Probabilistic embedding, clustering, and alignment for integrating spatial transcriptomics data with PRECAST

Wei Liu[1†], Xu Liao[1†], Ziye Luo[2,1], Yi Yang[1], Mai Chan Lau[3], Yuling Jiao[4], Xingjie Shi[5], Weiwei Zhai[6], Hongkai Ji[7], Joe Yeong[3,8], and Jin Liu[1*]

[1]Centre for Quantitative Medicine, Health Services & Systems Research, Duke-NUS Medical School

[2]School of Statistics, Renmin University, China,

[3]Institute of Molecular and Cell Biology (IMCB), Agency of Science, Technology and Research (A*STAR), Singapore,

[4]School of Mathematics and Statistics, Wuhan University, Wuhan, China,

[5]Academy of Statistics and Interdisciplinary Sciences, East China Normal University, Shanghai, China,

[6]Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, China

[7]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, USA

[8]Department of Anatomical Pathology, Singapore General Hospital, Singapore,

Abstract

Spatially resolved transcriptomics (SRT) involves a set of emerging technologies that enable the transcriptomic profiling of tissues with the physical location of expressions. Although a variety of methods have been developed for data integration, most of them are for single-cell RNA-seq datasets without any consideration of spatial information. Thus, methods that are capable of integrating spatial transcriptomics data from multiple tissue slides, possibly from multiple individuals, are sincerely needed. Here, we present PRECAST, an efficient data integration method for multiple spatial transcriptomics datasets with complex batch effects and/or biological effects between slides. PRECAST

*Corresponding author. Email: jin.liu@duke-nus.edu.sg

unifies spatial factor analysis simultaneously with spatial clustering and embedding alignment, while requiring only partially shared cell/domain clusters across datasets. Using both simulated and four real spatial transcriptomics datasets from the 10x Visium, ST, and Slide-seqV2 platforms, we showed improved cell/domain detection with outstanding visualization, and the estimated aligned embeddings of cellular biological effects and cell/domain labels facilitate many downstream analyses. Furthermore, the slide-specific embeddings estimated by PRECAST can be utilized to study the spatial impact of neighboring microenvironments. PRECAST is computationally scalable and applicable to SRT datasets from different platforms.

Introduction

Spatially resolved transcriptomics (SRT) encompass a set of recently developed technologies that characterize the gene expression profiles of tissues while retaining information on their physical location. The methodologies used for resolving spatial gene expression are primarily categorized into *in situ* hybridization (ISH) technologies (e.g., MERFISH [1–3], seqFISH [4, 5], seqFISH+ [6]), and *in situ* capturing technologies (e.g., ST [7], HDST [8], Slide-seq [9, 10] and 10x Genomics Visium [11]) [12]. The *in situ* capturing technologies are unbiased and involve transcriptome-wide expression measurements, while ISH-based methods are targeted and require prior knowledge of the genes of interest. These technologies have provided extraordinary new opportunities for researchers to characterize the transcriptomic landscape within a spatial context; explore how cells influence and are influenced by the cells around them [13]; identify genes with spatial variations other than cell/domain differences, e.g., cell morphology [14]; and identify spatial trajectories or RNA velocity in tissues [15, 16], among other applications [17].

Similar to single-cell RNA-sequencing (scRNA-seq) studies, in SRT studies of a single slide, identifying the cell/domain clusters for each spot with the collation of both spatial information and expression measurements is an important step [18–20]. Recently, multiple studies have involved the analysis of SRT datasets from multiple slides, requiring to further remove unwanted variations from different batches. For example, SRT profiles were characterized in 12 human cortex tissue slides from three adult donors using 10x Visium [21] and in multiple sections from a mouse olfactory bulb (OB) that were equally distributed along the anterior-posterior axis of the same mouse using Slide-seqV2 [22]. Moreover, when multiple SRT datasets from different clinical/biological conditions are available, integrative analysis to estimate shared embeddings of expressions representing variations between cell/domain types can provide the first step towards detecting genes that are differentially expressed between conditions [23]. Thus, it is important to develop rigorous methods that are capable of performing data integration across multiple SRT datasets by aligning shared embeddings of biological effects between cell/domain types while accounting for complex batch effects and/or biological effects between slides [24]

An ideal data integration method for SRT datasets should be capable of the following three tasks: (1) estimating the shared embeddings of biological effects between cell/domain types across SRT datasets and slide-specific embeddings that account for local microenvironments (spatial dimension reduction); (2) aligning the shared embeddings that capture cellular biological variation across datasets with heterogeneous batch effects and/or biological effects between slides (data alignment); (3) clustering the aligned embeddings to obtain cell/domain

clusters across datasets that promote spatial smoothness (spatial clustering). Most existing data integration methods, including MNN [24], Scanorama [25], Seurat V3 [26], Harmony [23], scVI [27], and scGen [28], have been developed for scRNA-seq datasets without any consideration of spatial information. More recently, MEFISTO was proposed as a way to analyze datasets with repeated spatio-temporal measurements [29], and PASTE was proposed as a method to stack and/or integrate SRT data from multiple adjacent tissue slices into a single slide, but it is not applicable to the integration of tissue sections from different individuals [30]. In addition, most existing methods perform data integration in low-dimensional space using the principal components (PCs) of conventional dimension reduction, e.g., principal component analysis (PCA), without considering the consistent loss functions of dimension reduction, alignment across datasets, or spatial clustering.

To address the challenges presented by SRT data integration and facilitate the downstream analyses of combinations of multiple tissue slides, we propose the use of a unified and principled probabilistic model, PRECAST, to simultaneously estimate low-dimensional embeddings for biological effects between cell/domain types, perform spatial clustering, and most importantly, align embeddings for normalized gene expression matrices from multiple tissue slides. As a result, PRECAST can resolve aligned representations, provide outstanding visualizations, and achieve higher spatial clustering accuracy for combined tissue slides. The resolved aligned representations and estimated labels from PRECAST can be used in multiple downstream analyses, e.g., removing batch effects, identifying differentially expressed genes under different conditions/stages, and recovering spatial trajectories/RNA velocity, etc. In addition, PRECAST uniquely estimates slide-specific embeddings that capture the spatial dependence in neighboring cells/spots, providing an opportunity to understand the spatial impact of various microenvironments. We illustrate the benefits of using PRECAST through extensive simulations and analysis of a diverse range of example datasets collated with different spatial transcriptomics technologies: 10x Visium datasets of 12 human dorsolateral prefrontal cortex (DLPFC) samples and four hepatocellular carcinoma (HCC) samples, ST datasets from eight mouse liver tissue sections and Slide-seqV2 datasets from 16 OB tissue slides.

Results

Spatial transcriptomics data integration using PRECAST

Unlike other integration methods that take as input the top (spatial) PCs and apply multiple steps to remove batch effects and merge the data, PRECAST takes as input the normalized gene expression matrices from multiple tissue slides, factorizes the input to each matrix into a latent factor with a shared distribution in each cell/domain cluster while simultaneously performing spatial dimension reduction and spatial clustering, and aligning and estimating joint embeddings for biological effects between cell/domain types across multiple tissue slides (Fig. 1a). In the dimension-reduction step, we used an intrinsic conditional autoregressive (CAR) component to capture the spatial dependence induced by neighboring microenvironments while in the spatial-clustering step, we used a Potts model to promote spatial smoothness within spot neighborhoods in the space of cluster labels. PRECAST applies a simple projection strategy to non-cellular biological effects, e.g., batch effects and/or biological effects between

slides, and implicitly accounts for shifts in the centroid of each cluster using the intrinsic CAR component in the dimension-reduction step (see Methods). These considerations of PRECAST mimic some of the recent explorations into self-supervised learning and domain adaptation in deep learning but in a parametric manner. We show that PRECAST outperforms existing data integration methods by more successfully aligning similar clusters across multiple tissue slides while separating clusters with outstanding visualization. Uniquely, PRECAST estimates slide-specific embeddings for spatial dependence among neighboring cells/spots, providing an opportunity to explore the impact of neighboring microenvironments.

PRECAST can be applied to SRT datasets of various resolutions. By increasing the resolution for SRT datasets, PRECAST can reveal fine-scale cell-type distributions. In the following analysis, datasets in Slide-seqV2 with near-single-cell resolution present spatial patterns with much more "noise" than those of Visium due to the heterogeneity of the cell-type distributions. When we merged nearby beads in the Slide-seqV2 datasets, the recovered spatial patterns resembled those from Visium.

We use the estimated cell/domain labels and finely aligned embeddings to showcase some downstream analyses, as depicted in Fig. 1b. First, users can visualize the inferred embeddings for biological effects between cell/domain types using two components from either tSNE [31] or UMAP [32]. Second, because the aligned embeddings obtained from PRECAST only carry information on the biological differences between cell/domain types, we provide a module to recover the gene expression matrices with batch effects removed and further identify genes differentially expressed either across different cell/domain clusters and/or under different conditions. Third, using these embeddings, we can identify genes whose spatial variability is not just due to biological differences between cell/domain types. Fourth, with the aligned embeddings estimated by PRECAST, we can perform trajectory inference/RNA velocity analysis to determine either the pseudotime or the pattern of dynamics in spatial spots across multiple tissue slides.

Validation using simulated data

We performed comprehensive simulations to evaluate the performance of PRECAST and compare it with that of several other methods (Fig. 1c and Supplementary Fig. S1-S2). Specifically, we considered the following eight integration methods: Harmony [23], Seurat V3 [26], fastMNN [24], scGen [28], Scanorama [25], scVI [27], MEFISTO [29] and PASTE [30], all of which can be used to estimate the aligned embeddings among samples, except PASTE, which can only estimate the embeddings of the center slice. The simulation details are provided in the Methods section. Briefly, we simulated either the normalized or count matrices of gene expression to compare PRECAST with other methods in terms of performance in data integration, dimension reduction, and spatial clustering. To mimic real data, we also considered the two ways to generate spatial coordinates and cell/domain labels, i.e., Potts models and real data in three DLPFC Visium slices. Using real data in DLPFC, we considered slides either from different or the same donor. In total, we considered five scenarios: (1) Potts + logCount; (2) Potts + Count; (3) DLPFC (slides from different donors) + logCount; (4) DLPFC (slides from different donors)+ Count; and (5) DLPFC (slides from the same donor) + Count. Scenario 5 was designed to compare PRECAST with PASTE, as PASTE can only handle SRT

data from multiple adjacent tissue slices (see Methods).

To quantify the performance of the data integration, we calculated the F1 scores of the average silhouette coefficients, which summarized two similar metrics of silhouette coefficients into a single quantity, and two versions of the local inverse Simpson's Index (LISI): integration LISI (iLISI) and cell-type LISI (cLISI). iLISI was employed to assess integration mixing, while cLISI was employed to assess the separation of each domain cluster. In Scenarios 1-4, PRECAST achieved the best performance in data integration, with the highest F1 scores for the average silhouette coefficients, and in Scenario 5, PRECAST was comparable to PASTE (Fig. 1c, top panel). Moreover, PRECAST was able to merge spots within common cell/domain clusters across datasets, while separating spots from different cell/domain clusters (small cLISI, Supplementary Fig. S1a, top panel) and, at the same time, maintaining a sufficient mix of different samples (large iLISI, Supplementary Fig. S1a, bottom panel). In contrast, the existing integration methods were either sensitive to the data generation process or failed to recover the spatial structure during integration (Fig. 1c, top panel, and Supplementary Fig. S1a).

We then evaluated the performance of estimating the embeddings induced by neighboring microenvironments in PRECAST. The estimated embeddings due to neighboring microenvironments were highly correlated with the underlying truth (Fig. 1c, middle panel), suggesting PRECAST was able to well recover the spatial dependence of spots with microenvironments. Next, we evaluated the performance of the models in obtaining embeddings for biological effects between cell/domain types. For the average canonical correlation between the estimated aligned embeddings and the true latent features, PRECAST ranked the highest in most scenarios (Supplementary Fig. S1b, top panel), suggesting the aligned embeddings estimated by PRECAST were more accurate. We also showed that Pearson's correlation coefficients between the observed expression and the estimated cell/domain labels conditioned with the aligned embeddings from PRECAST were lower than those from other methods (Supplementary Fig. S1b, bottom panel), suggesting PRECAST captured more relevant information regarding cell/domain clusters and, thus, facilitated downstream analysis.

Last, we compared the clustering performance of each method. As a unified method, PRECAST simultaneously estimates embeddings for biological effects between cell/domain types and cluster labels. For the other methods, we sequentially performed spatial clustering based on each of the estimated embeddings using SC-MEB/Louvain, except for Seurat V3, which has its own clustering pipeline based on Louvain. PRECAST achieved the highest adjusted Rand index (ARI) and normalized mutual information (NMI) in all considered scenarios (Fig. 1c, bottom panel, and Supplementary Fig. S2a), while the other methods, such as Seurat V3 and Harmony, were sensitive to the data generation process. Moreover, we observed that all methods correctly chose the number of clusters. Finally, we showed that PRECAST was computationally efficient and scalable to large sample sizes, and it only took ∼6 hours to analyze a dataset with 2,000 genes and 600,000 spots for a fixed number of clusters (Supplementary Fig. S2b).

Application to human dorsolateral prefrontal cortex Visium data

We applied PRECAST and the other methods to the analysis of four published datasets obtained via either Visium, ST, or Slide-seqV2 technologies (see Methods). By obtaining the

estimated aligned embeddings and cluster labels from PRECAST, we could perform many downstream analyses using all tissue slides. Here, we showcase the differential expression (DE) analyses across detected domains, spatial variation analysis (SVA) adjusting for aligned embeddings as covariates, and trajectory inference/RNA velocity analysis. To examine the clustering performance with low-resolution data, we performed deconvolution analysis to infer the cell compositions of the domains detected by PRECAST.

To quantitatively show that PRECAST outperforms existing data integration methods, we first analyzed LIBD human DLPFC data generated using 10x Visium [21] that contained 12 tissue slices from three adult donors, comprising four tissue slices from each donor. In all 12 tissue slices, the median number of spots was 3,844, and the median number of genes per spot was 1,716. The original study provided manual annotations for the tissue layers based on the cytoarchitecture that allowed us to evaluate the performance of both the data integration and accuracy of spatial domain detection by taking the manual annotations as ground truth. For each method, we summarized the inferred embeddings for biological effects between cell/domain types using three components from either tSNE or UMAP and visualized the resulting tSNE/UMAP components with red/green/blue (RGB) colors in the RGB plot (Fig. 2a, right-top panel, and Supplementary Fig. S3a & b to S5a & b). The resulting RGB plots from PRECAST showed the laminar organization of the human cerebral cortex, and PRECAST provided smoother transitions across neighboring spots and spatial domains than those from other methods. For each of the other methods, we further performed clustering analysis to detect spatial domains using different methods with the inferred embeddings (Fig. 2a, right-bottom panel, and Supplementary Fig. S3c-S5c). We observed that the results from PRECAST had stronger laminar patterns and the estimated aligned embeddings carried more information about the domain labels (Supplementary Fig. S6a).

A unique feature of PRECAST is its ability to estimate slide-specific embeddings capturing spatial dependence in the neighboring cells/spots due to various neighboring microenvironments in different regions. Supplementary Fig. S7 provides RGB plots of the inferred embeddings for spatial dependence using three components from either tSNE or UMAP. We observed that spots in the domain of white matter had similar microenvironments, while spots in layers 1 to 6 had two distinct microenvironment patterns from left to right, suggesting potentially distinct functions in the left and right regions of layers 1 to 6.

PRECAST can offer outstanding data visualization compared with other methods. We visualized the inferred embeddings for biological effects between cell/domain types using two components from tSNE for each method (Fig. 2b and Supplementary Fig. S6b). The tSNE plots for PRECAST show that spots from different slices were well mixed (Fig. 2b, top panel) while the domain clusters were well segregated (Fig. 2b, bottom panel), and there were significant improvements in visualization in comparison to the other methods, including the method applied with no corrections. Supplementary Fig. S6c shows that PRECAST achieved the best data integration in terms of F1 scores, iLISI, and cLISI. To evaluate the clustering accuracy, we used both ARI and NMI. As shown in Fig. 2c and Supplementary Fig. S6d, PRECAST achieved the highest ARI and NMI for the separate evaluation and combined evaluation: the median ARI was 0.434 for PRECAST, 0.382 for Scanorama, and 0.406 for scVI in the separate evaluation; and the ARI was 0.374 for PRECAST, 0.216 for Scanorama, and 0.301 for scVI in the combined evaluation. A heatmap of Pearson's correlation coefficients among the

detected domains shows the good separation of the estimated aligned embeddings across domains (Fig. 2d) and the correlations between deeper layers were high, e.g., there were high correlations between layers 5 and 6, while correlations among the separated layers were low.

A key benefit of PRECAST is its ability to estimate aligned embeddings for biological effects between cell/domain types and joint labels for all slides. We performed DE analysis for the combined 12 slices (see Methods). In total, we detected 1,331 DE genes with adjusted p-values of less than 0.001 among the 10 spatial domains identified by PRECAST, with 314 genes being specific to Domain 5, which corresponded to white matter (Supplementary Table S1). Many of these genes were reported to be enriched in different layers of DLPFC, i.e., *Pcp4* (Domain 2, layer 5) [19], *Hopx* (Domain 1, layer 1/3), and *Enc1* (Domain 4, layer 2/3) [33] (Fig. 2e and Supplementary Fig. S8-S10). Next, we performed trajectory inference using the aligned embeddings and domain labels estimated by PRECAST (Supplementary Fig. S11). In total, we identified 858 genes associated with the estimated pseudotime with adjusted p-values of less than 0.001. Among them, 373 were identified as DE genes in at least one domain (Fig. 2f and Supplementary Table S2). We further found that the pseudotime-associated genes identified by PRECAST were significantly enriched for nervous system development (Fig. 2g).

To further show that the estimated embeddings for biological effects between cell/domain types from PRECAST were well aligned across tissue slides, we performed SVA analysis with the aligned embeddings from PRECAST as covariates for each slice to identify spatially variable genes (SVGs) with nonlaminar patterns. A detailed list of the genes identified at a false discovery rate (FDR) of 1% is available in Supplementary Table S3. Interestingly, many of the identified genes were related to immune function. For example, *Isg15* encodes a ubiquitin-like molecule induced by type I interferon, and ISG15 deficiency increases antiviral responses in humans [34, 35]. Many studies in the literature have highlighted the importance of immune–brain interactions to the development of many disorders of the central nervous system [36, 37]. Further enrichment analysis showed the genes from each slice to be highly enriched in many common pathways, suggesting the embeddings for biological effects between cell/domain types were effectively aligned by PRECAST (Fig. 2h and Supplementary Fig. S12-S16).

Application to mouse liver ST data

We further applied PRECAST and other methods to analyze eight sections of wild-type adult, female mouse livers from the caudate and right liver lobes of three female mice using ST technology [38]. In all eight sections, the median number of spots was 640, and the average number of genes was 15,302. The original study provided manual annotations based on marker genes that allows us to evaluate the performance of both the data integration and accuracy of spatial domain detection by taking the manual annotations as ground truth. For each method, we summarized the clustering performance of each section and combined sections using both ARI and NMI (Fig. 3a, top panel, and Supplementary Fig. S17a). PRECAST achieved the highest ARI and NMI in both cases: in each separate section, the median ARI was 0.24 for PRECAST, 0.18 for Seurat V3, and 0.02 for scVI; and jointly, in the combined sections, the value of ARI was 0.23 for PRECAST, 0.18 for Seurat V3, and 0.02 for scVI. We visualized the cluster labels obtained by PRECAST and other methods as well as the manual annotations (Supplementary Fig. S18) and found PRECAST performed best for each

individual sample. On the other hand, PRECAST achieved better data integration than most of the other methods in terms of F1 score, iLISI, and cLISI (Fig. 3a, bottom panel, and Supplementary Fig. S17b) with comparable conditional correlations (Supplementary Fig. S17c). The tSNE plots for PRECAST show that spots from different sections were well mixed (Fig. 3b and Supplementary Fig. S17d, top panel), while the domain clusters were well segregated (Fig. 3b, bottom panel), and there were significant improvements in visualization over other methods, including the method applied with no corrections. We further visualized spatial dependence due to variations in neighboring microenvironments using RGB plots of the inferred slide-specific embeddings (Supplementary Fig. S17e & f) and observed various patterns of microenvironments in the different sections.

As a key feature, PRECAST estimates aligned embeddings for biological effects between cell/domain types and joint labels for the combined sections. We performed DE analysis of the combined sections (see Methods). In total, we detected 367 DE genes with adjusted p-values of less than 0.001 in all seven spatial domains detected by PRECAST (Supplementary Table S4). A heatmap of the findings shows the good separation of the DE genes across different spatial domains (Fig. 3c). Many of these genes are markers that define particular cellular regions in liver lobes, i.e., *Cyp2e1, Cyp2c37*, *Oat,* and *Slc1a2* for central veins (Domains 1-2) [38, 39]; *Cyp2f2, Hal, Sds,* and *Ctsc* for portal veins (Domains 6-7) [38]; and *Gsn, Vim,* and *Col3a1* for the mesenchyme (Domain 5) [38]. Further enrichment analysis shows that genes specific to both central veins and portal veins were highly enriched for metabolic processes, with central veins that were more enriched for fatty acid metabolism, while portal veins that were more enriched for amino acid metabolism (Supplementary Fig. S19). By performing enrichment analysis for DE genes unique to each of two subtypes in central/portal veins, we found pathways unique to each subtype of central/portal veins (Supplementary Fig. S20).

To examine the cell compositions of each spatial domain detected by PRECAST, we performed cell-type deconvolution analysis of all mouse liver datasets using scRNA-seq data in the Mouse Cell Atlas (MCA) [40] as the reference panel. To assess the performance in spatial clustering, we evaluated the associations between the cell type proportions obtained by cell-type deconvolution and the cluster labels estimated by PRECAST and other methods. The results displayed in Fig. 3d suggest that PRECAST retained the largest MacFadden's adjusted R^2 than the other methods. In addition, cell-type deconvolution enabled us to spatially map 17 cell types annotated in the MCA dataset for liver tissue sections (Fig. 3e and Supplementary Fig. S21-S22). We observed a high proportion of estimated values for periportal and pericentral hepatocytes, which is consistent with the findings in the existing literature [38].

Lastly, we performed trajectory inference using aligned embeddings and estimated domain labels from PRECAST, to examine the cell lineages in the detected domains. Fig. 3f shows the inferred pseudotime mapped to PRECAST-induced tSNE, which suggests the central veins differentiated earlier than portal veins, in accordance with findings that showed Wnts and R-spondin3 signals were released from the central veins and transited along the venular wall towards the perivenous hepatocytes [39]. Based on the inferred pseudotime, we identified differentially expressed genes along the cell pseudotime using TSCAN [41]. The heatmap of the expression of the top 20 most significant genes (Fig. 3g) suggested the occurrence of some interesting dynamic expression patterns over pseudotime.

Application to mouse olfactory bulb Slide-seqV2 data

In Visium, barcoded beads 55 μm in diameter and with a center-to-center distance of 100 μm are used to capture mRNA with an average of 1-10 cells per spot [11]. The Slide-seq technique was recently developed for high-resolution SRT using 10-μm barcoded beads with an average of 1–3 cells per spot [9, 10], and Slide-seqV2 further improved the detection sensitivity. To show the scalability of PRECAST, we analyzed a mouse OB dataset generated using Slide-seqV2 technology [22]. In this dataset, spatial transcriptomic information was obtained from a total of 20 OB sections distributed evenly along the anterior–posterior axis. We removed four slides due to the quality of the sections from the end of the mouse OB region, and analyzed 16 containing 21,571 genes, on average, from over a total of 693,863 spots using PRECAST and other methods. After quality control (QC), we obtained data of a lower resolution ($\sim$4,000 spots) by collapsing nearby spots in each slide (see Methods), and the resulting resolution was similar to that of Visium (average of 10 cells per spot). To further examine the structure of the mouse OB, we relied on the structural annotations in the Allen Brain Atlas [42] (Fig. 4a). With near-single-cell resolution, the identified aligned embeddings and cluster labels both showed more fine-scale cell-type distribution patterns in the mouse OB [43]. In comparison, when we lowered the resolution, PRECAST estimated the aligned embeddings and cluster labels with smoother spatial patterns at the expense of less detailed local spatial information (Fig. 4b and Supplementary Fig. S23-S26). Moreover, the resulting RGB plots from PRECAST showed the laminar organization of the mouse OB, and PRECAST provided smoother transitions across neighboring spots and spatial domains than the other methods (Supplementary Fig. S23a & b to S26a & b). We next visualized the inferred aligned embeddings for biological effects between cell/domain types using two components from tSNE from each method (Fig. 4c and Supplementary Fig. S27) and showed that, using PRECAST, the spatial spots were mixed well across the 16 slides, while the cell/domain clusters were well segregated, suggesting PRECAST was more effective at spatial data integration. We further visualized the inferred slide-specific embeddings for spatial dependence due to variations in the microenvironment using RGB plots (Supplementary Fig. S28) and observed a few microenvironmental patterns in the inner layers (e.g., the granule cell layer, GCL), middle layers (e.g., the glomerular layer, GL), and outer layers (e.g., the olfactory nerve layer, ONL).

At reduced resolution, PRECAST detected 12 spatial domains with laminar organization, including the rostral migratory stream (RMS, Domain 1), GCL (Domain 2), GCL/inner plexiform layer (GCL/IPL, Domain 3), mitral cell layer (MCL, Domain 4), outer plexiform layer (OPL, Domain 5), GL (Domains 6 and 7), and ONL (Domains 8), with Domains 9-12 belonging to low-quality regions or experiment artifacts. To characterize the transcriptomic properties of the spatial domains identified by PRECAST, we performed DE analysis of the combined 16 tissue slides (see Methods). In total, we detected 4,131 DE genes with adjusted p-values of less than 0.001 in all 12 spatial domains detected by PRECAST (Supplementary Table S5), including representative genes that define particular cellular layers in the mouse OB, e.g., *Sox2ot* and *Sox11* (RMS) [44, 45]. The heatmap of the findings shows the good separation of the DE genes across different spatial domains (Supplementary Fig. S29), and we found that genes specific to Domain 1 (RMS) were enriched for myelin sheath and structural constituents of myelin sheath (Supplementary Fig. S30). Compared with the reduced-resolution data, when near-single-cell

level resolution was used, PRECAST identified 24 cell clusters, including fine-scale cell-type clusters. To better visualize each detected cell cluster, we plotted a heatmap of each cluster assignment for all 16 slides (Supplementary Fig. S31-S32). A heatmap of DE genes across different cell clusters showed Clusters 1-3 were subtypes of granule cells, and Clusters 4-6 belonged to cells in MCL and GL (Supplementary Fig. S33).

To examine the cell compositions in each spatial domain detected by PRECAST at the reduced resolution, we performed cell-type deconvolution analysis of all 16 tissue slides using scRNA-seq data from adult mouse OB as the reference panel [46]. As shown in Fig. 4d and Supplementary Fig. S34a, Domain 1 (RMS) was enriched for immature neurons. Immature neurons reportedly migrate to the OB through RMS [47]. Unsurprisingly, we found that Domains 2-3 (GCL) were dominated by two primary subtypes of granule cell, with a larger proportion of immature neurons in Domain 2 (inner) and an enrichment of mitral and tufted cells in Domain 4 (MCL). Domain 8 (ONL) was primarily enriched in olfactory sensory neurons (OSNs); OSNs express odorant receptors in the olfactory epithelium [48]. We additionally quantified the association between the inferred cell type proportions and the domain labels estimated by PRECAST and the other methods using MacFadden's adjusted R^2 (Fig. 4e, top panel), and PRECAST achieved the highest R^2. Then, we manually annotated each spot with the cell type present at the highest proportion [49] and quantitatively evaluated the clustering performance of PRECAST and other methods. As shown in Fig. 4e (bottom panel) and Supplementary Fig. S34b & c, PRECAST achieved the highest ARI and NMI values assessed separately for each slide or jointly for the combined slides. Further analyses showed that PRECAST achieved better data integration, with the highest iLISI and the lowest cLISI (Supplementary Fig. S34d), and maintained comparable conditional Pearson's correlations with the other methods (Supplementary Fig. S34e).

We performed trajectory inference to examine the cell lineages among the detected domains using aligned embeddings and estimated domain labels from PRECAST. In general, the estimated trajectory showed an "inside-out" sequence, consistent with the general understanding that OB neurons migrate from the subventricular zone along the RMS to the OB, before migrating radially out of the RMS in an inside-out sequence [50] (Fig. 4f and Supplementary Fig. S35a). Further DE analysis identified genes along the pseudotime (Supplementary Fig. S35b).

Application to hepatocellular carcinoma Visium data

To study the dynamics of tumorigenesis in tumors and tumor-adjacent tissues, we further analyzed four slides of in-house HCC data generated using the 10x Visium platform, with two slides from tumors (HCC1 and HCC2) and two from tumor-adjacent tissues (HCC3 and HCC4) from an HCC patient. The median number of spots was 2,748, and the median number of genes per spot was 3,635. Fig. 5a shows a histology image (top panel) with manual annotations for tumor/normal epithelium (TNE) and stroma provided by a pathologist (bottom panel). Consistent with the DLPFC data, the RGB plots generated by PRECAST clearly segregated the tissue slices into multiple spatial domains (Fig. 5b, top panel), with neighboring spots and spots in the same domain across multiple slides more closely sharing similar RGB colors than those generated by other methods (Supplementary Fig. S36a & b). Similarly, the PRECAST spatial heatmaps of clustering assignment for the four tissue slides resembled the

corresponding RGB plots and presented more spatial patterns across the tissue slides than those from the other methods (Fig. 5b, bottom panel; Supplementary Fig. S36c). We further visualized the inferred embeddings for biological effects between cell/domain types using two components from tSNE for each method (Fig. 5c and Supplementary Fig. S37a), in which the tSNE plots for PRECAST showed that spots from different slices were mixed well (Fig. 5c, top panel) while the domain clusters were well segregated (Fig. 5c, bottom panel), with significant improvements in visualization. A heatmap of Pearson's correlation coefficients among the detected domains shows the good separation of estimated embeddings across domains (Supplementary Fig. S37b), in which correlations between regions in TNE were high, and correlations between regions for TNE and stroma were low. We further visualized the spatial dependence due to microenvironment variations using RGB plots of the inferred slide-specific embeddings (Supplementary Fig. S38a) and observed variations in microenvironmental patterns between Domains 1 and 3, and between Domains 4 and 5.

To characterize the transcriptomic properties of the spatial domains identified by PRECAST, we performed DE analysis of the combined four tissue slides (see Methods). In total, we detected 2,093 DE genes with adjusted p-values of less than 0.001 in all nine spatial domains detected by PRECAST, with 539 genes being specific to Domains 1-5, which corresponded to TNE regions (Supplementary Table S6). A heatmap and ridge plots of the findings showed the good separation of the DE genes across different spatial domains (Supplementary Fig. S38b & S39-S40). In TNE regions (Domains 1-5), we further found that genes specific to Domains 1 and 5 were highly enriched in pathways of chemical carcinogenesis DNA adducts and chemical carcinogenesis receptor activation. Genes specific to Domain 4 were enriched in signaling pathways of RAF1 mutants and signaling by RAS mutants, and genes specific to Domains 2 and 3 were highly enriched in complement and coagulation cascade pathways (Supplementary Fig. S41). Genes specific to Domains 6-9 were enriched in the angiogenesis pathway, and we found 40 out of 427, 34 out of 606, 25 out of 281, and 18 out of 240 angiogenesis signature genes, respectively, in Domains 6-9, from multiple studies [51–54]. Interestingly, Domains 1-3 were only present in the tumor tissues (HCC1 and HCC2), Domain 5 was only present in the tumor-adjacent tissues (HCC3 and HCC4), while Domain 4 was shared across the tumor and tumor-adjacent tissues.

To identify SVGs other than those that were merely relevant to domain differences, we performed SVA analysis with the embeddings estimated by PRECAST as covariates for each slice. A detailed list of genes identified at an FDR of 1% is available in Supplementary Table S7. By performing functional enrichment analysis of these SVGs, we detected SVGs adjusted for domain-relevant covariates to be highly enriched in many common pathways in the four HCC slices, e.g., cytoplasmic translation, and cytosolic ribosome (Supplementary Fig. S42-S43).

Next, to examine the cell compositions of each spatial domain detected by PRECAST, we performed cell-type deconvolution analysis of all four HCC slides using scRNA-seq data as the reference panel (see Methods). The scRNA-seq reference panel consisted of malignant and tumor microenvironment cells, including cancer-associated fibroblasts (CAFs), tumor-associated macrophages (TAMs), tumor-associated endothelial cells (TECs), cells of an unknown entity but expressing hepatic progenitor cell markers (HPC-like), and immune cells [55]. As shown in Fig. 5d & e and Supplementary Fig. S44a, the proportions of malignant cells were substantially higher in Domains 1-5, while HPC-like cells were seen at higher proportions in Domain

7. In Domain 6, we observed an increased proportion of TAMs and immune cells and genes specific to this domain included *Tgfb1* and *Mmp2*, which have been used for the classification of TAMs [56, 57].

The estimated aligned embeddings and cluster labels can also be used in RNA velocity analysis to investigate the directed transcriptional dynamics of tumorigenesis when spliced and unspliced mRNA are available (see Methods). Interestingly, two cell lineages were identified, with one that originated in Domain 2 (TNE) and spread to Domains 4 and 5 (TNE), followed by Domains 6, 7, and 9 (stroma), and the other that originated in Domains 1 and 3 (TNE) and spread to Domain 8 (stroma). The TNE regions in Domain 2 residing in HCC1 may play a key role in tumorigenesis with Domain 4 (TNE) shared among the four slides and Domain 5 (TNE) in the tumor-adjacent tissues (Fig. 5f and Supplementary Fig. S44b). To infer cell states in the identified TNE, we further performed RNA velocity analysis using the spots identified in Domains 1-5. An expression heatmap of the top genes associated with cell states with induction close to 0 and repression close to 1 is shown in Fig. 5g (see Methods). The TNEs in Domains 1 and 2 tended to be transcriptionally active, while TNEs in Domains 4 and 5 tended to be repressed and show no transcription. The top genes associated with cell states included *Spink1*, *Rpl30*, and *Il32*, which highlights the importance of genes associated with cell states in HCC [58–60].

Discussion

PRECAST takes, as input, matrices of normalized expression levels and the physical location of each spot across multiple tissue slides. The output of PRECAST comprises all aligned embeddings for cellular biological effect, slide-specific embeddings that capture spatial dependence in neighboring cells/spots, and estimated cluster labels. In contrast to other existing methods of data integration, PRECAST is a unified and principled probabilistic model that simultaneously estimates embeddings for cellular biological effects, performs spatial clustering, and more importantly, aligns the estimated embeddings across multiple tissue sections. Thus, we recommend applying PRECAST first before a comprehensive data analysis pipeline is deployed. By applying PRECAST, the aligned embeddings and estimated cluster labels can be used for many types of downstream analyses, such as visualization, trajectory analysis, and SVA and DE analysis for combined tissue slices. In more detail, we developed a module to further remove batch effects across multiple tissue slides based on housekeeping genes, making expression data comparable for different cell/domain clusters. This module is also applicable to the examination of expressional differences caused by multiple conditions when such information on tissue slides is obtainable.

PRECAST simultaneously performs dimension reduction and spatial clustering while using simple projections to align embeddings, and uniquely estimating embeddings that capture spatial dependence of neighboring cells/spots due to varied microenvironments. Recently, Liu et al. [61] showed that, compared with methods that perform dimension reduction and spatial clustering sequentially [18–20], joint methods can estimate embeddings for cellular biological effects more efficiently while accounting for the uncertainty in obtaining low-dimensional features from sequential analysis. A similar strategy has been described in previous self-supervised learning literature [62, 63]. PRECAST also takes advantage of CAR to account for the local

microenvironments of neighboring spots, and an intrinsic CAR component has been used to promote spatial smoothness in the observed expressions of SRT data [64]. We showed that, by projecting non-cellular biological effects onto cellular biological space, different sample slides exhibit a constant shift in the centroid of each cell/domain type. With the assistance of joint modeling, we can use a subclass of CAR, intrinsic CAR, to simultaneously account for both smoothness in neighboring embeddings and shift across batches due to non-cellular biological effects such as complex batch effects.

With the advent of high-throughput technologies for SRT, data integration is particularly relevant for analyzing SRT datasets from multiple tissue slides. Analysis of a single section with current state-of-the-art techniques, e.g., 10x Visium and Slide-seqV2, only covers a tiny area of the region of interest, and it takes a few or dozens of slides to cover the whole tissue/organ. In this case, biological variations between cell/domain types are often confounded by factors related to data generation processes. Methods in data integration, which serve as the first step before downstream analyses, not only align the embeddings but also estimate the shared cell/domain clusters across samples [23–26]. Most existing methods of data integration were designed to analyze scRNA-seq data without considering additional spatial information in the SRT data.

We examined the SRT data generated by three major platforms, 10x Visium, ST, and Slide-seqV2, with different spatial resolutions. With 10x Visium, each spot captures the mRNA of an average of 1-10 cells, whereas Slide-seqV2 achieves a higher resolution with an average of 1-3 cells per spot. As common sense implies, with low-resolution datasets, PRECAST recovers aligned embeddings for cellular biological effects and cluster labels with smoother spatial patterns while losing the detailed local spatial information. Whereas, at near-single-cell resolution, the identified aligned embeddings and cluster labels show more fine-scale cell-type distribution patterns. Using four datasets, we demonstrated that PRECAST can successfully perform data integration with aligned embeddings across tissue sections, such that spots across different tissue sections are well mixed while cell/domain clusters are well segregated, improve clutering performance, and detect varied microenvironments in tissues. When applied to an HCC dataset, PRECAST identified five spatial domains belonging to TNE cells that were consistent with both manual annotations and spatial deconvolution. We further performed RNA velocity analysis to show the potential velocity of spots in different regions, shedding light on the tumorigenesis in the context of the tissues. To demonstrate the scalability of PRECAST, we analyzed a Slide-seqV2 dataset of 16 slides equally distributed along the anterior–posterior axis of a mouse OB.

PRECAST provides opportunities for new exciting research routes. Firstly, when SRT datasets of single-cell resolution are available, its use can be extended to the integration of multimodal single-cell data. For example, integrating single-cell ATAC-seq will allow users to examine cell-type-specific regulatory mechanisms in the spatial context of tissues. Secondly, it would be interesting to integrate single-cell-resolution SRT datasets with CITE-seq data, thus integrating spatial transcriptomics with immunophenotyping, in which surface proteins are detected by antibody-derived tags. This would enable the exploration of surface proteins not measured in SRT datasets.

Materials and Methods

Methods

PRECAST overview

Here, we present a basic overview of PRECAST, and further details are available in the Supplementary Notes. PRECAST is a data integration method for SRT data from multiple tissue slides. The proposed method involves simultaneous dimension reduction and spatial clustering built on a hierarchical model with two layers, as shown in Fig. 1a. The first layer, the dimension-reduction step, relates gene expression to the shared latent embeddings, while the second layer, the spatial-clustering step, relates the shared latent embeddings and spatial coordinates to the cluster labels. In the dimension-reduction step, an intrinsic CAR model captures the spatial dependence induced by neighboring microenvironments in the low-dimensional embedding space, while in the spatial-clustering step, a Potts model promotes spatial smoothness in the cluster label space. Using simple projections of the batch effects and/or biological effects of slides onto the space of biological effects between cell/domain types, PRECAST aligns cell/domain clusters across multiple tissue slides with the shared distributions of embeddings of each cell/domain and detects their cell/domain labels. With M SRT datasets, we observe an $n_r \times p$ normalized expression matrix $X_r = (\mathbf{x}_{r1}^{\mathrm{T}}, \cdots, \mathbf{x}_{ri}^{\mathrm{T}}, \cdots, \mathbf{x}_{rn_r}^{\mathrm{T}})^{\mathrm{T}}$ for each sample r $(= 1, \cdots, M)$, where $\mathbf{x}_{ri} = (x_{ri1}, \cdots, x_{rip})^{\mathrm{T}}$ is a p-dimensional normalized expression vector for each spot $s_{ri} \in \mathbb{R}^2$ of sample r on square or hexagonal lattices, among others; while the cluster label of spot s_{ri}, $y_{ri} \in \{1, \cdots, K\}$, and q-dimensional shared embeddings, $\mathbf{z}_{ri}$'s, are unavailable. Without loss of generality, we assume that, for each sample r, $\mathbf{x}_{ri}$ is centered, and PRECAST models the centered normalized expression vector $\mathbf{x}_{ri}$ with its latent low-dimensional feature, $\mathbf{z}_{ri}$, and class label, y_{ri}, as

$$\mathbf{x}_{ri} = W(\mathbf{z}_{ri} + \mathbf{v}_{ri}) + \boldsymbol{\varepsilon}_{ri}, \boldsymbol{\varepsilon}_{ri} \sim N(0, \Lambda_r), \tag{1}$$

$$\mathbf{z}_{ri} | y_{ri} = k \sim N(\mu_k, \Sigma_k), \tag{2}$$

where $\Lambda_r = \mathrm{diag}(\lambda_{r1}, \cdots, \lambda_{rp})$ is a diagonal matrix for residual variance, $W \in \mathbb{R}^{p \times q}$ is a loading matrix that transforms the p-dimensional expression vector into q-dimensional embeddings shared across M datasets, $\mu_k \in \mathbb{R}^{q \times 1}$ and $\Sigma_k \in \mathbb{R}^{q \times q}$ are the mean vector and covariance matrix for the kth cluster, respectively, and $\mathbf{v}_{ri}$ is a q-dimensional slide-specific latent vector that captures the spatial dependence among neighboring spots and aligns embeddings across datasets. Eqn. (1) is related to the high-dimensional expression vector ($\mathbf{x}_{ri}$) in p genes with a low-dimensional feature ($\mathbf{z}_{ri}$) via a probabilistic PCA model [65] with consideration of spatial dependence while Eqn. (2) is a Gaussian mixture model (GMM) [66] for this latent feature among all spots across M datasets. To promote spatial smoothness in the space of cluster labels, we assume each latent class label, y_{ri}, is interconnected with the class labels of its neighborhoods via a discrete hidden Markov random field (HMRF). In detail, we use the following Potts model for the latent labels,

$$P(\mathbf{y}_r) = C_r(\beta_r)^{-1} \exp\{-\frac{1}{2} \sum_i \sum_{i' \in N_{ri}} \beta_r (1 - \delta(y_{ri}, y_{ri'}))\}, \tag{3}$$

where $C_r(\beta_r)$ is a normalization constant that does not have a closed form, N_{ri} is the neighborhood of spot s_{ri} in sample r, and β_r is the sample-specific smoothing parameter that captures the label similarity among the neighboring spots. However, we assume a continuous multivariate HMRF for the vector, $\mathbf{v}_{ri}$, which captures spatial dependence in the embedding space. In detail, we assume an intrinsic CAR model [67] for $\mathbf{v}_{ri}$

$$\mathbf{v}_{ri}|\mathbf{v}_{[n_r]\setminus i} \sim N(\mu_{v_{ri}}, m_{ri}^{-1}\Psi_r), \tag{4}$$

where subscript $_{[n_r]\setminus i}$ denotes all spots but s_{ri} in sample r, m_{ri} is the number of neighbors of spot i in sample r, $\mu_{v_{ri}} = m_{ri}^{-1}\sum_{i'\in N_{ri}}\mathbf{v}_{ri'}$ is the conditional mean relevant to the neighbors of that spot s_{ri}, and Ψ_r is a $q\times q$ conditional covariance matrix for the elements of $\mathbf{v}_{ri}$. Conventionally, the joint distribution of the intrinsic CAR model is non-identifiable, as the mean of the joint distribution of intrinsic CAR is not zero. As shown in the next section, non-cellular biological effects, e.g., batch effects, from each slide can be projected onto the cellular biological space, which can be corrected by the non-zero-mean property of intrinsic CAR.

Projections of non-cellular biological effects

For each tissue slide, we assume the normalized expressions in each sample r can be decomposed into additional parts with respect to non-cellular biological effects as follows:

$$\mathbf{x}_{ri} = W(\mathbf{z}_{ri} + \boldsymbol{\nu}_{ri}) + W_r\boldsymbol{\zeta}_{ri} + \boldsymbol{\varepsilon}_{ri}, \tag{5}$$

where $\boldsymbol{\nu}_{ri}$ is a q-dimensional vector that captures the spatial dependence among neighboring spots; $W_r \in \mathbb{R}^{p\times\tilde{q}}$ is a loading matrix for a factor related to non-cellular biological effects; and $\boldsymbol{\zeta}_{ri}$, independent of $(\mathbf{z}_{ri}, \boldsymbol{\nu}_{ri})$, is the corresponding $\tilde{q}$-dimensional vector. Assuming cell biological space (W) and non-cellular biological space (W_r) are non-orthogonal, we can project W_rs onto the column spaces of W, i.e., $\widehat{W_r} = W(W^TW)^{-1}W^{\mathrm{T}}W_r$, and then rewrite the normalized expressions as the following

$$\mathbf{x}_{ri} \approx W(\mathbf{z}_{ri} + \boldsymbol{\nu}_{ri} + \widetilde{W}^{\mathrm{T}}W_r\boldsymbol{\zeta}_{ri}) + \boldsymbol{\varepsilon}_{ri},$$

where $\widetilde{W} = W(W^TW)^{-1}$. We denote $\mathbf{v}_{ri} = \boldsymbol{\nu}_{ri} + \widetilde{W}^{\mathrm{T}}W_r\boldsymbol{\zeta}_{ri}$, $\mu'_{v_{ri}} = E(\boldsymbol{\nu}_{ri}|\boldsymbol{\nu}_{[n_r]/i})$, $m_{ri}^{-1}\Psi'_r = \mathrm{var}(\boldsymbol{\nu}_{ri}|\boldsymbol{\nu}_{[n_r]/i})$, then the conditional mean and variance in the intrinsic CAR component can be written as

$$\begin{aligned}
\mathrm{E}(\mathbf{v}_{ri}|\mathbf{v}_{[n_r]/i}) &= \mu'_{v_{ri}} + \widetilde{W}^{\mathrm{T}}W_r\mathrm{E}(\boldsymbol{\zeta}_{ri}) \equiv \mu_{v_{ri}}, \\
\mathrm{var}(\mathbf{v}_{ri}|\mathbf{v}_{[n_r]/i}) &= m_{ri}^{-1}\Psi'_r + \widetilde{W}^{\mathrm{T}}W_r\mathrm{var}(\boldsymbol{\zeta}_{ri})W_r^{\mathrm{T}}\widetilde{W} \equiv m_{ri}^{-1}\Psi_r,
\end{aligned}$$

where we assume $\mathrm{var}(\boldsymbol{\zeta}_{ri}) = m_{ri}^{-1}\Psi''_r$. Then, the projected approximated model can be written as

$$\mathbf{x}_{ri} \approx W(\mathbf{z}_{ri} + \mathbf{v}_{ri}) + \boldsymbol{\varepsilon}_{ri}, r = 1, 2, \tag{6}$$

where $\mathbf{z}_{ri}|y_{ri} = k \sim N(\mu_k, \Sigma_k)$, and $\mathbf{v}_{ri}|\mathbf{v}_{[n_r]/i} \sim N(\mu_{v_{ri}}, m_{ri}^{-1}\Psi_r)$.

Recovery of comparable gene expression matrices

Once the estimated cluster labels are obtained by PRECAST, we can remove unwanted variations using a set of housekeeping genes as negative control genes that are not affected by other biological effects [68]. In this study, we used a set of mouse/human housekeeping genes from the Housekeeping and Reference Transcript Atlas [69]. First, we obtained vectors, $\tilde{\mathbf{x}}_{ri} = (\tilde{\mathbf{x}}_{ri1}, \cdots, \tilde{\mathbf{x}}_{hrL})$, of the expression of L housekeeping genes by matching the set of housekeeping genes and genes passing QC for each dataset. By performing PCA, we obtained the top 10 PCs, $\widehat{\mathbf{m}}_{ri}$, as covariates to adjust for unwanted variation. One of the outputs of PRECAST, the posterior probability of y_{ri} ($\widehat{\mathbf{r}}_{ri} \in \mathbb{R}^K$), can be used as the design matrix to explain biological variation between cell/domain types. Finally, we used a linear model for the normalized gene expression vector

$$\mathbf{x}_{ri} = \widehat{\mathbf{r}}_{ri}^{\mathrm{T}}\boldsymbol{\alpha} + \widehat{\mathbf{m}}_{ri}^{\mathrm{T}}\boldsymbol{\gamma} + \varepsilon_{ri}, \tag{7}$$

where $\boldsymbol{\alpha}$ is a K-dimensional vector for biological effects between cell/domain types and $\boldsymbol{\gamma}$ is a 10-dimensional vector of regression coefficients associated with the unwanted factors. After obtaining the parameter estimates in Eqn. (7), users can remove batch effects from the original normalized gene expression using

$$\widehat{\mathbf{x}}_{ri} = \mathbf{x}_{ri} - \widehat{\mathbf{m}}_{ri}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}.$$

This strategy can also be applied to samples from multiple biological conditions when such information is available. This can be achieved by adding additional covariates for biological conditions in Eqn. (7).

Differential expression analysis and enrichment analysis

After removing unwanted variation of gene expression matrices for multiple slides, we performed DE analyses and enrichment analysis on all four datasets. In detail, we used the `FindAllMarkers` function in the R package *Seurat* with default settings to detect the differentially expressed genes for each domain detected by PRECAST. The DE analysis was considered to identify domain-specific DE genes with adjusted p-values of less than 0.001 and a log-fold change of greater than 0.25. After obtaining a set of genes specific to each domain, we performed gene set enrichment analysis (GSEA) on a set of detected DE genes using g:Profiler with the g:SCS multiple testing correction method and applying a significance threshold of 0.05 [70].

Conditional SVG analysis

After obtaining aligned embeddings and domain labels for DLPFC and HCC Visium data using PRECAST, we detected the SVGs by adjusting the estimated aligned embeddings as covariates to investigate the role of SVGs beyond differences between cell/domain types. In detail, we used the function `spark.vc` in the R package *SPARK* to identify SVGs adjusted for cell/domain-relevant covariates for each sample. Finally, an FDR of 1% was adopted to identify the significant SVGs.

Trajectory inference/RNA velocity analysis

To further investigate the development and differentiation of these identified spatial domains/cells by PRECAST, we used the aligned embeddings and domain clusters estimated by PRECAST to perform trajectory inference, or RNA velocity analysis if splicing and unsplicing information was available. For the DLPFC Visium data, mouse liver ST data and mouse OB Slide-seqV2 data, we conducted trajectory inference for the combined spots in all multiple tissue sections using Slingshot [71]. We inputted the aligned embeddings and domain clusters estimated by PRECAST into the function `slingshot` in the R package *slingshot* for implementation. Because the splicing and unsplicing information was available for the HCC Visium dataset, we ran RNA velocity analysis using the `scvelo.tl.velocity` function in the Python module *scvelo* based on the splicing and unsplicing matrices, then the aligned embeddings and domain clusters estimated by PRECAST were used to visualize the inferred RNA velocity and latent time.

Cell-type deconvolution analysis

We performed deconvolution analysis using Robust Cell Type Decomposition (RCTD)[49], a supervised learning method used to decompose each spatial transcriptomics pixel into a mixture of individual cell types while accounting for platform effects. We leveraged the results of deconvolution analysis for better biological interpretation of the real data analysis.

For mouse liver ST data, we used scRNA-seq data on liver tissue sections from an adult mouse from the MCA [40]. By removing cell types with cell numbers of less than 18, the reference data retained 4,640 cells belonging to 17 cell types: B cells with high *Jchain* expression ($n = 43$), erythroid cells with high *Hbb-bs* expression ($n = 555$), periportal hepatocytes ($n = 26$), hepatocytes with high *Fabp1* expression ($n = 149$), Kupffer cells ($n = 1,046$), erythroid cells with high *Hbb-bt* expression ($n = 62$), endothelial cells ($n = 1,196$), granulocytes ($n = 194$), NK cells ($n = 114$), macrophages ($n = 191$), dendritic cells ($n = 422$), B cells with high *Fcmr* expression ($n = 97$), T cells ($n = 219$), plasmacytoid dendritic cells ($n = 90$), hepatocytes with high *Spp1* expression ($n = 99$), pericentral hepatocytes ($n = 119$), and hepatic stellate cells ($n = 18$).

For the mouse OB Slide-seqV2 data, we used 10X Genomics Chromium scRNA-seq data collected from the wild-type OB [46]. The reference data included 17,453 cells belonging to 40 cell types: three astrocyte types (Astro1: $n = 1087$; Astro2: $n = 129$; and Astro3: $n = 599$), two endothelial types (EC1: $n = 897$ and EC2: $n = 445$), two mesenchymal types (Mes1: $n = 200$ and Mes2: $n = 65$), three microglials (MicroG1: $n = 255$; MicroG2: $n = 766$; and MicroG3: $n = 294$), monocytes ($n = 238$), two murals (Mural1: $n = 43$ and Mural2: $n = 272$), myelinating-oligodendrocyte cells ($n = 294$), macrophages ($n = 197$), five olfactory ensheathing cell types (OEC1: $n = 812$; OEC2: $n = 664$; OEC3:$n = 1073$; OEC4: $n = 1083$; and OEC5: $n = 281$), oligodendrocyte progenitor cells (OPC: $n = 137$), red blood cells (RBCs: $n = 103$), and 18 neural subtypes (OSNs: $n = 467$; PGC-1: $n = 437$; PGC-2: $n = 146$; PGC-3 $n = 106$; GC-1: $n = 540$; GC-2: $n = 1245$; GC-3: $n = 96$; GC-4: $n = 80$; GC-5: $n = 516$; GC-6: $n = 89$; GC-7: 293; Immature: $n = 1150$; Transition: $n = 733$; Astrocyte-Like: $n = 1078$; M/TC-1: $n = 24$; M/TC-2: $n = 44$; M/TC-3: $n = 411$; and EPL-IN: $n = 64$). For more details of the cell types, please refer to Tepe et al. [46].

For the HCC Visium data, we leveraged a droplet-based scRNA dataset collected from HCC and intra-hepatic cholangiocarcinoma patients to serve as the reference data for deconvolution [55]. After filtering out 92 cells of unclassified cell type, the reference data contained 5,023 cells belonging to seven cell types: B cells ($n = 598$), CAFs ($n = 724$), HPC-like cells ($n = 254$), malignant cells (702), T cells ($n = 1,429$), TAMs ($n = 437$), and TECs ($n = 879$). After acquiring the deconvolution results from the above-mentioned reference, we combined the proportions of B cells, T cells and TECs into a single set and referred to this as the immune cell proportion.

Comparisons of methods

We conducted comprehensive simulation and real data analyses to compare PRECAST with existing methods of data integration and clustering.

We applied the following single-cell integration methods to benchmark the data integration performance of PRECAST: (1) Seurat V3 [26]; (2) Harmony [23] implemented in the R package *harmony*; (3) fastMNN [24] implemented in the R package *batchelor*; (4) Scanorama [25] implemented in the Python module *scanorama*. (5) scGen [28]; (6) scVI [27] implemented in the Python module *scvi*; (7) MEFISTO [29] implemented in the Python module *mofax*; and (8) PASTE [30] implemented in the Python module *paste*.

The first six methods were designed for scRNA-seq integration, while MEFISTO can be used for scRNA-seq and SRT data integration, and PASTE is designed for integrating SRT data from multiple adjacent tissue slices into a single slice. In detail, the Seurat V3 method is based on finding the mutual nearest neighbors. First, anchor pairs between different batches were obtained using the function `FindIntegrationAnchors` in the *Seurat* R package. Then, the `IntegrateData` function in *Seurat* was used to integrate the datasets based on the anchor pairs and their confidence scores. Finally, we obtained the aligned 15-dimensional embeddings using the function `RunPCA` in *Seurat*. Compared with the original MNN method, fastMNN is a faster version based on PCA and is used to find the mutual nearest neighbors, greatly increasing the computational efficiency. We used `fastMNN` in the R package *batchelor* with the top 15 PCs and other default parameters and obtained all aligned 15-dimensional embeddings. Another PCA-based method, Harmony corrects batch-contaminated PCs by iteratively performing maximum diversity clustering and linear mixture model correction. For the analysis, we first obtained the 15-dimensional PCs by performing PCA on the combined normalized expression matrix, then took the obtained PCs and batch information as input for the function `HarmonyMatrix` in the R package *harmony*, and lastly, obtained the aligned 15-dimensional embeddings as output. Scanorama is a generalization of mutual nearest-neighbor matching used to find similar elements among many datasets. Instead of searching for similar elements in high-dimensional gene space, Scanorama applies randomized singular value decomposition to extract the aligned embeddings from the combined cell-by-gene expression matrix. We conducted the analysis following the author's pipeline `https://scanpy-tutorials.readthedocs.io/en/latest/spatial/integration-scanorama.html` and set the parameter *dimred=15* in function `correct_scanpy` to extract aligned 15-dimensional embeddings. The pipelines for scGen and scVI, which are two deep-learning-based methods, were followed using the two respective links, `https://scgen.readthedocs.io/en/latest/tutorials/scgen_batch_removal.html` and

https://docs.scvi-tools.org/en/stable/tutorials/notebooks/tabula_muris.html, and we set the number of latent embeddings to 15 for both methods in the analysis. MEFISTO is a factor-model-based integration method that integrates multiple datasets by simultaneously identifying and aligning the underlying patterns of variation. To implement this, we followed the author's pipeline provided at https://github.com/bioFAM/MEFISTO_tutorials/blob/master/MEFISTO_microbiome.ipynb and https://github.com/bioFAM/MEFISTO_tutorials/blob/master/MEFISTO_ST.ipynb, and set the number of factors to 15. PASTE is a pairwise alignment method that integrates adjacent slices based on transcriptional and spatial similarity using optimal transport and nonnegative matrix factorization; we followed the author's pipeline https://github.com/raphael-group/paste/blob/main/docs/source/notebooks/getting-started.ipynb for implementation and set the number of low-dimensional representation of center slice to 15.

To evaluate clustering performance, we considered the following two clustering methods that use extracted aligned embeddings as input. (1) SC-MEB implemented in the R package *SC.MEB* [20] and (2) Louvain implemented in the R package *igraph* [72]. SC-MEB was recently developed to perform spatial clustering based on a discrete Markov random field [20], while Louvain is a conventional non-spatial clustering algorithm based on community detection in large networks [72]. In the implementation, we set the default values for SC-MEB and Louvain clustering in the respective packages, except that the aligned embeddings and spatial coordinates were used as input for SC-MEB while only the aligned embeddings were the input for Louvain. Similar to PRECAST, SC-MEB uses MBIC [73] to determine the number of clusters in a data-driven manner, while Louvain uses a community-modularity maximizing rule [72].

Evaluation metrics

We evaluated the methods' performances in data integration, obtaining embeddings for cellular biological effects, and spatial clustering using the following metrics.

Local inverse Simpson's index. To assess performance in batch-effect removal, we used the cell-type/integration LISI, cLISI and iLISI [23] to quantify the performance in merging the shared cell populations among tissue slides and mixing spots from M tissue slides. cLISI assigns a diversity score to each spot that represents the effective number of cell types in the neighborhoods of that spot. For M datasets with a total of K cell types, accurate integration should maintain a cLISI value close to 1, reflecting the purity of the unique cell types in the neighborhood of each spot, as defined by the aligned embeddings. Erroneous embedding includes neighborhoods with a cLISI of more than 1, while the worst cases have cLISI close to K, suggesting that neighbors have K different types of cells. iLISI has a similar form and implication as cLISI but is based on a sample index set rather than clusters of cell types. Thus, a larger iLISI value means the different samples have more sufficient mixing.

Silhouette coefficient. To simultaneously evaluate the separation of each cell/domain cluster and mixing of multiple datasets, we calculated the average silhouette coefficient of the SRT datasets using two different groupings: (1) grouping using known cell types as the cell/domain-type silhouette coefficient ($\text{silh}_{cluster}$) and (2) grouping using different datasets as the batch silhouette coefficient (silh_{batch}). In the data integration, a larger value of $\text{silh}_{cluster}$ indicates better preservation of the biological signals between cell/domain types, while a smaller value

of silh_{batch} suggests better mixing of datasets. These two metrics can be summarized using the F1 score as follows [74]:

$$\text{F1_score_silh} = \frac{2(1-\text{silh}'_{\text{batch}})\text{silh}'_{\text{cluster}}}{\text{silh}'_{\text{cluster}} + (1-\text{silh}'_{\text{batch}})} \in [0,1],$$

where $\text{silh}'_{\text{batch}} = \frac{1+\text{silh}_{\text{batch}}}{2}$ and $\text{silh}'_{\text{cluster}} = \frac{1+\text{silh}_{\text{cluster}}}{2}$. A larger F1 score suggests better data integration that preserves the biological variations between cell/domain types while removing other non-cellular biological variations across multiple tissues.

Canonical correlation coefficients and/or conditional correlation. For dimension reduction, we applied two measurements to assess the performance of true latent feature recovery in the simulation studies. The first was the mean canonical correlation between the estimated features and the true one, defined as

$$\text{CCor} = \frac{1}{q}\sum_{l=1}^{q} \zeta_l(\mathbf{z}_i, \hat{\mathbf{z}}_i),$$

where ζ_l is the l-th canonical correlation coefficient. The second measure was the mean conditional correlation between gene expression $\mathbf{x}_i$ and cell type label y_i given the estimated latent features $\hat{\mathbf{z}}_i$, defined as

$$\text{ConCor} = \frac{1}{p}\sum_{j=1}^{p} corr(y_i, resid_{ij}),$$

where $resid_i$ is the residual of x_{ij} regressing on $\hat{\mathbf{z}}_i$, and $corr(y_i, resid_{ij})$ is the Pearson correlation coefficient between y_i and $resid_{ij}$.

Adjusted Rand index and/or normalized mutual information. To evaluate performance in spatial clustering, we used ARI [75] and NMI [76]. ARI is the corrected version of the Rand index (RI) [77] and is used avoid some of the drawbacks of RI [75]. ARI measures the similarity between two different partitions and ranges from -1 to 1. A larger value of ARI means a higher degree of similarity between two partitions. ARI takes a value of 1 when the two partitions are equal up to a permutation. Whereas NMI is a variant of the mutual information (MI) that normalizes the value of MI to within the range of 0 and 1. When the two partitions are equal up to a permutation, NMI takes a value of 1.

Simulations

Scenario 1. Normalized gene expression data: domain labels and spatial coordinates from the Potts model. For this scenario, we generated the normalized gene expression data for three samples as well as the spatial coordinates based on Potts models with four neighborhoods. All quantities, such as the labels y_{ri}s and latent features $\boldsymbol{\nu}_{ri}$s, of spots were randomly simulated from the generative model (5). In detail, we generated the class label y_{ri} for each $r = 1, 2, 3$ and $i = 1, \cdots, n_r$ with $n_r \in \{65 \times 65, 60 \times 60, 60 \times 60\}$ corresponding to rectangular lattices from a K-state (K=7) Potts model with the smoothing parameter $\beta_r = 0.8 + 0.2(r-1)$ using the function `sampler.mrf` in R package *GiRaF*. Then,

we generated latent features $\boldsymbol{\nu}_{ri}$s from the CAR model with the same N_{ri} as defined in the Potts model and covariance matrix $\Psi'_r = (\sigma_{r,ij})$ with $\sigma_{r,ij} = r(0.2r)^{|i-j|}$ using the function `rmatrixnorm` in R package *LaplacesDemon*. We set different values for the smoothing parameter β_r and covariance matrices Ψ'_r across r to mimic the heterogeneity of three samples. The domain labels y_{ri}s and latent features $\boldsymbol{\nu}_{ri}$s were fixed once they were obtained.

Then, we generated domain-relevant latent features $\mathbf{z}_{ri}$ in the model (5) from the conditional Gaussian, such that $\mathbf{z}_{ri}|y_{ri} = k \sim (\mu_k, \Sigma_k)$, where $\mathbf{z}_{ri} \in R^q$ with $q = 10$ and structures for μ_k and Σ_k are shown in Supplementary Table S8. Next, we generated $\widetilde{W} = (\widetilde{w}_{ij}, i \leq p, j \leq q)$ with each $\widetilde{w}_{ij} \overset{i.i.d.}{\sim} N(0,1)$, performed QR decomposition on $\widetilde{W}$ such that $\widetilde{W} = \widetilde{Q}\widetilde{R}$, and assigned $W = \widetilde{Q}$, which is a column orthogonal matrix. Then, we generated a batch-loading matrix W_r by generating $\bar{W}_r = \bar{W} + \bar{E}_r$ with $\bar{W} = (\bar{w}_{ij}, i \leq p = 2,000, j \leq q_r), q_r = 2$ and $\bar{w}_{ij} \overset{i.i.d.}{\sim} N(0,1)$ and $\bar{E}_r = (\bar{e}_{rij}, i \leq p, j \leq q_r)$ and $\bar{e}_{rij} \sim N(0, \bar{\sigma}_r^2)$ with $\bar{\sigma}_1 = 0.5, \bar{\sigma}_2 = 0.8, \bar{\sigma}_3 = 1$. Similar to the generation of W, we performed orthogonalization of $\bar{W}_r$ to generate W_r. Next, we generated $\boldsymbol{\zeta}_{ri}$ in the model (5) by $\boldsymbol{\zeta}_{ri} = (\zeta_{ri1}, \cdots, \zeta_{riq_r})^{\mathrm{T}}$ such $\zeta_{rik} \overset{i.i.d.}{\sim} N(0, \sigma_r^2)$ with $\sigma_1 = 1, \sigma_2 = 2$ and $\sigma_3 = 0.5$. Finally, we generated a high-dimensional normalized expression matrix using $\mathbf{x}_{ri} = W(\mathbf{z}_{ri} + \boldsymbol{\nu}_{ri}) + W_r\boldsymbol{\zeta}_{ri} + \boldsymbol{\varepsilon}_{ri}, \boldsymbol{\varepsilon}_{ri} \sim N(\mathbf{0}, \Lambda_r)$, where $\Lambda_r = \mathrm{diag}(\lambda_{rj}), j = 1, \ldots, p$, $\lambda_{1j} = 2(1 + 1.5|z_{1j}|)$ with $z_{1j} \overset{i.i.d.}{\sim} N(0,3)$; $\lambda_{2j} = 2(1 + z_{2j})$ with $z_{2j} \overset{i.i.d.}{\sim} U[0,1]$; and $\lambda_{3j} = 2(1 + 2z_{3j})$ with $z_{3j} \overset{i.i.d.}{\sim} U[0,1]$.

Scenario 2. Raw gene expression count data: domain labels and spatial coordinates from Potts model. In Scenario 2, we generated the raw gene expression data. The class label y_{ri}, loading matrices W, W_r, latent features $\boldsymbol{\nu}_{ri}$ and $\mathbf{z}_{ri}$, and error term $\boldsymbol{\varepsilon}_{ri}$ were generated in the same way as in Scenario 1, except that μ_k had a different value (Supplementary Table S8). The other difference was that we generated normalized gene expressions $\mathbf{x}_{ri}$ using $\mathbf{x}_{ri} = \tau_r + W(\mathbf{z}_{ri} + \boldsymbol{\nu}_{ri}) + W_r\boldsymbol{\zeta}_{ri} + \boldsymbol{\varepsilon}_{ri}, \tau_{rj} \sim N(0,4)$ and raw gene expressions $\tilde{\mathbf{x}}_{ri} = (\tilde{x}_{ri1}, \cdots, \tilde{x}_{rip})^{\mathrm{T}}$ using $\tilde{x}_{rij} \sim Poisson(exp(x_{rij}))$, where τ_{rj} is the j-th element of τ_r. The term $\boldsymbol{\varepsilon}_{ri}$ makes the distribution of $\tilde{\mathbf{x}}_{rij}$ over-dispersed, which can better imitate the properties of count expression. In this scenario, for each sample r, we only observed the raw expression $\tilde{x}_{rij}$ of gene j and spot i and spatial coordinates s_{ri} for spot i.

Scenario 3. Normalized gene expression data: domain labels and spatial coordinates from DLPFC data. To validate the generality of PRECAST, we also generated data based on three DLFPC datasets (ID: 151507, 151669 and 151673) from each of three donors (Visium platform). The domain labels y_{ri}s for spots were obtained from the expert annotations from Maynard et al. [21] together with the spatial coordinates. To generate the latent features $\boldsymbol{\nu}_{ri}$s, PRECAST was used to fit the three datasets, then we used the estimated features $\hat{\boldsymbol{\nu}}_{ri}$s as input. The cluster labels y_{ri}s and latent features $\boldsymbol{\nu}_{ri}$s were fixed once they were obtained. The other quantities were generated in the same way as in Scenario 1 but with a different μ_k (Supplementary Table S8).

Scenario 4. Raw gene expression data: domain labels and spatial coordinates from DLPFC data. In this scenario, we also generated the raw gene expression data. The class label y_{ri}, loading matrices W, W_r, latent features $\boldsymbol{\nu}_{ri}$ and $\mathbf{z}_{ri}$ were generated as in Scenario 3. The only difference was that we generated normalized gene expression $\mathbf{x}_{ri}$ using $\mathbf{x}_{ri} = \tau_r + W(\mathbf{z}_{ri} + \boldsymbol{\nu}_{ri}) + W_r\boldsymbol{\zeta}_{ri} + \boldsymbol{\varepsilon}_{ri}, \tau_{rj} \sim N(0,4)$ and $\boldsymbol{\varepsilon}_{ri} \sim N(\mathbf{0}, \Lambda_r)$, where $\Lambda_r = \mathrm{diag}(\lambda_{rj}), j = 1, \ldots, p$, $\lambda_{1j} = 2(1 + 0.5|z_{1j}|)$ with $z_{1j} \overset{i.i.d.}{\sim} N(0,3)$; $\lambda_{2j} = 2(1 + 0.2z_{2j})$ with

$z_{2j} \overset{i.i.d.}{\sim} U[0,1]$; and $\lambda_{3j} = 2(1 + z_{3j})$ with $z_{3j} \overset{i.i.d.}{\sim} U[0,1]$. Then the raw expressions $\tilde{\mathbf{x}}_{ri}$ were simulated using $\tilde{x}_{rij} \sim Poisson(exp(x_{rij}))$, where τ_{rj} is the j-th element of τ_r. The term $\boldsymbol{\varepsilon}_{ri}$ makes the distribution of $\tilde{\mathbf{x}}_{rij}$ over-dispersed, which can better imitate the properties of count expression.

Scenario 5. Raw gene expression data: count matrix, domain labels and spatial coordinates from DLPFC data. To make PASTE comparable to the other methods, we obtained the raw gene expressions, domain labels and spatial coordinates from all three tissue slides (ID: 151673, 151674 and 151675) from the last donor, where the domain labels were from the expert annotations in Maynard et al. [21]. In this case, the true shared embeddings were unknown, so we did not evaluate the canonical correlations. To generate the count matrix, we randomly added a pseudocount for each raw count from a binomial distribution with size three and probability parameter 0.3, similar to Zeira et al. [30]. Finally, we obtained the count matrix and spatial coordinates as input for the compared methods.

Data resources

Human dorsolateral prefrontal cortex Visium data

We downloaded spatial transcriptomic data for human DLPFC obtained on the 10x Visium platform from `https://github.com/LieberInstitute/spatialLIBD`. These data were collected from 12 human postmortem DLPFC tissue sections from three independent neurotypical adult donors, and the raw expression count matrix contained 33,538 genes for each sample, with a total of 47,681 spatial locations. Before conducting the analysis, we first performed QC on each sample to filter out genes with non-zero expression levels for fewer than 20 spots and the spots with nonzero expression levels for less than 20 genes. The filtering step led to sets of 14,535 genes on average in a total of 47,680 spatial locations. For each sample, we performed log-normalization, followed by centering, and selected the top 2,000 SVGs using SPARK [78]. Next, we prioritized genes based on the number of times they were selected as SVGs in all 12 samples and chose the top 2,000 genes as input for the analysis. The annotated spatial domains in all 12 samples based on the cytoarchitecture in the original study [21] were layer 1 (n=5,321), layer 2 (n=2,858), layer 3 (n= 17,587), layer 4 (n=3,547), layer 5 (n=7,300), layer 6 (n=6,201), white matter (n=4,514), and undetermined spots (n=352). In the analysis, we treated these manual annotations as the ground truth to evaluate the clustering and data integration performance of the different methods.

Mouse liver ST data

We downloaded eight sets of mouse liver ST data from `https://zenodo.org/record/4399655`. The eight datasets contained 15,302 genes, on average, measured over 4,865 spatial spots in total. In the QC steps, we first filtered out genes with non-zero expression levels in fewer than

20 spots and spots with non-zero expression levels for fewer than 20 genes. The filtering step led to a set of 9,221 genes, on average, from a total of 4,865 locations. For each tissue slide, we performed log-normalization, followed by centering, and selected the top 2,000 SVGs using SPARK. Next, we prioritized genes based on the number of times they were selected as SVGs in all eight datasets, and chose the top 2,000 genes as input for the analysis. The annotated spatial domains in all eight samples, based on marker genes in the original study [38] were portal veins ($n = 1,223$), central veins ($n = 720$), haemoglobin ($n = 464$), immune-related domain ($n = 163$), mesenchymal-related domain ($n = 110$), and undetermined spots ($n = 2,183$), which were used to evaluate the clustering and data integration performances of the different methods in the analyses.

Mouse olfactory bulb Slide-seqV2 data

We obtained the data for 20 mouse OB Slide-seqV2 sections in Replicate 2 from CNBI accession number GSE169021 `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE169021`. We used the first 16 slides because of the low quality of the last four slides. The data contained 21,571 genes, on average, over all 693,863 spatial locations across 16 sections, which were equally distributed along the anterior-posterior axis of the same mouse. In the QC steps, we first filtered out genes with non-zero expression levels in fewer than 20 spots and spots with non-zero expression levels for fewer than 20 genes. The filtering step led to a set of 14,307 genes, on average, for a total of 594,890 locations. For each tissue slide, we performed log-normalization followed by centering, and selected the top 2,000 SVGs using SPARK-X [79] for computational efficiency. Next, we prioritized genes based on the number of times they were selected as SVGs in all 16 tissue slides, and we chose the top 2,000 genes as input for the analysis. To evaluate the impact of spatial resolution, we further collapsed nearby spots in each tissue slide using square grids of size 70×70. Then, the same QC steps were performed to obtain normalized expression for the analysis. To examine the structure of the mouse OB, we relied on the structural annotation in the Allen Brain Atlas [42] and used the results of downstream analyses to determine the specific regions of OB.

Human hepatocellular carcinoma Visium data

These data were from two tissue sections each from tumor and tumor-adjacent regions of an HCC patient, and contained 36,601 genes from over 9,813 spatial locations. In the QC steps, we first filtered out genes with non-zero expression levels in fewer than 20 spots and spots with non-zero expression levels for fewer than 20 genes. The filtering step led to a set of 14,851 genes on average from a total of 9,813 locations. For each tissue section, we performed log-normalization, followed by centering, and selected the top 2,000 SVGs using SPARK. Next, we prioritized genes based on the number of times they were selected as SVGs in all four samples, and we chose the top 2,000 genes as input for the analysis. In this data, manual annotations for the TNE and stroma regions were provided by a pathologist using the Visium companion H&E images. We further performed spatial deconvolution to examine the spatial distribution

of malignant cells using RCTD [49].

Data availability

All datasets used in this study are publicly available. These include the 12 human dorsolateral prefrontal cortex Visium datasets (`https://github.com/LieberInstitute/spatialLIBD`), eight mouse liver ST datasets (`https://zenodo.org/record/4399655`), 16 mouse OB Slide-seqV2 datasets (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE169021`) and four human hepatocellular carcinoma Visium datasets (access available upon reviewer's request).

Code availability

The PRECAST methods were implemented in an open-source, publicly available R package that is available at `https://cran.r-project.org/package=PRECAST` and `https://github.com/feiyoung/PRECAST`. Code for reproducing the analysis can be found at `https://github.com/feiyoung/PRECAST_Analysis`.

Acknowledgment

We thank Dr. Juan Zhou for critical reading and feedback.

Author contributions

J.L. initiated and designed the study, W.L., X.L. and Z.L. implemented the model and performed simulation studies and benchmarking evaluation; J.L. wrote the manuscript, and all authors edited and revised the manuscript.

References

[1] Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233), 2015.

[2] Jeffrey R Moffitt, Junjie Hao, Guiping Wang, Kok Hao Chen, Hazen P Babcock, and Xiaowei Zhuang. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences*, 113(39):11046–11051, 2016.

[3] Guiping Wang, Jeffrey R Moffitt, and Xiaowei Zhuang. Multiplexed imaging of high-density libraries of rnas with merfish and expansion microscopy. *Scientific reports*, 8(1): 1–13, 2018.

[4] Eric Lubeck, Ahmet F Coskun, Timur Zhiyentayev, Mubhij Ahmad, and Long Cai. Single-cell in situ rna profiling by sequential hybridization. *Nature Methods*, 11(4):360–361, 2014.

[5] Sheel Shah, Eric Lubeck, Wen Zhou, and Long Cai. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron*, 92(2): 342–357, 2016.

[6] Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, et al. Transcriptome-scale super-resolved imaging in tissues by rna seqfish+. *Nature*, 568(7751): 235–239, 2019.

[7] Patrik L Strahl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

[8] Sanja Vickovic, Gökcen Eraslan, Fredrik Salmén, Johanna Klughammer, Linnea Stenbeck, Denis Schapiro, Tarmo Äijö, Richard Bonneau, Ludvig Bergenstrrahle, José Fernandéz Navarro, et al. High-definition spatial transcriptomics for in situ tissue profiling. *Nature Methods*, 16(10):987–990, 2019.

[9] Samuel G Rodriques, Robert R Stickels, Aleksandrina Goeva, Carly A Martin, Evan Murray, Charles R Vanderburg, Joshua Welch, Linlin M Chen, Fei Chen, and Evan Z Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, 2019.

[10] Robert R Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L Marshall, Daniela J Di Bella, Paola Arlotta, Evan Z Macosko, and Fei Chen. Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqv2. *Nature Biotechnology*, 39(3):313–319, 2021.

[11] 10x Genomics. Visium spatial gene expression., 2019. URL `https://www.10xgenomics.com/products/spatial-gene-expression`.

[12] Anjali Rao, Dalia Barkley, Gustavo S Fran**c**ca, and Itai Yanai. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220, 2021.

[13] Erick Armingol, Adam Officer, Olivier Harismendy, and Nathan E Lewis. Deciphering cell–cell interactions and communication from gene expression. *Nature Reviews Genetics*, 22(2):71–88, 2021.

[14] Isar Nassiri and Matthew N McCall. Systematic exploration of cell morphological phenotypes associated with a transcriptomic query. *Nucleic acids research*, 46(19):e116–e116, 2018.

[15] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(5):547–554, 2019.

[16] Xiaojie Qiu, Yan Zhang, Jorge D Martin-Rufino, Chen Weng, Shayan Hosseinzadeh, Dian Yang, Angela N Pogson, Marco Y Hein, Kyung Hoi Joseph Min, Li Wang, et al. Mapping transcriptomic vector fields of single cells. *Cell*, 185(4):690–711, 2022.

[17] Giovanni Palla, David S Fischer, Aviv Regev, and Fabian J Theis. Spatial components of molecular tissue biology. *Nature Biotechnology*, pages 1–11, 2022.

[18] Edward Zhao, Matthew R Stone, Xing Ren, Jamie Guenthoer, Kimberly S Smythe, Thomas Pulliam, Stephen R Williams, Cedric R Uytingco, Sarah EB Taylor, Paul Nghiem, et al. Spatial transcriptomics at subspot resolution with bayesspace. *Nature Biotechnology*, pages 1–10, 2021.

[19] Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J Irwin, Edward B Lee, Russell T Shinohara, and Mingyao Li. Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature Methods*, 18(11):1342–1351, 2021.

[20] Yi Yang, Xingjie Shi, Wei Liu, Qiuzhong Zhou, Mai Chan Lau, Jeffrey Chun Tatt Lim, Lei Sun, Cedric Chuan Young Ng, Joe Yeong, and Jin Liu. Sc-meb: spatial clustering with hidden markov random field using empirical bayes. *Briefings in bioinformatics*, 23 (1):bbab466, 2022.

[21] Kristen R Maynard, Leonardo Collado-Torres, Lukas M Weber, Cedric Uytingco, Brianna K Barry, Stephen R Williams, Joseph L Catallini, Matthew N Tran, Zachary Besich, Madhavi Tippani, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience*, 24(3):425–436, 2021.

[22] I-Hao Wang, Evan Murray, Greg Andrews, Hao-Ching Jiang, Sung Jin Park, Elisa Donnard, Violeta Durán-Laforet, Daniel M Bear, Travis E Faust, Manuel Garber, et al. Spatial transcriptomic reconstruction of the mouse olfactory glomerular map suggests principles of odor processing. *Nature Neuroscience*, pages 1–9, 2022.

[23] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, 16(12):1289–1296, 2019.

[24] Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, 2018.

[25] Brian Hie, Bryan Bryson, and Bonnie Berger. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nature Biotechnology*, 37(6):685–691, 2019.

[26] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

[27] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

[28] Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, 2019.

[29] Britta Velten, Jana M Braunger, Ricard Argelaguet, Damien Arnol, Jakob Wirbel, Danila Bredikhin, Georg Zeller, and Oliver Stegle. Identifying temporal and spatial patterns of variation from multimodal data using mefisto. *Nature Methods*, pages 1–8, 2022.

[30] Ron Zeira, Max Land, Alexander Strzalkowski, and Benjamin J Raphael. Alignment and integration of spatial transcriptomics data. *Nature Methods*, pages 1–9, 2022.

[31] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[32] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[33] Felix L Yeh, Yuanyuan Wang, Irene Tom, Lino C Gonzalez, and Morgan Sheng. Trem2 binds to apolipoproteins, including apoe and clu/apoj, and thereby facilitates uptake of amyloid-beta by microglia. *Neuron*, 91(2):328–340, 2016.

[34] Xianqin Zhang, Dusan Bogunovic, Béatrice Payelle-Brogard, Véronique Francois-Newton, Scott D Speer, Chao Yuan, Stefano Volpi, Zhi Li, Ozden Sanal, Davood Mansouri, et al. Human intracellular isg15 prevents interferon-α/β over-amplification and auto-inflammation. *Nature*, 517(7532):89–93, 2015.

[35] Mark Hermann and Dusan Bogunovic. Isg15: in sickness and in health. *Trends in immunology*, 38(2):79–93, 2017.

[36] Robert Dantzer, Jason C O'connor, Gregory G Freund, Rodney W Johnson, and Keith W Kelley. From inflammation to sickness and depression: when the immune system subjugates the brain. *Nature reviews neuroscience*, 9(1):46–56, 2008.

[37] Golam M Khandaker, Lesley Cousins, Julia Deakin, Belinda R Lennox, Robert Yolken, and Peter B Jones. Inflammation and immunity in schizophrenia: implications for pathophysiology and treatment. *The Lancet Psychiatry*, 2(3):258–270, 2015.

[38] Franziska Hildebrandt, Alma Andersson, Sami Saarenpää, Ludvig Larsson, Noémi Van Hul, Sachie Kanatani, Jan Masek, Ewa Ellis, Antonio Barragan, Annelie Mollbrink, et al. Spatial transcriptomics to define transcriptional patterns of zonation and structural components in the mouse liver. *Nature communications*, 12(1):1–14, 2021.

[39] Ki M Mak and CY Maximilian Png. The hepatic central vein: structure, fibrosis, and role in liver biology. *The Anatomical Record*, 303(7):1747–1767, 2020.

[40] Xiaoping Han, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, Ziming Zhou, Haide Chen, Fang Ye, et al. Mapping the mouse cell atlas by microwell-seq. *Cell*, 172(5):1091–1107, 2018.

[41] Zhicheng Ji and Hongkai Ji. Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. *Nucleic acids research*, 44(13):e117–e117, 2016.

[42] Ed S Lein, Michael J Hawrylycz, Nancy Ao, Mikael Ayres, Amy Bensinger, Amy Bernard, Andrew F Boe, Mark S Boguski, Kevin S Brockway, Emi J Byrnes, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445(7124):168–176, 2007.

[43] Vivien Marx. Method of the year: spatially resolved transcriptomics. *Nature Methods*, 18 (1):9–14, 2021.

[44] Paulo P Amaral, Christine Neyt, Simon J Wilkins, Marjan E Askarian-Amiri, Susan M Sunkin, Andrew C Perkins, and John S Mattick. Complex architecture and regulated expression of the sox2ot locus during vertebrate development. *Rna*, 15(11):2013–2027, 2009.

[45] Anja Haslinger, Tobias J Schwarz, Marcela Covic, and D Chichung Lie. Expression of sox11 in adult neurogenic niches suggests a stage-specific role in adult neurogenesis. *European Journal of Neuroscience*, 29(11):2103–2114, 2009.

[46] Burak Tepe, Matthew C Hill, Brandon T Pekarek, Patrick J Hunt, Thomas J Martin, James F Martin, and Benjamin R Arenkiel. Single-cell rna-seq of mouse olfactory bulb reveals cellular heterogeneity and activity-dependent molecular census of adult-born neurons. *Cell reports*, 25(10):2689–2703, 2018.

[47] Nader Sanai, Thuhien Nguyen, Rebecca A Ihrie, Zaman Mirzadeh, Hui-Hsin Tsai, Michael Wong, Nalin Gupta, Mitchel S Berger, Eric Huang, Jose-Manuel Garcia-Verdugo, et al. Corridors of migrating neurons in the human brain and their decline during infancy. *Nature*, 478(7369):382–386, 2011.

[48] Shin Nagayama, Ryota Homma, and Fumiaki Imamura. Neuronal organization of olfactory bulb circuits. *Frontiers in neural circuits*, 8:98, 2014.

[49] Dylan M Cable, Evan Murray, Luli S Zou, Aleksandrina Goeva, Evan Z Macosko, Fei Chen, and Rafael A Irizarry. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, pages 1–10, 2021.

[50] Candida Tufo, Subathra Poopalasundaram, Ana Dorrego-Rivas, Marc C Ford, Anthony Graham, and Matthew S Grubb. Development of the mammalian main olfactory bulb. *Development*, 149(3):dev200210, 2022.

[51] Jiangting Hu, Fabrizio Bianchi, Mary Ferguson, Alfredo Cesario, Stefano Margaritora, Pierluigi Granone, Peter Goldstraw, Michelle Tetlow, Cathy Ratcliffe, Andrew G Nicholson, et al. Gene expression signature for angiogenic and nonangiogenic non-small-cell lung cancer. *Oncogene*, 24(7):1212–1219, 2005.

[52] Stefan Bentink, Benjamin Haibe-Kains, Thomas Risch, Jian-Bing Fan, Michelle S Hirsch, Kristina Holton, Renee Rubio, Craig April, Jing Chen, Eliza Wickham-Garcia, et al. Angiogenic mrna and microrna gene expression signature predicts a novel subtype of serous ovarian cancer. *PloS one*, 7(2):e30269, 2012.

[53] Massimo Masiero, Filipa Costa Simões, Hee Dong Han, Cameron Snell, Tessa Peterkin, Esther Bridges, Lingegowda S Mangala, Sherry Yen-Yao Wu, Sunila Pradeep, Demin Li, et al. A core human primary tumor angiogenesis signature identifies the endothelial orphan receptor eltd1 as a key regulator of angiogenesis. *Cancer cell*, 24(2):229–241, 2013.

[54] Benoit Langlois, Falk Saupe, Tristan Rupp, Christiane Arnold, Michaël Van der Heyden, Gertraud Orend, and Thomas Hussenet. Angiomatrix, a signature of the tumor angiogenic switch-specific matrisome, correlates with poor prognosis for glioma and colorectal cancer patients. *Oncotarget*, 5(21):10529, 2014.

[55] Lichun Ma, Maria O Hernandez, Yongmei Zhao, Monika Mehta, Bao Tran, Michael Kelly, Zachary Rae, Jonathan M Hernandez, Jeremy L Davis, Sean P Martin, et al. Tumor cell biodiversity drives microenvironmental reprogramming in liver cancer. *Cancer cell*, 36(4): 418–430, 2019.

[56] Daria Capece, Mariafausta Fischietti, Daniela Verzella, Agata Gaggiano, Germana Cicciarelli, Alessandra Tessitore, Francesca Zazzeroni, and Edoardo Alesse. The inflammatory microenvironment in hepatocellular carcinoma: a pivotal role for tumor-associated macrophages. *BioMed research international*, 2013, 2013.

[57] Katarzyna Sawa-Wejksza and Martyna Kandefer-Szerszeń. Tumor-associated macrophages as target for antitumor therapy. *Archivum immunologiae et therapiae experimentalis*, 66(2):97–111, 2018.

[58] Kaijun Huang, Wenxuan Xie, Shutong Wang, Qiao Li, Xiangling Wei, Bin Chen, Yunpeng Hua, Shaoqiang Li, Baogang Peng, and Shunli Shen. High spink1 expression predicts poor prognosis and promotes cell proliferation and metastasis of hepatocellular carcinoma. *Journal of Investigative Surgery*, 34(9):1011–1020, 2021.

[59] Maryam Tahmasebi Birgani, Mohammadreza Hajjari, Arman Shahrisa, Atefeh Khoshnevisan, Zahra Shoja, Paria Motahari, and Baharak Farhangi. Long non-coding rna snhg6 as a potential biomarker for hepatocellular carcinoma. *Pathology & Oncology Research*, 24(2):329–337, 2018.

[60] Yun Hee Kang, Mi-Young Park, Do-Young Yoon, Seung Ro Han, Chung Il Lee, Na Young Ji, Pyung-Keun Myung, Hee Gu Lee, Jae Wha Kim, Young Il Yeom, et al. Dysregulation of overexpressed il-32α in hepatocellular carcinoma suppresses cell growth and induces apoptosis through inactivation of nf-κb and bcl-2. *Cancer letters*, 318(2):226–233, 2012.

[61] Wei Liu, Xu Liao, Yi Yang, Huazhen Lin, Joe Yeong, Xiang Zhou, Xingjie Shi, and Jin Liu. Joint dimension reduction and clustering analysis of single-cell RNA-seq and spatial transcriptomics data. *Nucleic Acids Research*, 03 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac219. URL `https://doi.org/10.1093/nar/gkac219`. gkac219.

[62] Yao-Hung Hubert Tsai, Martin Q Ma, Muqiao Yang, Han Zhao, Louis-Philippe Morency, and Ruslan Salakhutdinov. Self-supervised representation learning with relative predictive coding. *arXiv preprint arXiv:2103.11275*, 2021.

[63] Yingxin Lin, Tung-Yu Wu, Sheng Wan, Jean YH Yang, Wing H Wong, and YX Wang. scjoint integrates atlas-scale single-cell rna-seq and atac-seq data with transfer learning. *Nature Biotechnology*, pages 1–8, 2022.

[64] Carter Allen, Yuzhou Chang, Brian Neelon, Won Chang, Hang J Kim, Zihai Li, Qin Ma, and Dongjun Chung. A bayesian multivariate mixture model for spatial transcriptomics data. *bioRxiv*, 2021.

[65] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[66] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[67] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

[68] Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. Normalization of rna-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, 32(9): 896–902, 2014.

[69] Bidossessi Wilfried Hounkpe, Francine Chenou, Franciele de Lima, and Erich Vinicius De Paula. Hrt atlas v1. 0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive rna-seq datasets. *Nucleic acids research*, 49(D1):D947–D955, 2021.

[70] Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research*, 47(W1):W191–W198, 2019.

[71] Kelly Street, Davide Risso, Russell B Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*, 19(1):1–16, 2018.

[72] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[73] Hansheng Wang, Bo Li, and Chenlei Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683, 2009.

[74] Yingxin Lin, Shila Ghazanfar, Kevin YX Wang, Johann A Gagnon-Bartsch, Kitty K Lo, Xianbin Su, Ze-Guang Han, John T Ormerod, Terence P Speed, Pengyi Yang, et al. scmerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell rna-seq datasets. *Proceedings of the National Academy of Sciences*, 116 (20):9775–9784, 2019.

[75] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2 (1):193–218, 1985.

[76] Thomas M Cover and Joy A Thomas. *Elements of information theory 2nd edition (wiley series in telecommunications and signal processing)*. Wiley-Interscience, 2006.

[77] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[78] Shiquan Sun, Jiaqiang Zhu, and Xiang Zhou. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature Methods*, 17(2):193–200, 2020.

[79] Jiaqiang Zhu, Shiquan Sun, and Xiang Zhou. Spark-x: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biology*, 22(1):1–25, 2021.

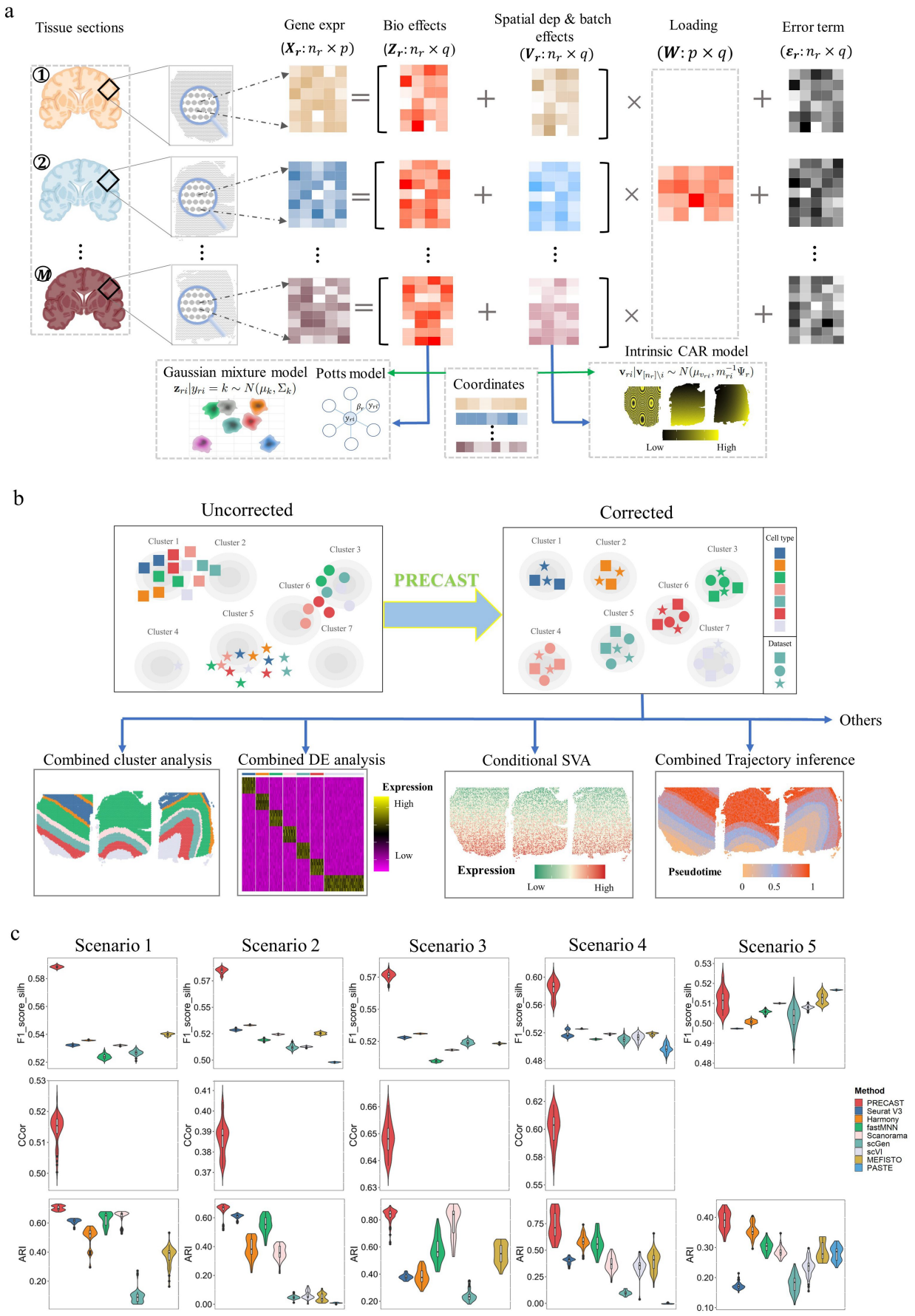
a
Tissue sections
Gene expr
$(X_r: n_r \times p)$
Bio effects
$(Z_r: n_r \times q)$
Spatial dep & batch effects
$(V_r: n_r \times q)$
Loading
$(W: p \times q)$
Error term
$(\varepsilon_r: n_r \times q)$
Gaussian mixture model
$z_{ri}|y_{ri} = k \sim N(\mu_k, \Sigma_k)$
Potts model
Coordinates
Intrinsic CAR model
$v_{ri}|v_{[n_r]\setminus i} \sim N(\mu_{v_{ri}}, m_{ri}^{-1}\Psi_r)$
Low
High
b
Uncorrected
Corrected
PRECAST
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cell type
Dataset
Others
Combined cluster analysis
Combined DE analysis
Conditional SVA
Combined Trajectory inference
Expression
High
Low
Pseudotime
0
0.5
1
c
Scenario 1
Scenario 2
Scenario 3
Scenario 4
Scenario 5
F1_score_silh
CCor
ARI
Method
PRECAST
Seurat V3
Harmony
fastMNN
Scanorama
scGen
scVI
MEFISTO
PASTE

Figure 1: Schematic overview of PRECAST and simulation results. (a) PRECAST is a unified probabilistic factor model that simultaneously estimates aligned embeddings and cluster labels with consideration of spatial smoothness in both the cluster label and low-dimensional embedding spaces. It takes, as input, the normalized gene expression matrices from multiple tissue slides. (b) Examples of PRECAST downstream analyses. (c) In the simulations, we considered two ways to generate gene expression, continuous or count matrix, followed by log-normalization. We also considered two ways to generate spatial coordinates and cell/domain labels: Potts model and three cortex tissues from the DLPFC data. Lastly, we considered a scenario favorable for PASTE. We evaluated the performance of the data integration, estimation of aligned embeddings, slide-specific embeddings due to neighboring microenvironments, and spatial clustering using F1 scores of average silhouette coefficients, canonical correlation coefficients (CCor), and the adjusted Rand index (ARI) and ARIs displayed for the other methods were evaluated based on the results of the spatial clustering method SC-MEB. PRECAST outperformed all other data integration methods, according to the assessments in Scenarios 1-4 and was comparable with PASTE in Scenario 5. In simulations, scVI and PASTE are only applicable to Scenarios 2, 4 and 5 with count matrix and only PRECAST estimates the slide-specific embeddings. We evaluate the CCor with underlying truth in Scenario 1-4.

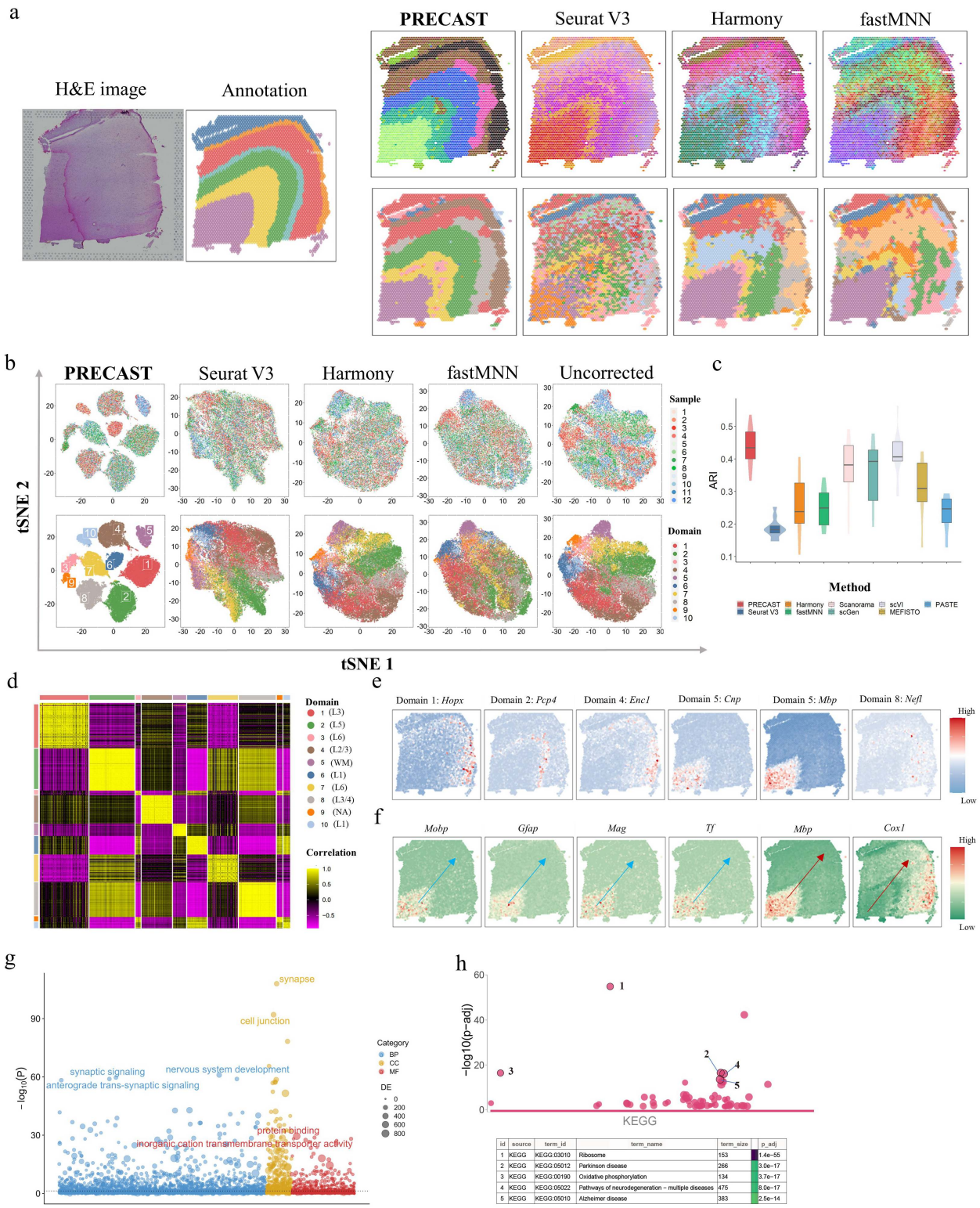
a
H&E image
Annotation
PRECAST
Seurat V3
Harmony
fastMNN
b
PRECAST
Seurat V3
Harmony
fastMNN
Uncorrected
Sample
Domain
tSNE 2
tSNE 1
c
ARI
Method
PRECAST
Seurat V3
Harmony
fastMNN
Scanorama
scGen
scVI
MEFISTO
PASTE
d
Domain
1 (L3)
2 (L5)
3 (L6)
4 (L2/3)
5 (WM)
6 (L1)
7 (L6)
8 (L3/4)
9 (NA)
10 (L1)
Correlation
e
Domain 1: Hopx
Domain 2: Pcp4
Domain 4: Enc1
Domain 5: Cnp
Domain 5: Mbp
Domain 8: Nefl
High
Low
f
Mobp
Gfap
Mag
Tf
Mbp
Cox1
High
Low
g
synapse
cell junction
synaptic signaling
anterograde trans-synaptic signaling
nervous system development
protein binding
inorganic cation transmembrane transporter activity
−log10(P)
Category
BP
CC
MF
DE
h
−log10(p−adj)
KEGG
id | source | term_id | term_name | term_size | p_adj
1 | KEGG | KEGG:03010 | Ribosome | 153 | 1.4e−55
2 | KEGG | KEGG:05012 | Parkinson disease | 266 | 3.0e−17
3 | KEGG | KEGG:00190 | Oxidative phosphorylation | 134 | 3.7e−17
4 | KEGG | KEGG:05022 | Pathways of neurodegeneration − multiple diseases | 475 | 8.0e−17
5 | KEGG | KEGG:05010 | Alzheimer disease | 383 | 2.5e−14

Figure 2: Analysis of data for 12 human DLPFC sections. (a) Left panel: H&E image and manual annotation of sample ID151674. Top panel: UMAP RGB plots of sample ID151674 for PRECAST, Seurat V3, Harmony, and fastMNN. Bottom panel: Clustering assignment heatmaps for these four methods. Color scheme used in clustering assignment heatmap for PRECAST is the same with (b) and (d). (b) tSNE plots for these four data integration methods with right-most column showing the analysis without correction; domains are labeled as in (d). (c) Box/violin plot of ARI values for PRECAST and other methods; SC-MEB was used in the other methods for clustering based on their aligned embeddings. (d) Heatmap of Pearson's correlation coefficients among detected domains. L1-L6, Layer 1 - Layer 6; WM, white matter; NA, undetermined. (e) Spatial expression patterns of DE genes for Domain 1 (*Hopx*), Domain 2 (*Pcp4*), Domain 4 (*Enc1*), Domain 5 (*Cnp*), Domain 5 (*Mbp*), and Domain 8 (*Nefl*) for sample ID151674. (f) Spatial expression patterns of genes associated with pseudotime: *Mobp*, *Gfap*, *Mag*, *Tf*, *Mbp*, and *Cox1*, where the arrow represents the direction of the increased pseudotime. (g) Bubble plot of -log10(p-values) for GO enrichment analysis of genes associated with pseudotime. (h) Bubble plot of -log10(p-values) for KEGG enrichment analysis of SVGs while adjusting domain-relevant aligned embeddings by PRECAST for sample ID151674.

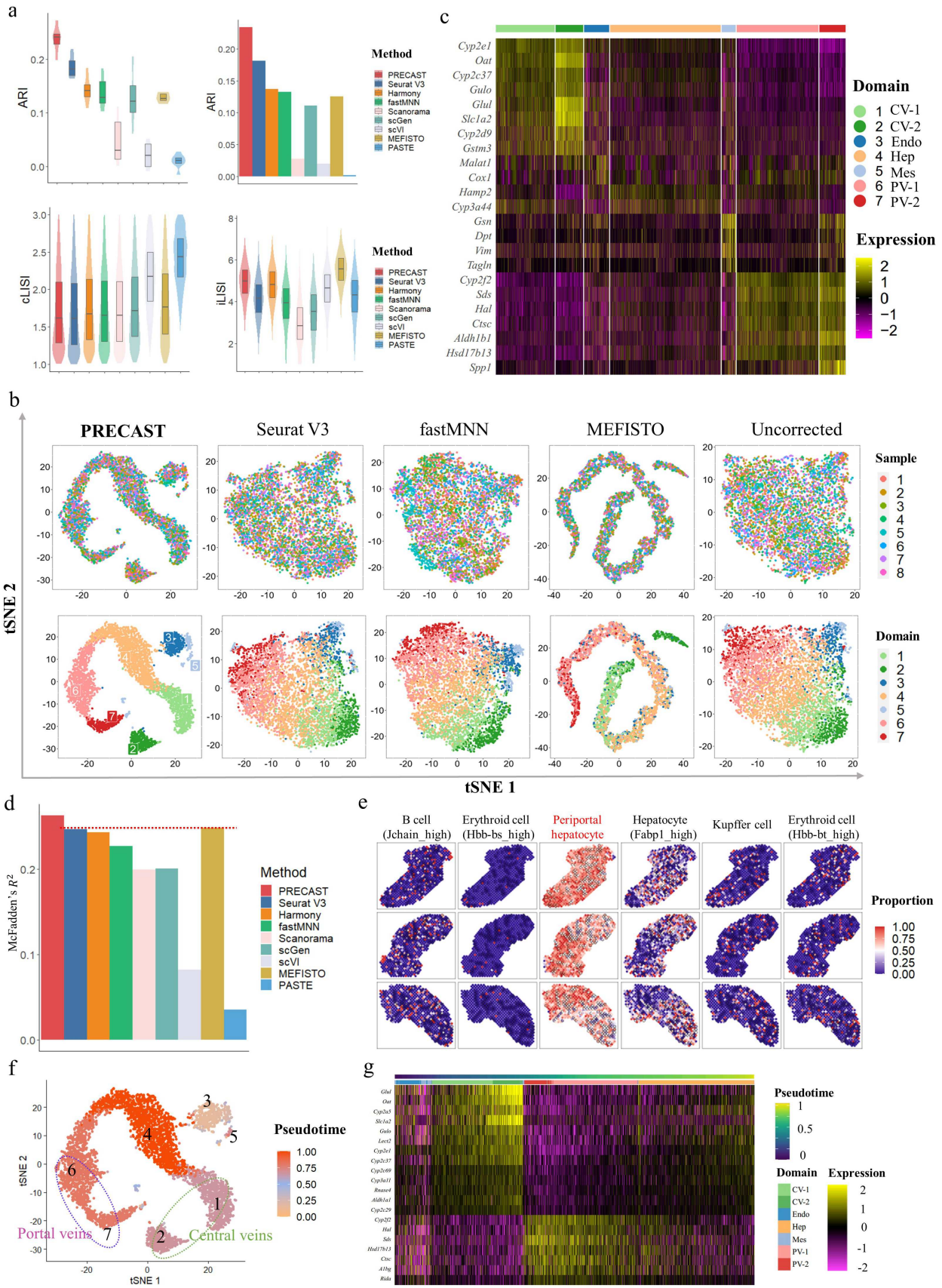
a
ARI
cLISI
iLISI
Method
PRECAST
Seurat V3
Harmony
fastMNN
Scanorama
scGen
scVI
MEFISTO
PASTE
c
Cyp2e1
Oat
Cyp2c37
Gulo
Glul
Slc1a2
Cyp2d9
Gstm3
Malat1
Cox1
Hamp2
Cyp3a44
Gsn
Dpt
Vim
Tagln
Cyp2f2
Sds
Hal
Ctsc
Aldh1b1
Hsd17b13
Spp1
Domain
1 CV-1
2 CV-2
3 Endo
4 Hep
5 Mes
6 PV-1
7 PV-2
Expression
b
PRECAST
Seurat V3
fastMNN
MEFISTO
Uncorrected
Sample
tSNE 2
tSNE 1
Domain
d
McFadden's R^2
Method
e
B cell (Jchain_high)
Erythroid cell (Hbb-bs_high)
Periportal hepatocyte
Hepatocyte (Fabp1_high)
Kupffer cell
Erythroid cell (Hbb-bt_high)
Proportion
f
Pseudotime
Portal veins
Central veins
g
Pseudotime
Domain
Expression

Figure 3: Analysis of data for eight mouse liver sections. (a) Top panel: Box/violin plot of ARI values of each sample for PRECAST and the other methods (left); bar plot of ARI value of combined samples for PRECAST and other methods (right). Bottom panel: Box/violin plots of cLISI and iLISI values for PRECAST and other methods. Color scheme of each method is the same as in (d). (b) tSNE plots for four data integration methods, with the right-most column showing analysis without correction. Color scheme of each domain is the same as in (c) and (g). (c) Heatmap of differentially expressed genes for each domain identified by PRECAST. CV-1, central veins 1; CV-2, central veins 2; Endo, endothelial cells; Hep, hepcidin-related cells; Mes, mesenchymal-related cells; PV-1, portal veins 1; PV-2, portal veins 2. (d) Bar plot of McFadden's adjusted R^2 values for PRECAST and other methods. McFadden's adjusted R^2 measures the association between the cluster label obtained by each method and the cell proportion obtained by RCTD cell-type deconvolution, and the a larger value indicates a stronger association. (e) Visualization of the cell type proportions mapped to spatial coordinates for six cell types of the first three samples. (f) Visualization of the combined trajectory inferred by PRECAST in tSNE plot of all samples. Domains 1-2 and Domains 6-7 representing central veins and portal veins, respectively, are circled. (g) Heatmap for genes with expression change in the Slingshot pseudotime inferred by PRECAST.

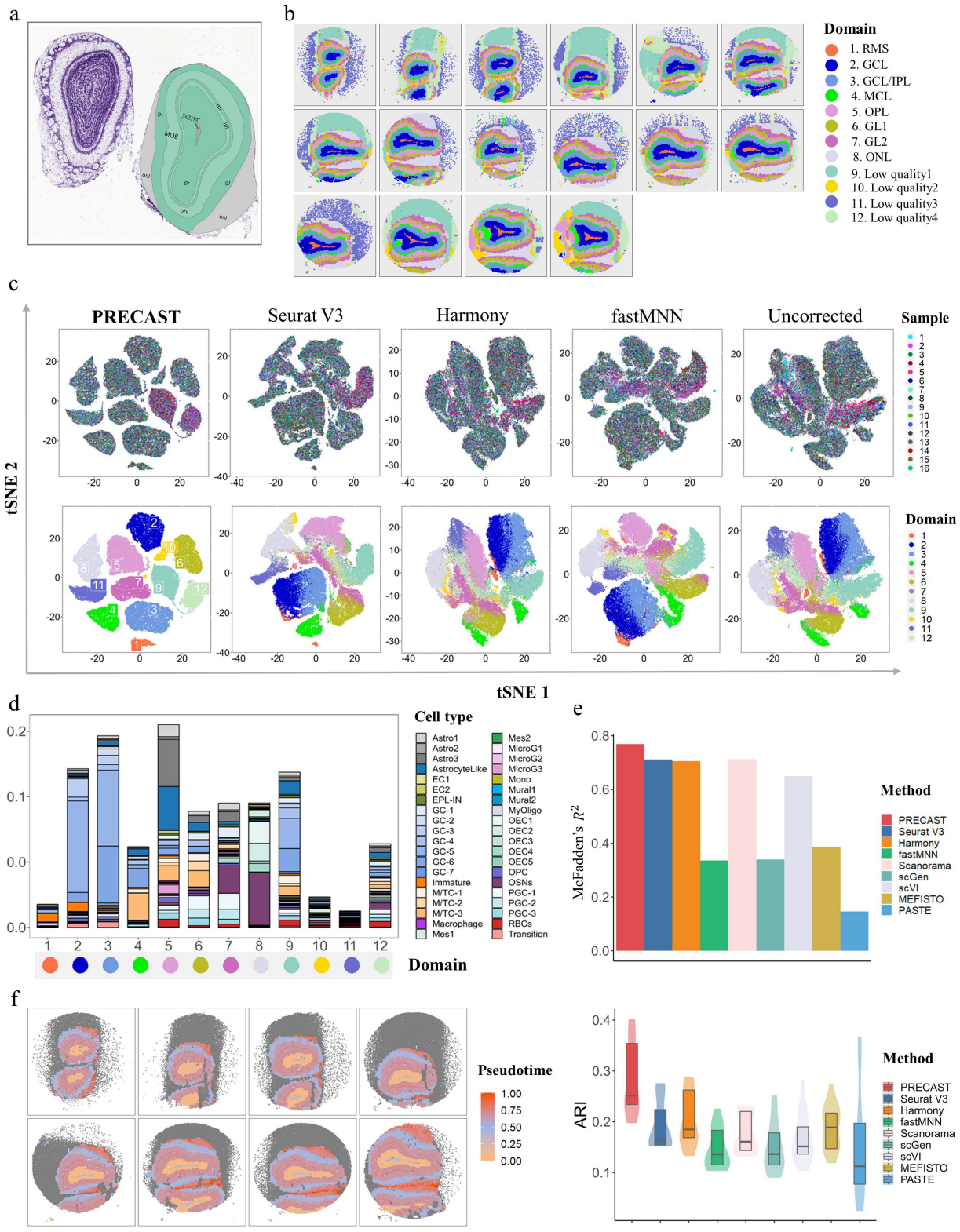
a
b
Domain
1. RMS
2. GCL
3. GCL/IPL
4. MCL
5. OPL
6. GL1
7. GL2
8. ONL
9. Low quality1
10. Low quality2
11. Low quality3
12. Low quality4
c
PRECAST
Seurat V3
Harmony
fastMNN
Uncorrected
Sample
Domain
tSNE 2
tSNE 1
d
Cell type
Astro1
Astro2
Astro3
AstrocyteLike
EC1
EC2
EPL-IN
GC-1
GC-2
GC-3
GC-4
GC-5
GC-6
GC-7
Immature
M/TC-1
M/TC-2
M/TC-3
Macrophage
Mes1
Mes2
MicroG1
MicroG2
MicroG3
Mono
Mural1
Mural2
MyOligo
OEC1
OEC2
OEC3
OEC4
OEC5
OPC
OSNs
PGC-1
PGC-2
PGC-3
RBCs
Transition
Domain
e
McFadden's R^2
Method
PRECAST
Seurat V3
Harmony
fastMNN
Scanorama
scGen
scVI
MEFISTO
PASTE
f
Pseudotime
1.00
0.75
0.50
0.25
0.00
ARI
Method
PRECAST
Seurat V3
Harmony
fastMNN
Scanorama
scGen
scVI
MEFISTO
PASTE

Figure 4: Analysis of data for 16 mouse olfactory bulb sections. (a) Structure of the mouse olfactory bulb annotated using the Allen Brain Atlas. (b) Clustering assignment heatmaps for 16 tissue slides by PRECAST, where the first row shows samples 1-6, the second row samples 7-12, and the last row samples 13-16 (RMS, rostral migratory stream; GCL, granule cell layer; IPL, inner plexiform layer; MCL, mitral cell layer; OPL, outer plexiform layer; GL, glomerular layer; ONL, olfactory nerve layer). Color scheme for domains detected in PRECAST is as in (c) and (d), and the order of domain labels in (b) is the same as in (c), (d), and (e). (c) tSNE plots for four data integration methods with the right-most column showing analysis without correction. (d) Percentage of different cell types in each domain detected by PRECAST with scaling. (e) MacFadden's adjusted R^2 between the inferred cell type proportions and the estimated domain labels by PRECAST and other methods (top panel); bar plot of ARI values of combined samples for PRECAST and other methods, where each spot is annotated using the cell type, with the highest proportion from the spatial deconvolution (bottom panel). (f) Visualization of the trajectory inferred by PRECAST in spatial heatmap for samples 1-8 (left panel) and tSNE plot of all samples(right panel).

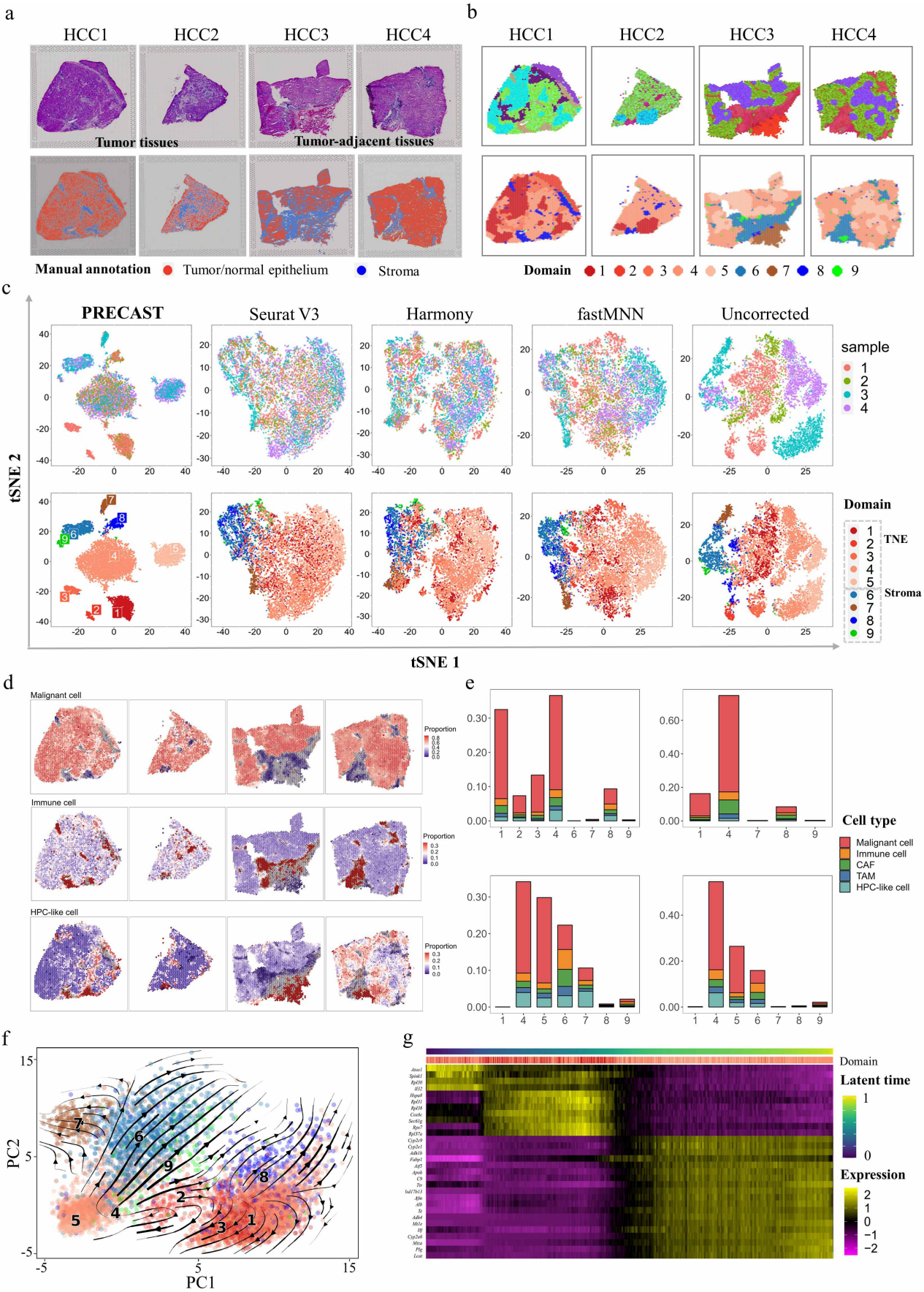
a
HCC1
HCC2
HCC3
HCC4
Tumor tissues
Tumor-adjacent tissues
Manual annotation
Tumor/normal epithelium
Stroma
b
HCC1
HCC2
HCC3
HCC4
Domain
1 2 3 4 5 6 7 8 9
c
PRECAST
Seurat V3
Harmony
fastMNN
Uncorrected
sample
1
2
3
4
tSNE 2
Domain
TNE
Stroma
tSNE 1
d
Malignant cell
Immune cell
HPC-like cell
Proportion
e
Cell type
Malignant cell
Immune cell
CAF
TAM
HPC-like cell
f
PC1
PC2
g
Domain
Latent time
Expression

Figure 5: Analysis of data for four human HCC sections. (a) Top panel: H&E images from four tissue slides. Bottom panel: Manual annotation by a pathologist of four tissue slides. (b) Top panel: UMAP RGB plots of PRECAST for four tissue slides. Bottom panel: Clustering assignment heatmaps for four tissue sections by PRECAST. Color scheme for clustering assignment heatmap in PRECAST is the same as in (c), (f), and (g). (c) tSNE plots for four data integration methods with the right-most column showing analysis without correction; domains are labeled as in (e) and (f). TNE, tumor/normal epithelium. (d) Spatial heatmap of deconvoluted cell proportions in malignant cells, immune cells, and HPC-like cells. (e) Percentage of different cell types in each domain detected by PRECAST, with (left panel) or without (right panel) scaling to the summation of all cell types across all domains equal to 100%. (f) PC plot of estimated RNA velocity. (g) Heatmap of genes with expression change in latent time.