notably, in the breast cancer dataset containing over half a million cells, RASP readily analyzed the full dataset, a task that caused other methods to fail due to memory limits (requiring over 1 TB of RAM to train the model).

The computational speed of RASP allows users to systematically and interactively explore the effects of the kNN threshold and β exponent, supporting robust identification of optimal domain solutions tailored to the driving biological questions that motivated the experiment. For instance, in the ovary dataset, local smoothing (small kNN) and larger β yielded superior performance for cell type annotation, while higher kNN was necessary to delineate tissue-scale domains in breast cancer. Because RASP's runtime is so low relative to alternative techniques, researchers are not forced to compromise on parameter search or multi-scale analysis, but can instead comprehensively interrogate tissue structure at all relevant spatial scales. Importantly, the rPCA embedding

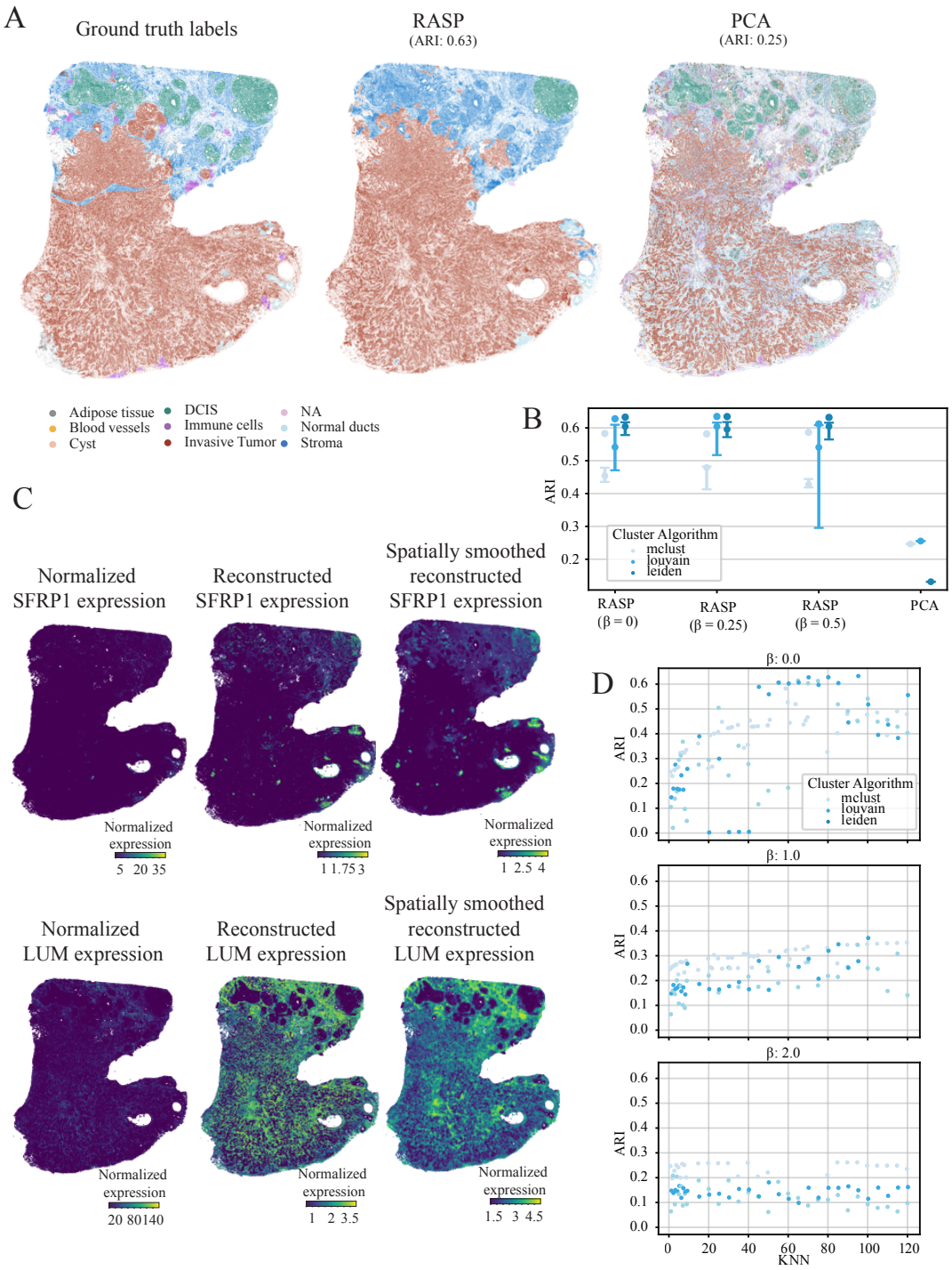

Figure 7: **Human breast cancer analysis (10x Xenium). A:** Ground truth spatial domain annotations (top left) and corresponding predictions for RASP, and normal PCA. **B:** Quantification of ARI for all methods, different colors indicate the clustering algorithm used to assign labels. Interquartile 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.

need only be computed once: subsequent parameter sweeps simply rebuild the sparse distance matrix, apply the inverse-distance weighting to the fixed rPCA coordinates, and rerun clustering. To guide choice of optimal settings in a label-free manner, users can compute spatial-autocorrelation statistics (e.g. Moran's I) or boundary-coherence measures like the CHAOS score for each domain solution, and then select the (kNN,) combination that maximizes these metrics. This workflow makes it trivial to identify biologically meaningful multi-resolution domains without ever repeating the costly dimension-reduction step. In sum, RASP's computational efficiency not only matches or exceeds state-of-the-art clustering accuracy, but also fundamentally enables the multi-resolution, exploratory analyses needed to reveal the full complexity of spatial tissue architectures.

2.11 PC number impacts RASP performance

For the real ST analyses detailed in Sections 2.3-2.8, 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, analyses typically use a default of between 20 and 50 PCs. Tools like **Scanpy** and **Seurat** default to retaining 30 and 50 PCs respectively, serving as a useful benchmarks.

To evaluate robustness, we ran 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 an increase number of retained PCs, from $5-60$, to understand the correlation between PC count and cluster label accuracy (Extended Data Fig. 15). For the DLPFC dataset, stable ARI values were observed with increasing PC number for the Louvain, Leiden, and Mclust clustering methods, and an increase for Walktrap. This invariance suggests these clustering models describe data distribution well regardless of dimensions. Conversely, in the mouse ovary dataset, we noted an upward ARI trend with more PCs for 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.16A, middle). Interestingly, Mclust improved at the lowest PC count, likely because Mclust assumes clusters have consistent size and shape, which isn't the case here. In the breast cancer dataset, Louvain and Leiden algorithms showed increased ARIs up to 30 PCs, and then values decreased, while Mclust ARIs decreased as the number of PCs increased. This reduction in classification accuracy aligns with the nature of the classification task in this dataset, which deals with more extensive spatial domains rather than finer cellular granularity, thereby indicating that broader spatial domains might be more appropriately captured with fewer PCs (Extended Data Fig.16A, right).

2.12 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 Figs. 17 and 18). 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 down weighing 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. 17B). 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.

2.13 Differential distance metrics and weighting analysis

For the RASP results presented in Sections 2.3-2.8, spatial smoothing was performed using a sparse weighting matrix where each element is computed as the inverse Euclidean distance (raised to the power of β) between ST locations, with diagonal elements set to either the minimum distance (or half the minimum distance for Visium technology). We choose Euclidean distance for a number of reasons. First, the raw spatial coordinates generated by any imaging-based ST platform (Visium, Xenium, MERFISH, Stereo-Seq, etc.) are expressed in real-space units (microns), and Euclidean distance is the natural metric by which physical proximity on a 2D tissue section is

measured. Second, Euclidean geometry is both isotropic and translation-invariant: cells or spots that are offset by the same vector are treated identically regardless of orientation, ensuring unbiased smoothing across the tissue. Third, from a computational standpoint Euclidean distances can be computed efficiently and are straightforward to threshold or sparsify—critical for scaling to high-resolution, subcellular datasets. The inverse distance weighting of the Euclidean distance captures the biologically plausible principle that the strength of spatial relationships decays continuously with physical separation, with the β parameter allowing flexible modeling of interaction ranges, from highly local (large β) to more diffuse (small β).

To systematically evaluate the impact of both the distance metric and the weighting kernel, we evaluated Euclidean, Manhattan, and Chebychev distances as well as kernel-based weightings (Gaussian, quadratic), in addition to the standard inverse distance method, on both spot-based (DLPFC Visium) and an imaging-based (ovary MERFISH) ST datasets (Extended Data Figs. 19 and 20). For each combination, we conducted clustering on the spatially smoothed principal components, and computed ARI relative to ground truth annotations, sweeping across a range of kNN and β parameters (as well as kernel decay settings for the alternative weightings).

Across both datasets, our results demonstrate that no alternative metric or kernel provided a consistent or significant improvement over inverse Euclidean distance weighting. In all cases, ARI values were highly similar for different distance and kernel choices across the full parameter grid, indicating that the simpler and biologically motivated approach of inverse Euclidean weighting is both effective and robust. These findings justify our default choice of Euclidean distance and inverse distance weighting for RASP, balancing biological plausibility, computational efficiency, and empirical performance.

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.

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 2 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. This is important because it allows users to leverage a priori knowledge to incorporate biologically relevant covariates that are not present in the transcriptomic data in a flexible manner, functionality other tools lack.

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 Table 2). 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 SpatialPCA 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 [60–62]. PALMO employs variance decomposition analysis and incorporates some spatial modeling capabilities [63]. In contrast, methods such as PASTE2 and DeST-OT utilize optimal transport principles to align sections from the same tissue sample, enabling 3D models [64, 65]. 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 [66, 67]. 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 [68]. 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 [34]. 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 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 [34].

4.4 Application of RASP to existing ST datasets

4.4.1 Dataset descriptions

We applied RASP on four publicly available ST datasets, as well as two simulated ST, that were generated using different techniques and resolutions on human and mouse tissues. Details of data preprocessing can be found in Supplementary Information, details on data simulation can be found in 4.4.4. In brief, raw count matrices where filtered to remove cells/locations with mitochondrial and ribosomal RNA $> 10\%$, cells/locations with < 100 reads

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

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 $\mathbf{X_r}$

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 [49]. 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 [53]. 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 [48].

The Sagittal adult mouse brain dataset (Vizgen MERFISH technology) consisted of 81,452 cells and 1,122 genes. The brain section is part of the Allen Institute's Mouse Brain Atlas originally described in [1],[47]. The dataset includes region, cell class and cell type annotations. For ARI calculations we used the cell type level 'class', and division level 'parcellation division' labels. See 4.4.4 for detailed instructions to access this data resource.

The two simulated datasets (*Stripes*, and *Dots*) consisted of 10,000 cells and 150 genes each.

4.4.2 RASP evaluation

To evaluate the RASP method, we applied the algorithm to the processed gene expression matrices of the six 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 and simulated datasets. 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), STAGATE pyG (Python, version 1.0.0), and CellCharter (Python, version 0.3.5)[24, 26–30]. 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. 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. Finally, CellCharter is a domain detection framework that utilizes Variational Autoencoders (VAEs) to reduce dimentionality and remove batch effects, and then clusters the latent space using iterative Gaussian Mixture Modeling (GMM) to determine k clusters and cluster identities. 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 25 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.

4.4.4 Generation of simulated ST data

We simulated ST data according to two different tissue models, *Stripes* and *Dots*, using the **SRTsim** package's GUI [59], following the reference free example provided at `https://xzhoulab.github.io/SRTsim/03_Reference_Free_Example/`. The *Stripes* model represents a laminar tissue with eight vertical domains (Extended Data Fig 8A). The *Dots* dataset represents a more complex tissue architecture that includes circular clusters of cells of varying sizes, small striped regions, and a large proportion of background cells making up a background region (Extended Data Fig 11A). Both models include eight annotated spatial domains. Count data was simulated according to these models for $10,000$ cells, 50 low signal genes, 50 high signal genes and 50 noise genes using a zero inflated negative binomial distribution with $\mu = 2$, zero $\% = 0.5$, and $\theta = 0.5$. Each annotated tissue region was assigned a unique fold change value (relative to μ) to simulate differences in spatial domains seen in real tissue. We passed the *Stripes* and *Dots* models to the `reGenCountshiny` function with different *seed_number* values to generate 100 replicate count matrices for each. For model details, spatial coordinates and domain labels for the simulated datasets please refer to 4.4.4.

For details on how the simulated data was processed prior to running RASP see SI 1.6.

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 [69] (`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`).

Sagittal mouse brain:

We downloaded the mouse brain dataset from the Allen Brain Atlas via their **abc_atlas_access Python** package. Specifically we followed the *MERFISH whole mouse brain spatial transcriptomics (Xiaowei Zhuang)* tutorial `https://alleninstitute.github.io/abc_atlas_access/notebooks/zhuang_merfish_tutorial.html` to access section 3.010 from the 'Zhuang-ABCA-3' dataset.

Simulated datasets and RASP implementation: Please visit our GitHub to access the simulated datasets and their associated models.

Please visit our GitHub for details on RASP implementation: `https://github.com/Goods-Lab/RASP`.

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. Varrone, M., Tavernari, D., Santamaria-Martínez, A., Walsh, L. A. & Ciriello, G. CellCharter reveals spatial cell niches associated with tissue remodeling and cell plasticity. *Nature Genetics* **56.** Publisher: Nature Publishing Group, 74–84. ISSN: 1546-1718. https://www.nature.com/articles/s41588-023-01588-4 (2025) (Jan. 2024).

31. Singhal, V. *et al.* BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis. *Nature Genetics* **56.** Publisher: Nature Publishing Group, 431–441. ISSN: 1546-1718. https://www.nature.com/articles/s41588-024-01664-3 (2025) (Mar. 2024).

32. Schaub, D. P. *et al.* PCA-based spatial domain identification with state-of-the-art performance. *Bioinformatics* **41,** btaf005. ISSN: 1367-4811. https://doi.org/10.1093/bioinformatics/btaf005 (2025) (Jan. 1, 2025).

33. Yuan, Z. MENDER: fast and scalable tissue structure identification in spatial omics data. *Nature Communications* **15.** Publisher: Nature Publishing Group, 207. ISSN: 2041-1723. https://www.nature.com/articles/s41467-023-44367-9 (2025) (Jan. 5, 2024).

34. 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).

35. 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).

36. 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).

37. 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).

38. 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).

39. 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).

40. 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).

41. 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).

42. 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).

43. 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).

44. 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).

45. 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).

46. 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).

47. Yao, Z. *et al. A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain* Pages: 2023.03.06.531121 Section: New Results. Mar. 6, 2023. https://www.biorxiv.org/content/10.1101/2023.03.06.531121v1 (2025).

48. 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).

49. 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).

50. 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).

51. 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).

52. 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).

53. 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).

54. 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).

55. 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).

56. 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).

57. 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).

58. 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).

59. Zhu, J., Shang, L. & Zhou, X. SRTsim: spatial pattern preserving simulations for spatially resolved transcriptomics. *Genome Biology* **24,** 39. ISSN: 1474-760X. https://doi.org/10.1186/s13059-023-02879-z (2025) (Mar. 3, 2023).

60. 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).

61. 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).

62. 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).

63. 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).

64. 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).

65. 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).

66. 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).

67. 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).

68. 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).

69. 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).