FAST: a fast and scalable factor analysis for spatially aware dimension reduction of multi-section spatial transcriptomics data

Wei Liu[1,2†], Xiao Zhang[1†], Xiaoran Chai[3], Zhenqian Fan[4], Huazhen Lin[5], Jinmiao Chen[6], Lei Sun[2], Tianwei Yu[1], Joe Yeong[7,8*], and Jin Liu[1*]

[1]School of Data Science, The Chinese University of Hong Kong-Shenzhen, Shenzhen, China

[2]Cardiovascular and Metabolic Disorders Program, Duke-NUS Medical School, Singapore

[3]Cancer and Stem Cell Biology Program, Duke-NUS Medical School, Singapore

[4]Department of Endocrinology, the Second Hospital of Tianjin Medical University, Tianjin, China

[5]Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Chengdu, China

[6]Singapore Immunology Network, Agency of Science, Technology and Research (A*STAR), Singapore

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

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

Abstract

Biological techniques for spatially resolved transcriptomics (SRT) have advanced rapidly in both throughput and spatial resolution for a single spatial location. This progress necessitates the development of efficient and scalable spatial dimension reduction methods that can handle large-scale SRT data from multiple sections. Here, we developed FAST as a fast and efficient generalized probabilistic factor analysis for spatially aware dimension reduction, which simultaneously accounts for the count nature of SRT data and extracts a low-dimensional representation of SRT data across multiple sections, while preserving biological effects with consideration of spatial smoothness among nearby locations.

Compared with existing methods, FAST uniquely models the count data across multiple sections while using a local spatial dependence with scalable computational complexity. Using both simulated and real datasets, we demonstrated the improved correlation between FAST estimated embeddings and annotated cell/domain types. Furthermore, FAST exhibits remarkable speed, with only FAST being applicable to analyze a mouse embryo Stereo-seq dataset with >2.3 million locations in only 2 hours. More importantly, FAST identified the differential activities of immune-related transcription factors between tumor and non-tumor clusters and also predicted a carcinogenesis factor *CCNH* as the upstream regulator of differentially expressed genes in a breast cancer Xenium dataset.

Introduction

Spatially resolved transcriptomics (SRT) encompasses a set of breakthrough technologies that enable gene expression profiling with spatial information on tissues. Spatial location information is of paramount significance in comprehending the mechanisms underlying processes such as cell biology [1], tumor biology [2], and developmental biology [3]. Among the many factors that influence the choice of these technologies [4, 5], throughput in profiling and spatial resolution are two of the most important. Technologies based on in situ hybridization (e.g., MERFISH [6], seqFISH [7], and seqFISH+ [8]) and in situ sequencing (ISS) (e.g., FISSEQ [9], and Xenium) provide single-molecule/single-cell resolution, but are for targeted genes that require prior knowledge. While next-generation sequencing (NGS)-based technologies, such as Visium, Slide-seq [10, 11], and Stereo-seq [12], are unbiased and involve high-throughput expression measurements, most do not provide single-cell resolution. The diverse range of SRT technologies has enabled the exploration of intricate transcriptional structure across heterogeneous tissues and will revolutionize the worlds of cell biology and molecular biology and advance our understanding in many areas of biology [13, 14].

For “high-dimensional”, often noisy, expression measurements obtained using SRT technologies, dimensionality reduction is a key step in generating a low-dimensional data representation that enriches biological signals by aggregating gene expression relevant to biological effects [15]. Moreover, transformation in dimension reduction circumvents the curse of dimensionality usually present in “high-dimensional” expression profiles constructed in genomic studies, including those based on SRT [16, 17]. A plethora of dimension reduction methods have been developed, including the most popular method, principal component analysis (PCA) [18], which is routinely used in many software pipelines, such as Seurat [19] and Cell Ranger [20] for single-cell RNA sequencing (scRNA-seq) analysis and BayesSpace [21], SpaGCN [22], and SC-MEB [23] for SRT data analysis. However, PCA does not consider the spatial nature of SRT data in the process of estimating low-dimensional embeddings, omitting the influence of the microenvironment in neighboring locations.

In SRT data analysis, expression patterns among neighboring locations exhibit the “similarity” induced by the shared microenvironment. Recently, SpatialPCA [17] and non-negative spatial factorization (NSF) [24] were proposed for spatially aware dimension reduction using Gaussian-type kernels over spatial locations. To reduce the computational burden, SpatialPCA applies a low-rank approximation, while NSF implements a sparse Gaussian process but is not applicable to spatial locations from multiple sections [24, 25]. With improved spatial resolution, the number of spatial locations profiled increases substantially, while multiple sections are needed to either generate a spatial map or recover the spatiotemporal transcriptomics atlas of a whole organ [12, 26, 27]. More recently, we proposed PRECAST as a method to unify the tasks of dimension reduction, cluster allocation, and embedding alignment for SRT data from multiple sections [28]. However, PRECAST did not model the count nature of SRT data, while it was designed for a unified task of dimension reduction and clustering. When facing downstream tasks other than clustering (e.g., trajectory inference and cell-cell interaction; CCI), PRECAST may not be optimal. Although the computational order for PRECAST is linear to location number n, it still takes a few days to complete the analysis if the number of spatial locations goes into the millions due to its unified framework. Ideally, an efficient method that allows for spatial dimension reduction across multiple sections and is capable of capturing information for both biological effects and spatial correlation structure is also required. This method should also account for the count nature of SRT data, be applicable to multiple downstream tasks, and be scalable to millions of spatial locations.

To address the limitations of existing methods and facilitate scalability, we proposed FAST as a generalized probabilistic factor analysis model for spatial transcriptomics that efficiently estimates embeddings across multiple sections intrinsic to biological effects, taking into account local expression similarities induced by the shared microenvironment. Uniquely, FAST explicitly allows simultaneous spatially aware dimensionality reduction across multiple sections while modeling the count nature of many existing SRT datasets. Moreover, scalability is promoted by modeling of local spatial dependence using a conditional autoregressive component with an improved computational complexity linear to n, facilitating its applicability to the analysis of multi-section high-resolution SRT data. FAST is versatile in analyzing various multi-section SRT datasets obtained from distinct spatial transcriptomics technologies and tissue structures.

Results

Spatial dimension reduction using FAST

Similar to scRNA-seq analysis, dimension reduction is an essential step for many downstream analyses (Fig. 1a, left panel). We describe FAST in the "Methods" section and provide its technical details in the Supplementary Notes. Briefly, FAST is a generalized probabilistic factor analysis for spatially aware dimension reduction across multi-section spatial transcriptomics data with millions of spatial locations (Fig. 1b). Taking the normalized/count gene expression matrices from multiple sections as input data, FAST factorizes the expression matrices into factor matrices with a shared loading matrix, while assuming a conditional autoregressive (CAR) component for factors from each section (Fig. 1a, right panel). Instead of applying a global kernel for all spatial locations, FAST models local spatial dependence induced by neighboring microenvironments using CAR components. We showed that this consideration in FAST not only reduces the computational complexity in linear form to n, scalable to millions of spatial locations, but also estimates meaningful embeddings that correlate more closely with biological effects. Subsequently, by performing an integrative analysis to remove batch effects via iSC-MEB, the aligned embeddings can be paired with many existing software/tools developed in scRNA-seq studies to enhance the effectiveness of downstream analyses in SRT studies (Fig. 1c). FAST is implemented as an R package available at `https://github.com/feiyoung/FAST`.

Validation using simulated data

We conducted extensive simulations to assess the performance of FAST and compare it to several other methods (Fig. 1d; Supplementary Fig. S1). The methods compared included a range of spatially aware and non-spatially aware dimension reduction methods such as SpatialPCA [17], PRECAST [28], DR-SC [29], scVI [30], PCA, multiBatchPCA [31], NMF and LIGER [32]. Among these methods, SpatialPCA provides a full-rank, SpatialPCA-F, and a low-rank approximation version, SpatialPCA-L. We performed simulations using three sections from the human dorsolateral prefrontal cortex (DLPFC) Visium dataset with eight predefined spatial domains [33], comprising layers 1–6, whiter matter (WM), and unknown cells (Unknown). Supplementary Fig. S1a provides a visual representation of the spatial distribution of the eight domains for each section. The simulation details are provided in the "Methods" section. Briefly, we considered four simulation scenarios in combinations of high and low values for both biological effects and batch effects, reflecting both shared and section-specific effects.

To quantify the performance of the embedding estimation, we calculated the adjusted McFadden's pseudo R^2 (adjusted R^2_{McF}) between the estimated embeddings and the true labels for each section. As shown in Fig. 1d (top panel), FAST outperformed other spatially aware methods such as PRECAST and SpatialPCA, and performed much better than the non-spatially aware methods such as PCA and scVI, across all scenarios. In general, spatially aware methods performed better than the non-spatially aware methods. With biological effects between domains varying from strong to weak, the performance of all methods decreased, but the Poisson version of FAST, FAST-P, was the least sensitive compared with the other methods. However, all methods suffered slightly from the increased batch effects in each section. Despite its inferiority to FAST-P, the Gaussian version of FAST, FAST-G, was more computationally efficient. In all scenarios, both versions of FAST were more computationally efficient than other spatially aware methods such as SpatialPCA and PRECAST (Fig. 1d, bottom panel). By applying iSC-MEB, we aligned embeddings estimated using different methods to detect spatial domains. Not surprisingly, FAST-P achieved the highest adjusted Rand index (ARI) and normalized mutual information (NMI) across all scenarios (Supplementary Fig. S1b).

To evaluate the scalability of FAST, we compared its computational efficiency by varying the number of spatial locations analyzed while fixing the number of genes to 2000 (Fig. 1e). Clearly, both FAST-P and FAST-G demonstrated superior computational efficiency. The computational complexity orders for both FAST and PRECAST were linear to the number of spatial locations while they both used less memory. In our study, SpatialPCA-F required approximately 19 hours and 287 GB of memory to analyze a dataset with around 20,000 locations, but experienced breakdowns when reaching 40,000 locations. Meanwhile, SpatialPCA-L took approximately 6 hours and 108 GB of memory to analyze a dataset with 78,000 locations, but experienced breakdowns at 80,000 locations. However, FAST-P and FAST-G exhibited impressive scalability. FAST-P was able to analyze 200,000 locations in just 30 minutes with 7 GB of memory usage, while FAST-G achieved the same task in only 10 minutes with 5 GB of memory usage.

Application to the human dorsolateral prefrontal cortex Visium dataset

We applied FAST and the other methods to the analysis of four published datasets obtained via either Visium, Xenium, or Stereo-seq technologies (see "Methods"). The four datasets included a DLPFC dataset [33] and a hepatocellular carcinoma dataset [34] generated using 10× Visium, a breast cancer dataset [35] generated using 10× Xenium, and a mouse embryo dataset [12] generated using Stereo-seq. First, we examined the dimension reduction performance of FAST in comparison with SpatialPCA, PRECAST, DR-SC, scVI, PCA, multiBatchPCA, NMF, and LIGER, followed by the application of iSC-MEB to perform clustering analysis and align embeddings from multiple sections. For some or all four datasets, downstream analyses were also performed, including DE analysis, CCI analysis, cell-type deconvolution, and somatic mutation.

To quantify the ability of FAST to outperform existing methods for dimension reduction, we first analyzed the LIBD human DLPFC dataset generated using 10× Visium [33], which contained a total of 47,681 spatial locations across 12 tissue sections from three donors. Taking the manual annotations for the tissue layers based on the cytoarchitecture provided by the original study as ground truth, we were able to evaluate the performance of the dimension reduction. For this purpose, we used the adjusted McFadden's pseudo R^2 (adjusted $\mathrm{R}^2_{\mathrm{McF}}$) between the estimated embeddings and manual annotations for each section. As shown in Fig. 2a, FAST achieved the highest adjusted $\mathrm{R}^2_{\mathrm{McF}}$ while it was 50 times faster and required only 3% memory usage of SpatialPCA with a low-rank approximation, SpatialPCA-L. In detail, FAST-P required 661 seconds with 2 GB memory usage to complete the analysis compared with the 37,172 seconds and 80 GB memory usage required by SpatialPCA-L. After obtaining estimated embeddings from a variety of dimension reduction methods, we performed integrative clustering analysis to detect spatial domains by applying iSC-MEB. For each method, we summarized the aligned embeddings using three components extracted from UMAP. We then visualized the resulting UMAP components using red/green/blue (RGB) colors in the RGB plot (Fig. 2b, upper right panel; Supplementary Fig. S2), accompanied by the corresponding spatial heatmap of cluster assignment (Fig. 2b lower right panel; Supplementary Fig. S3) and the tSNE plot (Supplementary Fig. S4). The results using FAST embeddings exhibited stronger laminar patterns while presenting a harmonious blending of locations from various sections. These findings illustrated the utility of FAST for estimating embeddings of high-dimensional expression profiles among spatial locations. To evaluate the clustering accuracy of the methods, we used both ARI and NMI. As shown in Fig. 2c, FAST achieved the highest ARI and NMI, with median ARIs for FAST-P, FAST-G, SpatialPCA-L, for PRECAST, DR-SC, scVI, PCA, multiBatchPCA and NMF of 0.56, 0.52, 0.20, 0.45, 0.39, 0.45, 0.42, 0.42 and 0.41, respectively.

A key feature of FAST is its ability to estimate low-dimensional embeddings for spatial locations across multiple sections, facilitating many downstream analyses requiring cross-section embedding alignment such as DE analysis and CCI analysis. First, we performed DE analysis for all 12 sections by removing unwanted variations between expression profiles in multiple sections (see "Methods"). In total, we detected 1069 differentially expressed genes (DEGs) with adjusted p-values < 0.001 among the eight domains identified by FAST-P, with 163 genes specific to Domain 1, which corresponds to layer 1 (Supplementary Data 1). A dot plot of normalized expression aligned across 12 sections showed good separation of the DEGs across the detected spatial domains, many of which are human layer-specific markers such as *PCP4* [36], *DIRAS2* [37], *MBP* [33], and *MOBP* [33] (Fig. 2d).

Next, we performed CCI analysis for all spatial locations across multiple sections using CellChat [38] (see "Supplementary Notes"). We observed strong spatial patterns in both the number and the strength of interactions (Fig. 2e), with WM showing substantial interactions with other layers, consistent with its crucial role in transmitting messages between different regions of the brain [39]. We further examined the signaling pathways enriched in each layer during cell-cell communications and found that WM had the highest score in both incoming and outgoing signals, sending electrical signals across different layers via 11 signaling pathways (Supplementary Fig. S5). These pathways included somatostatin, a known presynaptic modulator of glutamatergic signaling in the central nervous system, and the neuregulins signaling pathway, which is crucial in various aspects of the nervous system, including the development, maintenance, and repair processes [40]. Moreover, an estimated PAGA graph [41] generated using FAST-P embeddings demonstrated an almost linear development trajectory from WM to layer 1 in many of the 12 DLPFC sections (Fig. 2f, top panel; Supplementary Fig. S6). But, when employing PRECAST embeddings for the PAGA graph, the resulting topology showed a messy structure (Fig. 2f, bottom panel; Supplementary Fig. S7).

Application to the breast cancer Xenium dataset

We further applied FAST and the other methods to analyze two breast cancer sections generated using 10× Xenium [35] and containing a total of 72,651 spatial locations with expression profiling for 313 genes and their corresponding H&E images (Fig. 3a). In detail, we estimated 15-dimensional embeddings using FAST and the other methods followed by the application of iSC-MEB for clustering analysis and alignment of the embeddings for the two breast cancer sections. For each method, we summarized the aligned embeddings using three components extracted from UMAP. We then visualized the resulting UMAP components using RGB colors in the RGB plot (Fig. 3b and Supplementary Fig. S8). As shown in Fig. 3c and Supplementary Fig. S9, we detected a total of 17 domains in two sections using FAST-P, with the analysis completed in 133 seconds using 0.33 GB memory usage while SpatialPCA-L required 105,417 seconds and 11.31 GB memory usage (Fig. 3d and Supplementary Fig. S10). The estimated cluster proportions matched well in two adjacent sections (Fig. 3e).

After removing unwanted variations in expression profiles, we performed DE analysis for the two sections and detected a total of 1416 DEGs with adjusted p-values < 0.001 in all 17 of the spatial domains detected by FAST-P (Supplementary Data 2), including marker genes for breast cancer such as *TACSTD2* and *FOXA1* [35]. The dot plot of average expression levels across all 17 of the detected domains showed that many of the highly expressed genes specific to Domains 1–6 were marker genes for breast cancer (Fig. 3f). Further detailed examination revealed many marker genes for other cell types present in breast cancer, including *MYLK* [42] (Cluster 7, myoepithelial cells), *KRT5, KRT14* [43] (Cluster 8, basal cells), *KRT15* [44] (Cluster 9, luminal progenitor cells), *APOC1* [45] (Cluster 10, macrophages), *FCER1A, MRC1* [46] (Clusters 11 and 12, dendritic cells), *ADH1B* [47], *MMP2* [48] (Clusters 13–16, fibroblasts), and *IL7R, CD3D* [49] (Cluster 17, T cells). Further KEGG enrichment analysis revealed that genes specific to Clusters 1–6 were significantly enriched in many cancer-related pathways (Fig. 3g and Supplementary Fig. S11). For example, Clusters 1–3 were enriched in endocrine resistance, platinum drug resistance, and some cancer-related pathways. These findings provide valuable insights into the molecular mechanisms underlying cancer development and progression, and have important implications for the development of targeted therapies.

To investigate the transcription factors that regulate gene expression, we performed aberrant protein activity analysis for the two sections. As shown in Fig. 3h and Supplementary Fig. S12, the activities of immune-related transcription factors, such as *IL16*, *CD86*, *TNFRSF4*, and *POU2AF1*, in non-tumor regions (Clusters 10–17) were significantly higher than those in tumor regions (Clusters 1–6). Among these transcription factors, the activity levels of *POU2AF1* and *TNFRSF4* were much higher in the T cell cluster (Cluster 17). This observation is consistent with the active transcription of *POU2AF1* [50] and the function of *TNFRSF4* as a tumor necrosis factor [51] in activated T cells. Although *CCNH* was not measured among the 313 gene profiles, we still predicted a higher activity level for *CCNH* in tumor regions (Clusters 1–6). Further deconvolution analysis revealed that Clusters 1–6 were enriched for cancer cells, which was consistent with the H&E images (Fig. 3i&j and Supplementary Fig. S13&14). Moran's I values of the top 15 components from FAST-P were on average 155% higher than those with no spatial consideration in estimating embeddings and 17% higher than those from SpatialPCA-L (Fig. 3k).

Application to the hepatocellular carcinoma Visium dataset

To study mutational patterns in tumor and tumor-adjacent tissues, we also analyzed four sections of the HCC dataset [34] generated using the 10× Visium platform, with two sections from tumors (HCC1 and HCC2) and two from tumor-adjacent tissues (HCC3 and HCC4) collected from a patient with HCC, containing a total of 9813 spatial locations with a median number of 3635 genes per location. Fig. 4a shows a histology image (top panel), accompanied by manual annotations made by a pathologist for tumor/normal epithelium (TNE) and stroma regions, and the spatial heatmap of nine spatial domains detected by FAST-P (bottom panel). By performing integrative clustering via iSC-MEB, we aligned embeddings across multiple sections and allocated domain labels for each location. The aligned embeddings were visualized using two components of tSNE for each method (Fig. 4b; Supplementary Fig. S15). Both FAST and PRECAST achieved better data integration performance, while the computational speed for FAST was much faster than that for PRECAST and SpatialPCA (Fig. 4c; Supplementary Fig. S16). For each method, we summarized the aligned embeddings using three components extracted from UMAP. We then visualized the resulting UMAP components using RGB colors in the RGB plot (Supplementary Fig. S17), with the one for FAST showing clear segregation of the TNE and stroma.

After removing unwanted variations in the expression profiles, we performed DE analysis and detected a total of 2745 DEGs with adjusted p-values < 0.001 in all nine spatial domains detected by FAST-P (Supplementary Data 3). These DEGs included HCC marker genes such as *GPC3* and *CYP2A6* [52, 53]. The dot plot of the average expression levels across all nine detected domains revealed many genes specific to tumor regions in tumor/tumor-adjacent tissues and immune regions (Fig. 4d). In detail, Domains 1–3 comprised tumor regions present only in tumor tissues, among which *CYP2A6* was downregulated, while *THY1* [54], *GPC3*, and *CYP3A7* [55] were upregulated. Domains 4–6 comprised tumor regions present primarily in tumor-adjacent tissues, among which *SCD* [56] and *PCK1* [57] were upregulated. Domains 7-9 were enriched in many immune-related marker genes, including *IGLC1*, *IGHG3*, and *IGKC* [58].

To study the somatic mutation landscapes, we performed location-level mutation detection in the spatial transcriptomics data for all four sections (see "Supplementary Notes"). As shown in Fig. 4e, we detected more somatic mutations in TNE (Domains 1–6) than in stroma (Domains 7–9). In detail, we identified four genes containing the top four SNPs with somatic mutations and visualized their spatial expression across four HCC sections (Supplementary Fig. S18–20), in which *CESR2* and *ETS2* showed higher expression in TNE than in stroma (Fig. 4f). The Moran's I values of the top 15 components of FAST-P were, on average, 19%

higher than those of the other methods (Fig. 4g; Supplementary Fig. S21). In total, FAST-P completed the analysis in 110 seconds and used 0.31 GB memory, while SpatialPCA-F required 3680 seconds and 1.37 GB of memory usage (Fig. 4c).

Application to the mouse embryo Stereo-seq dataset

The Stereo-seq technology was recently developed for high-resolution spatial transcriptomics using 0.22-μm-diameter DNA nanoball (DNB)-patterned arrays [12]. To demonstrate its scalability, we compared the application of FAST with other methods to learn a low-dimensional representation of all spatial locations from a large-scale spatial transcriptomics study on C57BL/6 mouse embryos. In the analysis, we first binned the data as 50bin, with manual annotations based on the expression of marker genes [12]. In total, we analyzed Stereo-seq data from 26 sagittal sections of mouse embryos collected at one-day intervals from E12.5 to E16.5, containing an average of 27,295 genes over a total of 2,323,044 spatial locations. Among all the methods capable of being used for spatial dimension reduction, only FAST was applicable to the analysis of this dataset, while SpatialPCA, PRECAST, and DR-SC were unable to analyze data at this scale. For the analysis of these data, we exclusively utilized FAST-G due to its high computational efficiency compared with the other methods. FAST-G completed the analysis in approximately 2 hours, requiring 94 GB of memory.

First, we showed that FAST achieved the highest adjusted $\mathrm{R}^2_{\mathrm{McF}}$ (Fig. 5a), with a median of 0.79 for FAST-G, 0.71 for PCA and 0.67 for LIGER. To assess the accuracy of clustering, we conducted integrative clustering analysis using iSC-MEB to detect spatial clusters across sections. As shown in Fig. 5b, FAST-G achieved the highest ARI and NMI. FAST-G also provided a more accurate representation of the annotated spatial domains compared to alternative methods (Fig. 5c, left panel). Moreover, our summary of the aligned embeddings using three components from UMAP and their visualization as RGB colors in the RGB plot (Fig. 5c, right panel) further highlighted the superior performance of FAST.

After removing unwanted variations in expression profiles, we detected DEGs and recovered the temporal patterns for all 26 sections across five time points (see "Supplementary Notes"). In the DE analysis, we identified a total of 3663 DEGs with adjusted p-values < 0.001 across the 20 spatial domains identified by FAST-G (Supplementary Data 4). Among these genes, 194 were specific to Domain 1, corresponding to a brain-related subregion. The heatmap illustrates the findings and demonstrates the effective separation of DEGs across various spatial domains (Fig. 5d). By further investigation of mouse embryo spatial gene expression database, EMAGE [59], we detected marker genes for different cell types in mouse embryo, including *Fabp7* [60], *Hes5* [61], *Nefl*, *Nefm* [62] (Domains 1–6, brain), *Ptgds, Slc6a13* [63] (Domain 7, meninges), *Myl2*, *Myl3*, *Myl4*, *Myl7* [64] (Domains 8–10, heart), *Alb*, *Afp* [65] (Domain 11, liver), *Sftpc* [66], *Tcf21* [67] (Domain 12, lung), *Myh3*, *Myh8*, *Myl1* [68] (Domains 13–15, muscle), *Col19a1*, *Col2a1*, *Igf1*, *Sfrp2* (Domains 16–17, trunk somite), *Krt10*, *Krt15* [69] (Domain 18, epidermis/cavity), *Trps1* [70], *Wnt5a* [71] (Domain 19, jaw and tooth), *Krt1* [72] and *Perp* [73] (Domain 20, mucosal epithelium). Furthermore, we generated heatmaps to visualize the expression patterns of marker genes specific to the brain region based on spatial coordinates. These heatmaps confirmed the prominent expression levels within the corresponding brain region (Supplementary Fig. S22).

Next, we detected genes exhibiting temporal expression trends in the brain region (see "Supplementary Notes"; Supplementary Fig. S23). We observed decreased expression of *Hbb-y*, *Hba-x*, and *Hbb-bh1*, which plays a crucial role for cell development by enhancing oxygen transport and aligns with their expression patterns during the early stages of embryonic development [74]. We found an increasing trend in the expression of *Cd44* and *Cdk8*, which encodes a protein that belongs to the cyclin-dependent protein kinase (CDK) family and is known to play a crucial role in regulating the progression of the cell cycle [75].

By the DE genes of the six subclusters within the brain region, we were able to detect three cell types (astrocytes, oligodendrocytes1/2, neurons1/2/3). By inferring the trajectory using embeddings and cluster labels estimated by FAST, we identified the trajectory from glia cells (oligodendrocytes, astrocytes) to neurons (Fig. 5e), which is consistent with the findings in the existing literature [76]. By leveraging temporal information, we successfully detected genes that exhibit specific expression trends within each local region of the brain (Fig. 5f and Supplementary Fig. S24-S28). Notably, some of these genes serve as brain markers, such as *Mt3*, *Dbi* and *Ina*, while others are associated with the development of the nervous system, including *Zic1* and *Hes6*. Among these, we observed that *Mt3* initiates expression during embryo development in brain subcluster 1 at E13.5, suggesting a significant role for *Mt3* in the differentiation process of subcluster 1 (astrocytes; Fig. 5f). These findings provide valuable information on the intricate mechanisms that govern brain development.

Discussion

We present FAST, a probabilistic generalized factor analysis for spatially aware dimension reduction of spatial transcriptomics across multiple sections. Uniquely, FAST takes into consideration the count nature of many existing SRT technologies and models the local spatial correlation structure for "factors" from a common loading matrix during dimension reduction across multiple sections, promoting its computational efficiency and enabling its scalability up to millions of spatial locations. As a result, the "factors" estimated in FAST can be taken as low-dimensional embeddings intrinsic to biological effects with preservation of spatial correlation structure and, thus, can be paired with many existing software/tools developed in scRNA-seq studies to facilitate downstream analyses in SRT studies. Notably, only FAST was found to be applicable to a mouse embryo Stereo-seq dataset across 26 sections varying from E12.5 to E16.5 with >2.3 million spatial locations. Moreover, the analysis was completed in approximately 2 hours, which cannot be achieved by other spatial dimension reduction methods. After obtaining low-dimensional embeddings, we applied iSC-MEB for integrative clustering analysis to align embeddings among multiple sections, followed by a module designed to remove unwanted variations prior to downstream analyses. Using the DLPFC Visium dataset with manual annotation as our benchmark, we have illustrated the benefits of using FAST for data visualization, DE analysis, CCI, and trajectory analyses across multiple sections.

FAST provides a useful tool that can interconnect with aberrant protein activity and somatic mutation analyses to delineate the spatial genomic landscape of cancer. In our analysis of the breast cancer Xenium dataset, FAST not only identified biologically separable clusters for many cell types, such as subtypes for breast cancer cells, fibroblasts, and T cells, but also, through further analysis of aberrant protein activity, showed that these identified clusters presented differential protein activities for immune-related transcription factors. Furthermore, FAST identified a carcinogenesis factor, *CCNH*, which plays an important role in carcinogenesis [77] and was shown to be associated with many signaling pathways in breast cancer [78]. When applied to a HCC Visium dataset, FAST provided estimated embeddings similar to those from PRECAST; however, the computational process was completed in only 142 seconds. Further mutation analysis showed that somatic mutations were more prevalent in TNE than in stroma and identified four genes containing top SNPs with somatic mutations. Among these genes, *Cers2* plays a vital role in preserving hepatic chromosome polyploidization during cell division by regulating the expression of Mad2 in mouse [79], and is closely associated with the progression of liver cancer [80]. Takeda et al. [81] reported that *ETS2* functions as a key downstream transcription factor and assembles a transcription complex with *MLL* in HCC cells that directly activates *MMP1* and *MMP3*.

FAST has some caveats that may require further exploration. First, it would be interesting to perform joint dimension reduction for both single-cell multimodal omic data and SRT data with single-cell resolution, thus promoting inference for cis-regulatory interactions and/or defining gene-regulatory networks for transcription factors. Second, FAST performs spatial dimension reduction in an unsupervised manner. When manual annotations are available for some sections, it may be preferable to adopt a semi-supervised method by utilizing partial information about annotation labels across all sections. Finally, in the era of deep learning, performing a deep spatial encoding for SRT datasets would help extract a low-dimensional representation capturing nonlinear biological effects. In this study, we demonstrate the potential of FAST as a prototype for scalable extraction of cross-section embeddings among a large variety of spatial locations. Future exploration of all these issues is warranted to further confirm the important value of FAST as a method for spatially aware dimension reduction.

Methods

FAST model

Here, we present a basic overview of FAST; further details are available in the Supplementary Notes. For the M SRT sections, we observed a count expression matrix for each section. FAST has the ability to model these count matrices or log-normalized matrices of gene expressions using either a log link or identity link for multiple sections (Fig. 1a). We primarily focus on introducing the Poisson version of FAST (FAST-P) for count matrices, while details about the Gaussian version (FAST-G) can be found in the Supplementary Notes.

Specifically, we observe an $n_m \times p$ count expression matrix $\mathbf{X}_m = (\mathbf{x}_{m1}, \cdots, \mathbf{x}_{mi}, \cdots, \mathbf{x}_{mn_m})^{\mathrm{T}}$ for section m $(= 1, \cdots, M)$, where $\mathbf{x}_{mi} = (x_{mi1}, \cdots, x_{mip})^{\mathrm{T}}$ is a p-dimensional expression vector for each location $s_{mi} \in \mathbb{R}^2$ in the section m on square or hexagonal lattices, among others. FAST models the count expression level of gene j, x_{mij}, with its latent low-dimensional features $\mathbf{v}_{mi}$ via a log link as

$$x_{mij} | f_{mij} \sim Poisson(f_{mij}), \tag{1}$$

$$\ln f_{mij} = \nu_{mj} + \mathbf{w}_j^{\mathrm{T}} \mathbf{v}_{mi} + \varepsilon_{mij}, \quad \varepsilon_{mij} \sim N(0, \lambda_{mj}), \tag{2}$$

where f_{mij} is an unknown Poisson rate that represents the underlying gene-expression level,

$\nu_m = (\nu_{m1}, \cdots, \nu_{mp}) \in \mathbb{R}^p$ is a sample-specified intercept term, $W = (\mathbf{w}_1, \cdots, \mathbf{w}_p)^{\mathrm{T}} \in \mathbb{R}^{p\times q}$ is a sample-shared loading matrix that captures the shared information of sections, $\mathbf{v}_{mi}$ is a q-dimensional factors that captures the biological information, and $\boldsymbol{\epsilon}_{mi} = (\varepsilon_{mi1}, \cdots, \varepsilon_{mip})^{\mathrm{T}} \sim N(\mathbf{0}, \Lambda_m)$ is an error term that captures the overdispersion, and $\Lambda_m = \mathrm{diag}(\lambda_{m1}, \cdots, \lambda_{mp})$ is the covariance matrix of the error term. To account for the spatial dependence among spatial locations within each section, we adopted a continuous multivariate Hidden Markov Random Field (HMRF) model for the factor $\mathbf{v}_{mi}$, which captures spatial dependencies in the embedding space, referred to as a spatial factor. Specifically, we assumed an intrinsic CAR model [82] for $\mathbf{v}_{mi}$

$$\mathbf{v}_{mi}|\mathbf{v}_{[n_m]\setminus i} \sim N(\mu_{v_{mi}}, L_{mi}^{-1}\Psi_m), \tag{3}$$

where subscript $[n_m] \setminus i$ denotes all locations but s_{mi} in the section m, L_{mi} is the number of neighbors of location s_{mi} in section m, $\mu_{v_{mi}} = L_{mi}^{-1}\sum_{i'\in N_{mi}} \mathbf{v}_{mi'}$ is the conditional mean relevant to the neighbors of the location s_{mi}, and Ψ_m is a $q \times q$ conditional covariance matrix for the elements of $\mathbf{v}_{ri}$. The intrinsic CAR model (3) models the local dependence of the spatial factor $\mathbf{v}_{mi}$ for the location s_{mi} in section m via the factors of its neighboring locations (Fig. 1a).

Embedding alignment and spatial clustering

Once we obtained the uncorrected embeddings of FAST or other compared methods, we applied iSC-MEB [83] for joint embedding alignment and spatial clustering. However, before running iSC-MEB, we needed to determine the optimal number of clusters. Therefore, we first utilized Harmony to obtain batch corrected embeddings, followed by Louvain clustering to select the appropriate number of clusters. After determining the optimal number of clusters, we ran iSC-MEB; the implementation details can be found in the Supplementary Notes. After obtaining aligned embeddings and clusters, we compared the performance of FAST and other methods in visualization of the aligned embeddings, batch removal of the aligned embeddings, and spatial clustering.

Unwanted variation removal for gene expressions

To remove unwanted variations in gene expression from the multiple sections, we applied iSC-MEB [83] to $\widehat{V}$, where $\widehat{V}$ is the combined unaligned embeddings obtained by FAST. By doing so, we obtained the posterior probabilities of the cluster labels, $\widehat{\mathbf{r}}_{mi}$. To further address unwanted variations, we used a set of housekeeping genes, which are not influenced by other biological factors [84], as negative control genes; for more information on the selection of housekeeping genes, please refer to the Supplementary Notes.

Using the selected L housekeeping genes, we performed PCA and obtained the top 10 principal components, $\widehat{\mathbf{h}}_{mi}$, which we used as covariates to adjust for unwanted variation. Let $\tilde{x}_{mij}$ be the log-normalized expression value for gene j of spot i in section m. Then, we applied a spatial linear model to the normalized expression level of gene j,

$$\tilde{x}_{mij} = \widehat{\mathbf{r}}_{mi}^{\mathrm{T}}\boldsymbol{\alpha}_j + \widehat{\mathbf{h}}_{mi}^{\mathrm{T}}\boldsymbol{\gamma}_j + t_m\zeta_j + u_{mij} + \epsilon_{mij}, \tag{4}$$

where $\boldsymbol{\alpha}_j$ is a K-dimensional vector for biological effects between cell/domain types, $\boldsymbol{\gamma}_j$ is a 10-dimensional vector of regression coefficients associated with the unwanted factors, and ζ_j is the regression coefficient denoting the temporal effect if the temporal information is available (e.g., in the embryo dataset), u_{mij} follows a univariate intrinsic CAR model for retaining the spatial dependence of locations in the same section such that $u_{mij}|\mathbf{u}_{[n_m]\backslash i,j} \sim N(\mu_{u_{mij}}, L_{mi}^{-1}\phi_{mj})$, and $\epsilon_{mij} \sim N(0, \sigma_{mj}^2)$ is an i.i.d (with respect to i) error term.

An ICM-EM algorithm was designed to solve the model parameters; details are provided in the Supplementary Notes. After obtaining the parameter estimates in Eqn. (4), users can remove batch effects from the original normalized gene expression using

$$\widehat{\tilde{x}}_{mij} = \tilde{x}_{mij} - \widehat{\mathbf{h}}_{mi}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_j.$$

This strategy can also be applied to multiple sections from multiple biological conditions by adding additional covariates in the linear model when such information is available.

Other analyses

After removing unwanted variation in the gene expression matrices of multiple sections, we performed DE analysis for the combined sections. To present biological discovery in different tissues, we applied various analyses including KEGG pathway analysis, CCI analysis, transcription factor analysis, cell-type deconvolution analysis, somatic mutation analysis, and analysis to detect genes with temporal trend of expressions. Details of these analyses are provided in the Supplementary Notes.

Comparison of methods

Through simulations and with real-world data, we conducted a comprehensive comparison of FAST with existing methods for dimension reduction and spatial clustering.

To evaluate the biological embedding estimation performance of FAST, we used a range of spatially aware and non-spatially aware dimension reduction methods as benchmarks. Specifically, we compared FAST with (1) SpatialPCA [17] implemented in the R package *SpatialPCA* (version 1.2.0); (2) PRECAST [28] implemented in the R package *PRECAST* (version 1.6.1); (3) DR-SC [29] implemented in the R package *DR.SC* (version 3.2); (4) scVI [30] implemented in the Python module *scvi-tools* (version 0.20.3); (5) PCA; (6) multiBatchPCA [31] implemented in the R package *batchelor* (version 1.10.0); (7) NMF implemented in the R package *scater* (version 1.25.1); and (8) LIGER [32] implemented in the R package *rliger* (version 1.0.0). The first two methods are for spatial dimension reduction, whereas the remaining methods are for dimension reduction without consideration of spatial information. SpatialPCA has both full-rank and low-rank versions, denoted as SpatialPCA-F and SpatialPCA-L, respectively. We extracted 15-dimensional embeddings for all datasets and methods for comparison. See the Supplementary Notes for the details of each method compared.

To compare the clustering performance, after obtaining the low-dimensional embeddings for all methods except for LIGER, we applied iSC-MEB to conduct the joint embedding alignment and spatial clustering. Since LIGER had already addressed batch effects in its embeddings, we did not compare the spatial clustering performance using iSC-MEB.

Evaluation metrics

McFadden's adjusted R^2. For DLPFC Visium and Embryo Stereo-seq datasets that have manual domain annotations serving as ground truth, we evaluated the performance of different methods in dimension reduction by measuring the association between the embeddings and ground truth for each section using a multinomial regression model with the ground truth as the response variable and the extracted embeddings as covariates for each slice. For the fitted model, we then calculated the adjusted $\mathrm{R}^2_{\mathrm{McF}}$ [85], which provides a measure of the amount of biological information contained in the extracted embeddings, and a higher value indicates better performance in dimension reduction.

ARI and NMI. For the DLPFC Visium and Embryo Steore-seq datasets, we evaluated the performance of different methods in spatial domain clustering by comparing the detected spatial domains with the ground truth. For this purpose, we employed standard clustering evaluation metrics, including ARI [86] and NMI [87].

Moran's I As a reflection of spatial autocorrelation, Moran's I was used to measure the spatial information contained in the embeddings of each section obtained by FAST and other methods. It is defined as

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})},$$

where n is the number of spatial units indexed by i and j, x is the variable of interest, $\bar{x}$ is the mean of x, and w_{ij} is the (i, j)-th element of a matrix of spatial weights with zero on the diagonal. During the evaluation, we assigned a value of 1 to w_{ij} if the spot i is a neighbor of the spot j, and otherwise w_{ij} was set to 0.

Simulations

We simulated three sections based on a single section of the human DLPFC Visium dataset (sample ID: 151672) with eight spatial domains [33], including layers 1-6, WM, and unknown cells (Unknown). Using these data, we initially established the spatial coordinates and spatial domains for the three sections. The first section was set using the same coordinates as the panel data. For the second section, the coordinates were selected such that the x coordinates were below the 90th percentile. For the last section, the coordinates were chosen such that the y coordinates were below the 90th percentile. The spatial domains for all three sections remained the same as the panel data; see Supplementary Fig. S1a.

Subsequently, we used the R package *splatter* (version 1.18.2) [88] to simulate gene expressions at predefined spatial coordinates of each section. This package generates gene expression data based on a gamma-Poisson distribution. We explored four different scenarios by adjusting the parameters of both batch effects and biological effects to either low or high values (batch effect=low or high; biological effect=low or high). The *splatter* package offers two parameters, namely `de_facScale` and `de_prop`, that enable the control of batch effects and biological effects. To generate the four scenarios, we set `de_facScale=0.2` and `de_facScale=0.6` for batch effect=low and high, respectively, and `de_prop=0.15` and `de_prop=0.3` for biological effect=low and high, respectively. The four scenarios tested were as follows: Scenario 1 (batch effect=low, biological effect=high), Scenario 2 (batch effect=low, biological effect=low), Scenario 3 (batch effect=high, biological effect=high), and Scenario 4 (batch effect=high, biological effect=low).

Gene selection for multi-section analysis

During the quality control (QC) process, we applied filtering criteria to eliminate genes with zero expression in multiple locations and locations with zero expression for numerous genes (see "Data resources"). For our analyses, we used the `FindVariableFeatures` function of the *Seurat* (version 4.1.1) R package, with the default settings, to identify highly variable genes (HVGs) in all four datasets. Specifically, we selected the top 2000 HVGs for each section. We then prioritized genes based on the frequency of their selection as HVGs across all sections and selected the top 2000 genes based on this criterion. These 2000 genes were used as input for the comparison of FAST with other analytical methods.

Data resources

Human dorsolateral prefrontal cortex Visium dataset

We obtained spatial transcriptomic data for human DLPFC from the 10× Visium platform; data were downloaded from `https://doi.org/10.5281/zenodo.4730634`. The dataset comprised 12 postmortem DLPFC tissue sections from three independent neurotypical adult donors, and the raw expression count matrix contained 33,538 genes for each section, with a total of 47,681 spatial locations. To ensure data quality, we performed QC on each section, filtering genes with non-zero expression levels for $\leq$ 20 locations and locations with non-zero expression levels for $\leq$ 20 genes. This filtering step resulted in a set of 14,535 genes on average across a total of 47,680 spatial locations. The spatial domains annotated in all 12 sections were layer 1 (n=5321), layer 2 (n=2858), layer 3 (n=17,587), layer 4 (n=3547), layer 5 (n=7300), layer 6 (n=6201), WM (n=4514), and undetermined locations (n = 352), according to the cytoarchitecture of the original study [33]. In our analysis, we considered these manual annotations as the ground truth for evaluating the dimension reduction and clustering performance of the different methods.

Human breast cancer Xenium dataset

We collected a dataset from two tissue sections of a patient with breast cancer measured by Xenium In Situ technology [35]; data were downloaded from `https://www.dropbox.com/s/t05w7ccufh1v0h8/xenium_prerelease_jul12_hBreast_replicates.tar?dl=0`. The dataset consisted of 313 genes of interest, with 35,868 and 36,783 spatial locations from the two tissue sections, respectively. These target genes were selected and curated primarily based on single cell atlas data for human breast tissue [35]. To ensure data quality, we performed a QC step in which we filtered locations with non-zero expression levels for ≤ 15 genes. As a result, 35,015 and 35,932 locations were retained in the two tissue sections, respectively. Aberrant protein activity was inferred by VIPER to assess protein activity from gene expression data and the regulatory network. The spatial deconvolution was then performed to examine the spatial distribution of the compositions of different cells using RCTD.

Human hepatocellular carcinoma Visium dataset

The HCC dataset used in this study was derived from two tissue sections, one from the tumor and the other from the tumor-adjacent regions of a patient with hepatocellular carcinoma [34]. The dataset consisted of 36,601 genes from more than 9813 spatial locations. During the QC process, genes with non-zero expression levels for ≤ 20 locations and locations with non-zero expression levels for ≤ 20 genes were removed, resulting in a set of 14,851 genes on average from a total of 9813 locations. To identify the TNE and stroma regions, manual annotations were provided by a pathologist using the Visium companion H&E images. Somatic mutation analysis was performed to investigate differences in somatic mutations between the TNE and stromal regions.

Mouse embryo Stereo-seq dataset

We collected data for 26 mouse embryo Stereo-seq sections from `https://db.cngb.org/stomics/mosta/`. The dataset comprised 27,295 genes on average, with 2,323,044 spatial locations recorded across the 26 sections, from embryo days E12.5 to E16.5. During the QC process, we initially eliminated genes with non-zero expression levels for ≤ 20 locations, as well as locations with non-zero expression levels for ≤ 20 genes. As a result, we were left with 14,307 genes on average, totaling 2,318,423 locations.

Data availability

The four datasets used in this study are publicly available. These include the 12 human DLPFC Visium datasets (`https://doi.org/10.5281/zenodo.4730634`), two human breast cancer Xenium datasets (`https://www.dropbox.com/s/t05w7ccufh1v0h8/xenium_prerelease_jul12_hBreast_replicates.tar?dl=0`), four human HCC Visium datasets (Raw FASTQ data are available at `https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=858545`, and H&E images are available at `https://doi.org/10.6084/m9.figshare.`

`21280569.v1` and `https://doi.org/10.6084/m9.figshare.21061990.v1`), and 26 mouse embryo Stereo-seq datasets (`https://db.cngb.org/stomics/mosta/`).

Code availability

The FAST methods were implemented in an open-source R package that is publicly available at `https://github.com/feiyoung/FAST`. The code to reproduce the analysis can be found at `https://github.com/feiyoung/FAST_Analysis`.

Author contributions

J.L. initiated and designed the study, W.L. implemented the model and developed the software tool; W.L. and X.Z. performed the simulation studies and the benchmark evaluation; J.L. wrote the manuscript, and W.L., X.Z., X.C., Z.F., H.L., J.C., L.S., T.Y., J.Y. and J.L. edited and revised the manuscript.

Competing Interests

The authors declare no competing interests.

References

[1] Robert Foreman and Roy Wollman. Mammalian gene expression variability is explained by underlying cell state. *Molecular systems biology*, 16(2):e9146, 2020.

[2] Andrew L Ji, Adam J Rubin, Kim Thrane, Sizun Jiang, David L Reynolds, Robin M Meyers, Margaret G Guo, Benson M George, Annelie Mollbrink, Joseph Bergenstr**r**ahle, et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell*, 182(2):497–514, 2020.

[3] Chang Liu, Rui Li, Young Li, Xiumei Lin, Kaichen Zhao, Qun Liu, Shuowen Wang, Xueqian Yang, Xuyang Shi, Yuting Ma, et al. Spatiotemporal mapping of gene expression landscapes and developmental trajectories during zebrafish embryogenesis. *Developmental Cell*, 57(10):1284–1298, 2022.

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

[5] Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. *Nature Methods*, 19 (5):534–546, 2022.

[6] 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): aaa6090, 2015.

[7] 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.

[8] 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.

[9] Je Hyuk Lee, Evan R Daugharthy, Jonathan Scheiman, Reza Kalhor, Joyce L Yang, Thomas C Ferrante, Richard Terry, Sauveur SF Jeanty, Chao Li, Ryoji Amamoto, et al. Highly multiplexed subcellular rna sequencing in situ. *science*, 343(6177):1360–1363, 2014.

[10] 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.

[11] 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.

[12] Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, Jin Yang, Jiangshan Xu, Shijie Hao, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell*, 185(10):1777–1792, 2022.

[13] Michaela Asp, Joseph Bergenstr**r**ahle, and Joakim Lundeberg. Spatially resolved transcriptomes—next generation tools for tissue exploration. *BioEssays*, 42(10):1900221, 2020.

[14] Giovanni Palla, David S Fischer, Aviv Regev, and Fabian J Theis. Spatial components of molecular tissue biology. *Nature Biotechnology*, 40(3):308–318, 2022.

[15] F William Townes, Stephanie C Hicks, Martin J Aryee, and Rafael A Irizarry. Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome biology*, 20:1–16, 2019.

[16] Naomi Altman and Martin Krzywinski. The curse (s) of dimensionality. *Nat Methods*, 15 (6):399–400, 2018.

[17] Lulu Shang and Xiang Zhou. Spatially aware dimension reduction for spatial transcriptomics. *Nature Communications*, 13(1):7203, 2022.

[18] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[19] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.

[20] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):14049, 2017.

[21] 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*, 39 (11):1375–1384, 2021.

[22] 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.

[23] 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.

[24] F William Townes and Barbara E Engelhardt. Nonnegative spatial factorization applied to spatial genomics. *Nature Methods*, 20(2):229–238, 2023.

[25] Felix Leibfried, Vincent Dutordoir, ST John, and Nicolas Durrande. A tutorial on sparse gaussian processes and variational inference. *arXiv preprint arXiv:2012.13962*, 2020.

[26] Michaela Asp, Stefania Giacomello, Ludvig Larsson, Chenglin Wu, Daniel Fürth, Xiaoyan Qian, Eva Wärdell, Joaquin Custodio, Johan Reimegrard, Fredrik Salmén, et al. A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. *Cell*, 179(7):1647–1660, 2019.

[27] 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*, 25(4):484–492, 2022.

[28] Wei Liu, Xu Liao, Ziye Luo, Yi Yang, Mai Chan Lau, Yuling Jiao, Xingjie Shi, Weiwei Zhai, Hongkai Ji, Joe Yeong, et al. Probabilistic embedding, clustering, and alignment for integrating spatial transcriptomics data with precast. *Nature Communications*, 14(1):296, 2023.

[29] 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*, 50(12):e72–e72, 2022.

[30] 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.

[31] 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.

[32] Joshua D Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and Evan Z Macosko. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177(7):1873–1887, 2019.

[33] 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.

[34] Mai Chan Lau, Yang Yi, Denise Goh, Chun Chau Lawrence Cheung, Benedict Tan, Jeffrey Chun Tatt Lim, Craig Ryan Joseph, Felicia Wee, Justina Nadia Lee, Xinru Lim, et al. Case report: Understanding the impact of persistent tissue-localization of sars-cov-2 on immune response activity via spatial transcriptomic analysis of two cancer patients with covid-19 co-morbidity. *Frontiers in Immunology*, 13:5376, 2022.

[35] Amanda Janesick, Robert Shelansky, Andrew D Gottscho, Florian Wagner, Morgane Rouault, Ghezal Beliakoff, Michelli Faria de Oliveira, Andrew Kohlway, Jawad Abousoud, Carolyn A Morrison, et al. High resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and in situ analysis of ffpe tissue. *bioRxiv*, pages 2022–10, 2022.

[36] Hongkui Zeng, Elaine H Shen, John G Hohmann, Seung Wook Oh, Amy Bernard, Joshua J Royall, Katie J Glattfelder, Susan M Sunkin, John A Morris, Angela L Guillozet-Bongaarts, et al. Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell*, 149(2):483–496, 2012.

[37] Shuo Chen, Yuzhou Chang, Liangping Li, Diana Acosta, Yang Li, Qi Guo, Cankun Wang, Emir Turkes, Cody Morrison, Dominic Julian, et al. Spatially resolved transcriptomics reveals genes associated with the vulnerability of middle temporal gyrus in alzheimer's disease. *Acta Neuropathologica Communications*, 10(1):1–24, 2022.

[38] Suoqin Jin, Christian F Guerrero-Juarez, Lihua Zhang, Ivan Chang, Raul Ramos, Chen-Hsiang Kuan, Peggy Myung, Maksim V Plikus, and Qing Nie. Inference and analysis of cell-cell communication using cellchat. *Nature communications*, 12(1):1088, 2021.

[39] Huan Liu, Yuanyuan Yang, Yuguo Xia, Wen Zhu, Rehana K Leak, Zhishuo Wei, Jianyi Wang, and Xiaoming Hu. Aging of cerebral white matter. *Ageing research reviews*, 34: 64–76, 2017.

[40] Andrés Buonanno. The neuregulin signaling pathway and schizophrenia: from genes to synapses and neural circuits. *Brain research bulletin*, 83(3-4):122–131, 2010.

[41] F Alexander Wolf, Fiona K Hamey, Mireya Plass, Jordi Solana, Joakim S Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J Theis. Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology*, 20:1–9, 2019.

[42] Chris Jones, Alan Mackay, Anita Grigoriadis, Antonio Cossu, Jorge S Reis-Filho, Laura Fulford, Tim Dexter, Susan Davies, Karen Bulmer, Emily Ford, et al. Expression profiling of purified normal human luminal and myoepithelial breast cells: identification of novel prognostic markers for breast cancer. *Cancer research*, 64(9):3037–3045, 2004.

[43] Hirotoshi Soyama, Miki Nishio, Junji Otani, Toshiko Sakuma, Shintaro Takao, Shigeo Hara, Takaaki Masuda, Koshi Mimori, Shinya Toyokuni, John P Lydon, et al. Hippo-taz signaling is the master regulator of the onset of triple-negative basal-like breast cancers. *Proceedings of the National Academy of Sciences*, 119(29):e2123134119, 2022.

[44] Katharina Theresa Kohler, Nadine Goldhammer, Samuel Demharter, Ulrich Pfisterer, Konstantin Khodosevich, Lone Rønnov-Jessen, Ole William Petersen, René Villadsen, and Jiyoung Kim. Ductal keratin 15+ luminal progenitors in normal breast exhibit a basal-like breast cancer transcriptomic signature. *NPJ Breast Cancer*, 8(1):81, 2022.

[45] Liwen Ren, Jie Yi, Yihui Yang, Wan Li, Xiangjin Zheng, Jinyi Liu, Sha Li, Hong Yang, Yizhi Zhang, Binbin Ge, et al. Systematic pan-cancer analysis identifies apoc1 as an immunological biomarker which regulates macrophage polarization and promotes tumor metastasis. *Pharmacological Research*, 183:106376, 2022.

[46] Rongqun Guo, Mengdie Lü, Fujiao Cao, Guanghua Wu, Fengcai Gao, Haili Pang, Yadan Li, Yinyin Zhang, Haizhou Xing, Chunyan Liang, et al. Single-cell map of diverse immune phenotypes in the acute myeloid leukemia microenvironment. *Biomarker Research*, 9(1): 1–16, 2021.

[47] John A Grout, Philemon Sirven, Andrew M Leader, Shrisha Maskey, Eglantine Hector, Isabelle Puisieux, Fiona Steffan, Evan Cheng, Navpreet Tung, Mathieu Maurin, et al. Spatial positioning and matrix programs of cancer-associated fibroblasts promote t-cell exclusion in human lung tumors. *Cancer Discovery*, 12(11):2606–2625, 2022.

[48] Liuyang Zhao, Yan Sun, Yixuan Hou, Qiongle Peng, Liyang Wang, Haojun Luo, Xi Tang, Zongyue Zeng, and Manran Liu. Mirna expression analysis of cancer-associated fibroblasts and normal fibroblasts in breast cancer. *The international journal of biochemistry & cell biology*, 44(11):2051–2059, 2012.

[49] Hee Jin Lee, Jeong-Ju Lee, In Hye Song, In Ah Park, Jun Kang, Jong Han Yu, Jin-Hee Ahn, and Gyungyub Gong. Prognostic and predictive value of nanostring-based immune-related gene signatures in a neoadjuvant setting of triple-negative breast cancer: relationship to tumor-infiltrating lymphocytes. *Breast cancer research and treatment*, 151:619–627, 2015.

[50] Félix Lombard-Vadnais, Julie Lacombe, Geneviève Chabot-Roy, Mathieu Ferron, and Sylvie Lesage. Oca-b does not act as a transcriptional coactivator in t cells. *Immunology and Cell Biology*, 100(5):338–351, 2022.

[51] Lindsay K Ward-Kavanagh, Wai Wai Lin, John R **v**Sedỳ, and Carl F Ware. The tnf receptor superfamily in co-stimulating and co-inhibitory responses. *Immunity*, 44(5): 1005–1019, 2016.

[52] Yukihiro Haruyama and Hiroaki Kataoka. Glypican-3 is a prognostic factor and an immunotherapeutic target in hepatocellular carcinoma. *World journal of gastroenterology*, 22(1):275, 2016.

[53] Xiaojing Ren, Yuanyuan Ji, Xuhua Jiang, and Xun Qi. Downregulation of cyp2a6 and cyp2c8 in tumor tissues is linked to worse overall survival and recurrence-free survival from hepatocellular carcinoma. *BioMed research international*, 2018, 2018.

[54] Xin Chen, John Higgins, Siu-Tim Cheung, Rui Li, Veronica Mason, Kelli Montgomery, Sheung-Tat Fan, Matt van de Rijn, and Samuel So. Novel endothelial cell markers in hepatocellular carcinoma. *Modern Pathology*, 17(10):1198–1210, 2004.

[55] Nobuo Kondoh, Toru Wakatsuki, Akihide Ryo, Akiyuki Hada, Tsukasa Aihara, Sankichi Horiuchi, Narihide Goseki, Osamu Matsubara, Kenji Takenaka, Mizue Shichita, et al. Identification and characterization of genes associated with human hepatocellular carcinogenesis. *Cancer research*, 59(19):4990–4996, 1999.

[56] Wei Dong, Zhi-hui Dai, Fu-chen Liu, Xing-gang Guo, Chun-mei Ge, Jin Ding, Hui Liu, and Fu Yang. The rna-binding protein rbm3 promotes cell proliferation in hepatocellular carcinoma by regulating circular rna scd-circrna 2 production. *EBioMedicine*, 45:155–167, 2019.

[57] Meng-Xi Liu, Lei Jin, Si-Jia Sun, Peng Liu, Xu Feng, Zhou-Li Cheng, Wei-Ren Liu, Kun-Liang Guan, Ying-Hong Shi, Hai-Xin Yuan, et al. Metabolic reprogramming by pck1 promotes tca cataplerosis, oxidative stress and apoptosis in liver cancer cells and suppresses hepatocellular carcinoma. *Oncogene*, 37(12):1637–1653, 2018.

[58] Ajit J Nirmal, Tim Regan, Barbara B Shih, David A Hume, Andrew H Sims, and Tom C Freeman. Immune cell gene signatures for profiling the microenvironment of solid tumorsimmune cell gene signatures for profiling solid tumors. *Cancer immunology research*, 6(11):1388–1400, 2018.

[59] Lorna Richardson, Shanmugasundaram Venkataraman, Peter Stevenson, Yiya Yang, Julie Moss, Liz Graham, Nicholas Burton, Bill Hill, Jianguo Rao, Richard A Baldock, et al. Emage mouse embryo spatial gene expression database: 2014 update. *Nucleic acids research*, 42(D1):D835–D844, 2014.

[60] Yoko Arai, Nobuo Funatsu, Keiko Numayama-Tsuruta, Tadashi Nomura, Shun Nakamura, and Noriko Osumi. Role of fabp7, a downstream gene of pax6, in the maintenance of neuroepithelial cells during early embryonic development of the rat cortex. *Journal of Neuroscience*, 25(42):9752–9761, 2005.

[61] Masuko Katoh and Masaru Katoh. Integrative genomic analyses on hes/hey family: Notch-independent hes1, hes3 transcription in undifferentiated es cells, and notch-dependent hes1, hes5, hey1, hey2, heyl transcription in fetal tissues, adult tissues, or cancer. *International journal of oncology*, 31(2):461–466, 2007.

[62] Ze-Kai Cui, Shen-Yang Li, Kai Liao, Zhi-Jie Wang, Yong-Long Guo, Luo-Sheng Tang, Shi-Bo Tang, Jacey Hongjie Ma, and Jian-Su Chen. Characteristics of neural growth and cryopreservation of the dorsal root ganglion using three-dimensional collagen hydrogel culture versus conventional culture. *Neural Regeneration Research*, 16(9):1856, 2021.

[63] Adrian M Stankiewicz, Aneta Jaszczyk, Joanna Goscik, and Grzegorz R Juszczak. Stress and the brain transcriptome: Identifying commonalities and clusters in standardized data from published experiments. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 119:110558, 2022.

[64] Nobutoshi Ichise, Tatsuya Sato, Hiroyori Fusagawa, Hiroya Yamazaki, Taiki Kudo, Izaya Ogon, and Noritsugu Tohse. Ultrastructural assessment and proteomic analysis in myofibrillogenesis in the heart primordium after heartbeat initiation in rats. *Frontiers in Physiology*, 13, 2022.

[65] Vincent W Keng, Hideshi Yagi, Masahito Ikawa, Takashi Nagano, Zaw Myint, Kazuya Yamada, Takashi Tanaka, Ayuko Sato, Ikunobu Muramatsu, Masaru Okabe, et al. Homeobox gene hex is essential for onset of mouse embryonic liver development and differentiation

of the monocyte lineage. *Biochemical and biophysical research communications*, 276(3): 1155–1161, 2000.

[66] Pimchanok Pimton, Shimon Lecht, Collin T Stabler, Gregg Johannes, Edward S Schulman, and Peter I Lelkes. Hypoxia enhances differentiation of mouse embryonic stem cells into definitive endoderm and distal lung cells. *Stem cells and development*, 24(5):663–676, 2015.

[67] Juwon Park, Malina J Ivey, Yanik Deana, Kara L Riggsbee, Emelie Sörensen, Veronika Schwabl, Caroline Sjöberg, Tilda Hjertberg, Ga Young Park, Jessica M Swonger, et al. The tcf21 lineage constitutes the lung lipofibroblast population. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, 316(5):L872–L885, 2019.

[68] Ruth Amann, Stefan Wyder, Anne M Slavotinek, and Beat Trueb. The fgfrl1 receptor is required for development of slow muscle fibers. *Developmental biology*, 394(2):228–241, 2014.

[69] Mathilde Couteaudier, Laëtitia Trapp-Fragnet, Nicolas Auger, Katia Courvoisier, Bertrand Pain, Caroline Denesvre, and Jean-Fran**c**cois Vautherot. Derivation of keratinocytes from chicken embryonic stem cells: establishment and characterization of differentiated proliferative cell populations. *Stem Cell Research*, 14(2):224–237, 2015.

[70] Ikumi Michikami, Toshiya Fukushi, Shiho Honma, Seisuke Yoshioka, Shunji Itoh, Yasuteru Muragaki, Kojiro Kurisu, Takashi Ooshima, Satoshi Wakisaka, and Makoto Abe. Trps1 is necessary for normal temporomandibular joint development. *Cell and tissue research*, 348: 131–140, 2012.

[71] Minkui Lin, Lu Li, Chao Liu, Hongbing Liu, Fenglei He, Fuhua Yan, Yanding Zhang, and YiPing Chen. Wnt5a regulates growth, patterning, and odontoblast differentiation of developing mouse tooth. *Developmental Dynamics*, 240(2):432–440, 2011.

[72] Taishi Komori, Mitsuaki Ono, Emilio Satoshi Hara, Junji Ueda, Ha Thi Thu Nguyen, Ha Thi Nguyen, Tomoko Yonezawa, Takahiro Maeba, Aya Kimura-Ono, Takeshi Takarada, et al. Type iv collagen α6 chain is a regulator of keratin 10 in keratinization of oral mucosal epithelium. *Scientific reports*, 8(1):2612, 2018.

[73] Rebecca A Ihrie, Michelle R Marques, Bichchau T Nguyen, Jennifer S Horner, Cristian Papazoglu, Roderick T Bronson, Alea A Mills, and Laura D Attardi. Perp is a p63-regulated gene essential for epithelial integrity. *Cell*, 120(6):843–856, 2005.

[74] Andrew J King, Duantida Songdej, Damien J Downes, Robert A Beagrie, Siyu Liu, Megan Buckley, Peng Hua, Maria C Suciu, A Marieke Oudelaar, Lars LP Hanssen, et al. Reactivation of a developmentally silenced embryonic globin gene. *Nature communications*, 12(1):4439, 2021.

[75] Zsolt Szilagyi and Claes M Gustafsson. Emerging roles of cdk8 in cell cycle control. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1829(9):916–920, 2013.

[76] Sanjay R Srivatsan, Mary C Regier, Eliza Barkan, Jennifer M Franks, Jonathan S Packer, Parker Grosjean, Madeleine Duran, Sarah Saxton, Jon J Ladd, Malte Spielmann, et al. Embryo-scale, single-cell spatial transcriptomics. *Science*, 373(6550):111–117, 2021.

[77] Yuchan Wang, Fang Liu, Feng Mao, Qinlei Hang, Xiaodong Huang, Song He, Yingying Wang, Chun Cheng, Huijie Wang, Guangfei Xu, et al. Interaction with cyclin h/cyclin-dependent kinase 7 (ccnh/cdk7) stabilizes c-terminal binding protein 2 (ctbp2) and promotes cancer cell migration. *Journal of Biological Chemistry*, 288(13):9028–9034, 2013.

[78] Ying-Nan Yu, George Wai-Cheong Yip, Puay-Hoon Tan, Aye Aye Thike, Ken Matsumoto, Masafumi Tsujimoto, and Boon-Huat Bay. Y-box binding protein 1 is up-regulated in proliferative breast cancer and its inhibition deregulates the cell cycle. *International journal of oncology*, 37(2):483–492, 2010.

[79] Mingjun Cao, Shaohua Zhang, Sin Man Lam, and Guanghou Shui. Hepatic loss of cers2 induces cell division defects via a mad2-mediated pathway. *Clinical and Translational Medicine*, 12(1), 2022.

[80] Mengmeng Zhang, Zhangyun Li, Yuwei Liu, Xiao Ding, Yanyan Wang, and Shaohua Fan. The ceramide synthase (cers/lass) family: Functions involved in cancer progression. *Cellular Oncology*, pages 1–21, 2023.

[81] Shugaku Takeda, Han Liu, Satoru Sasagawa, Yiyu Dong, Paul A Trainor, Emily H Cheng, James J Hsieh, et al. Hgf-met signals via the mll-ets2 complex in hepatocellular carcinoma. *The Journal of clinical investigation*, 123(7):3154–3165, 2013.

[82] 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.

[83] Xiao Zhang, Wei Liu, Fangda Song, and Jin Liu. isc. meb: an r package for multi-sample spatial clustering analysis of spatial transcriptomics data. *Bioinformatics Advances*, 3(1): vbad019, 2023.

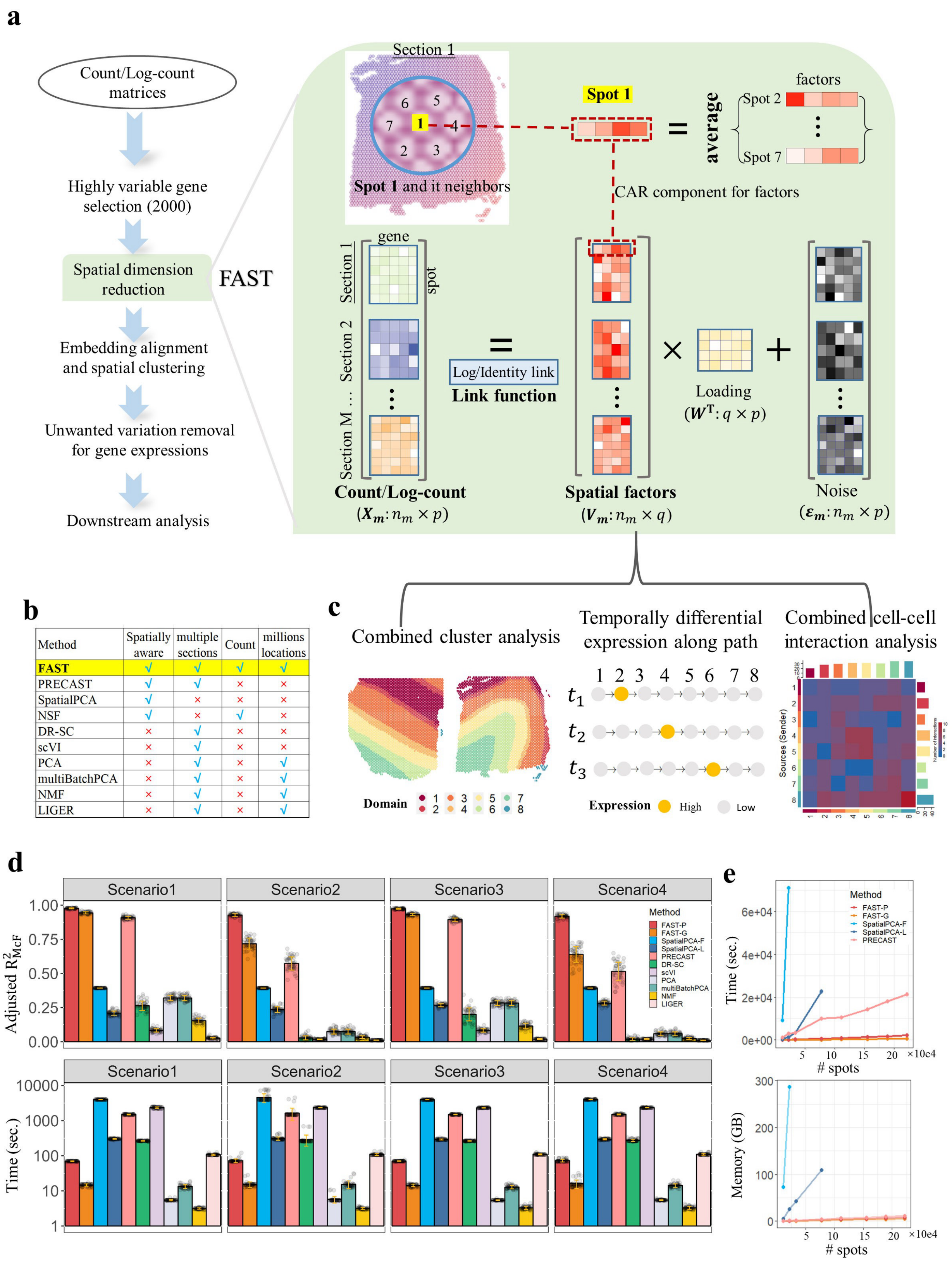
[84] Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. Normalization of rna-seq data using factor analysis of control genes or samples. *Nature biotechnology*, 32(9):896–902, 2014.

[85] Daniel McFadden. Regression-based specification tests for the multinomial logit model. *Journal of econometrics*, 34(1-2):63–82, 1987.

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

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

[88] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):174, 2017.



Method	Spatially aware	multiple sections	Count	millions locations
FAST	√	√	√	√
PRECAST	√	√	×	×
SpatialPCA	√	×	×	×
NSF	√	×	√	×
DR-SC	×	√	×	×
scVI	×	√	×	×
PCA	×	√	×	√
multiBatchPCA	×	√	×	√
NMF	×	√	×	√
LIGER	×	√	×	√

Figure 1: Schematic overview of FAST and simulation results. (a) Left panel: Workflow of multi-section SRT data analysis comprising the following steps: count or log-count as input, highly variable gene selection, spatial dimension reduction based on FAST, embedding alignment and spatial clustering, unwanted variation removal for combing gene expressions from multi-sections, and a series of downstream analyses. Right panel: FAST is a powerful generalized probabilistic factor model that efficiently estimates embeddings while incorporating spatial smoothness across multiple expression matrices. It has the flexibility to utilize either count or log-normalized gene expression matrices as input via log or identity link, enabling it to project all spots onto a shared low-dimensional space. This capability greatly enhances downstream analyses by effectively utilizing information from all sections. (b) A comparative analysis of four aspects: spatial awareness, multi-section applicability, count matrix utilization, and scalability to millions of locations in FAST and other dimension reduction methods. (c) Representative FAST downstream analyses: combined clustering analysis, temporally differential expression analysis along differentiation path, and combined CCI analysis. (d) In the simulations, we conducted tests in four different scenarios to assess the effectiveness of FAST by varying the parameters of batch effects and biological effects between low and high values. The four scenarios tested were as follows: Scenario 1 (batch effect=low, biological effect=high), Scenario 2 (batch effect=low, biological effect=low), Scenario 3 (batch effect=high, biological effect=high), and Scenario 4 (batch effect=high, biological effect=low). Top panel: Bar plots of the adjusted $\mathrm{R}^2_{\mathrm{McF}}$ values for FAST and other methods. Bottom panel: Bar plots of the running time (seconds) for FAST and other methods. (e) Comparison of computational cost in terms of running time and memory usage for FAST, SpatialPCA and PRECAST with regard to the number of spots. Note that the lines of both versions of SpatialPCA were truncated due to their inability to handle data of such scale.

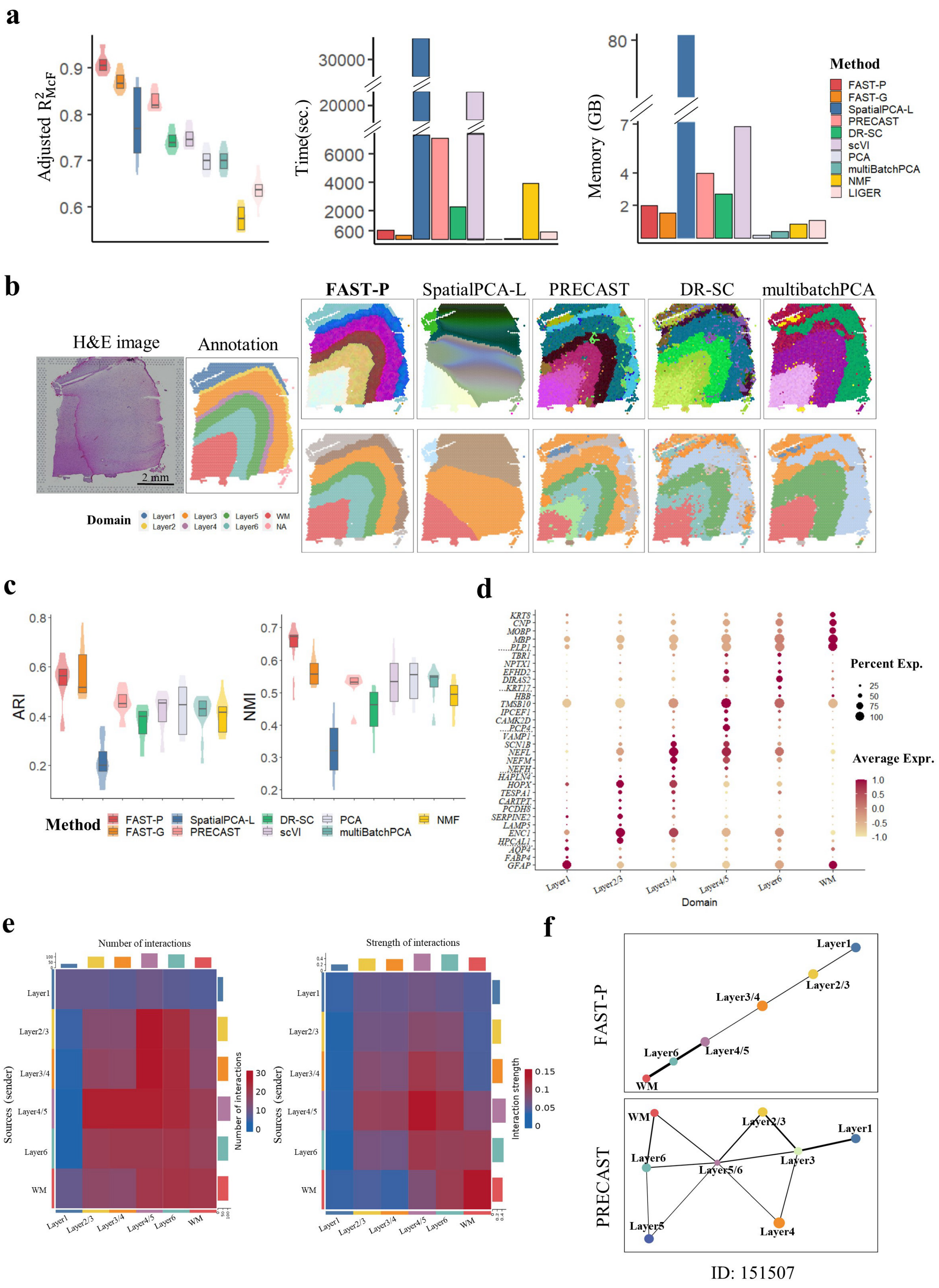

a
Adjusted R^2_{McF}
0.9
0.8
0.7
0.6
Time(sec.)
30000
20000
6000
4000
2000
600
Memory (GB)
80
7
4
2
Method
FAST-P
FAST-G
SpatialPCA-L
PRECAST
DR-SC
scVI
PCA
multiBatchPCA
NMF
LIGER
b
H&E image
2 mm
Annotation
FAST-P
SpatialPCA-L
PRECAST
DR-SC
multibatchPCA
Domain
Layer1
Layer2
Layer3
Layer4
Layer5
Layer6
WM
NA
c
ARI
0.8
0.6
0.4
0.2
NMI
0.7
0.6
0.5
0.4
0.3
0.2
Method
FAST-P
FAST-G
SpatialPCA-L
PRECAST
DR-SC
scVI
PCA
multiBatchPCA
NMF
d
KRT8
CNP
MOBP
MBP
PLP1
TBR1
NPTX1
EFHD2
DIRAS2
KRT17
HBB
TMSB10
IPCEF1
CAMK2D
PCP4
VAMP1
SCN1B
NEFL
NEFM
NEFH
HAPLN4
HOPX
TESPA1
CARTPT
PCDH8
SERPINE2
LAMP5
ENC1
HPCAL1
AQP4
FABP4
GFAP
Layer1
Layer2/3
Layer3/4
Layer4/5
Layer6
WM
Domain
Percent Exp.
25
50
75
100
Average Expr.
1.0
0.5
0.0
-0.5
-1.0
e
Number of interactions
Sources (sender)
Layer1
Layer2/3
Layer3/4
Layer4/5
Layer6
WM
Number of interactions
30
20
10
0
Strength of interactions
Interaction strength
0.15
0.1
0.05
0
f
FAST-P
Layer1
Layer2/3
Layer3/4
Layer4/5
Layer6
WM
PRECAST
WM
Layer2/3
Layer1
Layer6
Layer5/6
Layer3
Layer5
Layer4
ID: 151507

Figure 2: Analysis of human DLPFC data ($n = 47,680$ locations over 12 tissue sections). (a): Box/violin plot of adjusted $\mathrm{R}^2_{\mathrm{McF}}$ values for FAST and other methods, and bar plots of running time (seconds) and memory usage (GB) for FAST and other methods. In the boxplot, the center line and box lines denote the median, upper, and lower quartiles, respectively. (b): Left panel: H&E image and manual annotation of sample ID 151674. Top panel: UMAP RGB plots of sample ID 151674 for FAST, SpatialPCA-L, DR-SC, mutibatchPCA, and NMF. Bottom panel: Clustering assignment heatmaps for these five methods. (c): Box/violin plots of ARI/NMI values for FAST and other methods. In the boxplot, the center line and box lines denote the median, upper, and lower quartiles, respectively. (d): Dot plot of the normalized expression levels aligned across 12 sections for the marker genes of the layers detected by FAST-P. (e) Combined cell-cell interaction analysis using the normalized expression levels aligned across 12 sections and layer labels obtained by FAST-P. Heatmap of the number/strength of interactions among the different layers. (f) PAGA trajectory for sample ID 151507 using the embeddings and layer labels obtained by FAST-P and PRECAST, respectively.

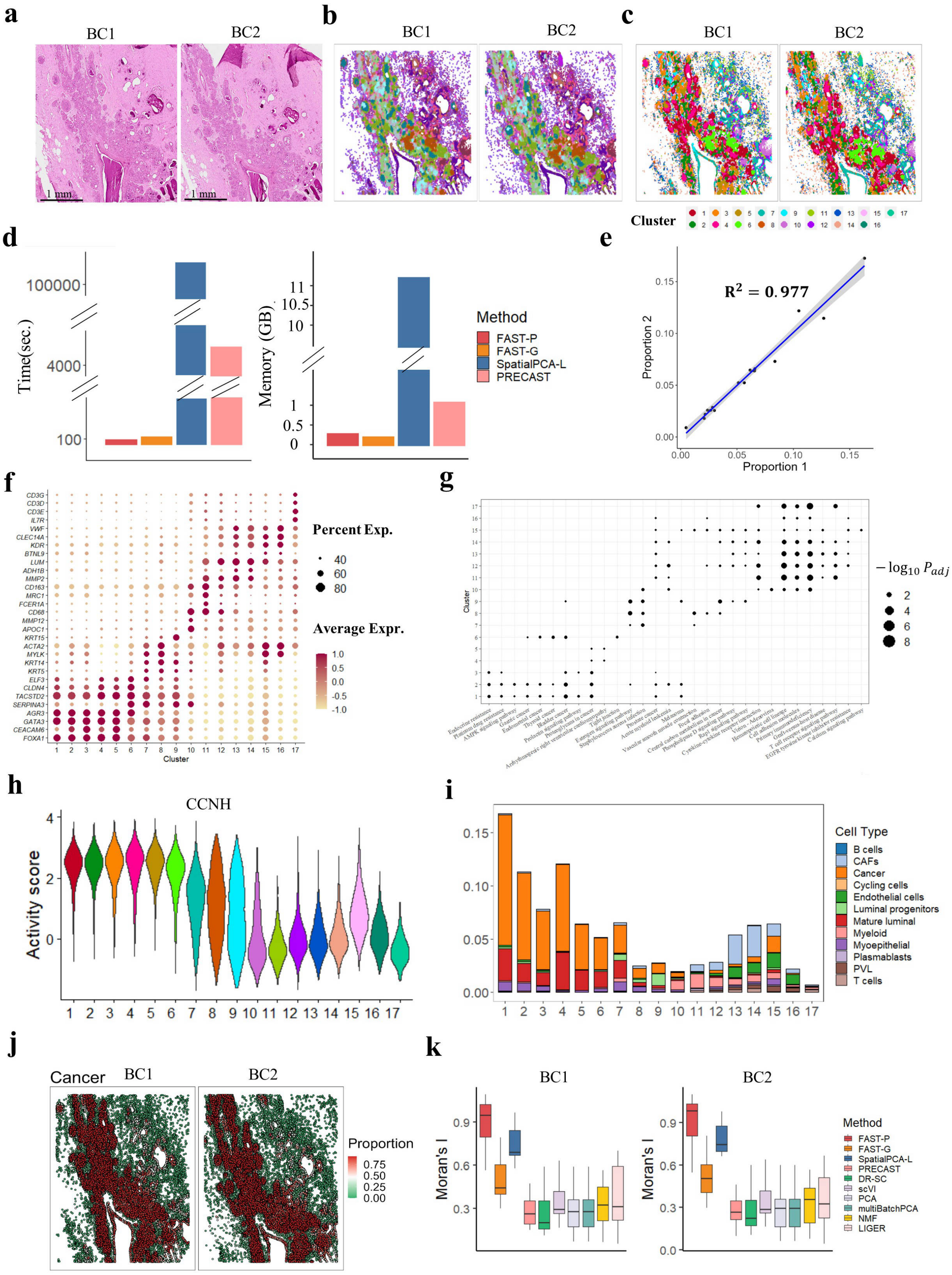

a
BC1
BC2
1 mm
1 mm
b
BC1
BC2
c
BC1
BC2
Cluster
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
d
Time(sec.)
100000
4000
100
Memory (GB)
11
10.5
10
1
0.5
0
Method
FAST-P
FAST-G
SpatialPCA-L
PRECAST
e
$R^2 = 0.977$
Proportion 2
Proportion 1
0.00
0.05
0.10
0.15
f
CD3G
CD3D
CD3E
IL7R
VWF
CLEC14A
KDR
BTNL9
LUM
ADH1B
MMP2
CD163
MRC1
FCER1A
CD68
MMP12
APOC1
KRT15
ACTA2
MYLK
KRT14
KRT5
ELF3
CLDN4
TACSTD2
SERPINA3
AGR3
GATA3
CEACAM6
FOXA1
Cluster
Percent Exp.
40
60
80
Average Expr.
1.0
0.5
0.0
-0.5
-1.0
g
Cluster
Endocrine resistance
Platinum drug resistance
AMPK signaling pathway
Gastric cancer
Endometrial cancer
Thyroid cancer
Bladder cancer
Prolactin signaling pathway
Proteoglycans in cancer
Arrhythmogenic right ventricular cardiomyopathy
Tight junction
Estrogen signaling pathway
Staphylococcus aureus infection
Prostate cancer
Acute myeloid leukemia
Melanoma
Vascular smooth muscle contraction
Focal adhesion
Central carbon metabolism in cancer
Phospholipase D signaling pathway
Rap1 signaling pathway
Cytokine-cytokine receptor interaction
Virion - Adenovirus
Hematopoietic cell lineage
Cell adhesion molecules
Primary immunodeficiency
Graft-versus-host disease
T cell receptor signaling pathway
EGFR tyrosine kinase inhibitor resistance
Calcium signaling pathway
$-\log_{10} P_{adj}$
2
4
6
8
h
CCNH
Activity score
4
2
0
i
0.15
0.10
0.05
0.00
Cell Type
B cells
CAFs
Cancer
Cycling cells
Endothelial cells
Luminal progenitors
Mature luminal
Myeloid
Myoepithelial
Plasmablasts
PVL
T cells
j
Cancer
BC1
BC2
Proportion
0.75
0.50
0.25
0.00
k
BC1
BC2
Moran's I
0.9
0.6
0.3
0.0
Method
FAST-P
FAST-G
SpatialPCA-L
PRECAST
DR-SC
scVI
PCA
multiBatchPCA
NMF
LIGER

Figure 3: Analysis of human breast cancer Xenium data ($n = 72,651$ locations over two tissue sections). (a): H&E images of the two sections. (b): UMAP RGB plots of these two sections for FAST-P. (c): Clustering assignment heatmaps of these two sections for the 17 clusters detected by FAST-P. (d): Bar plots of running time (seconds) and memory usage (GB) for FAST, SpatialPCA-L and PRECAST. (e): Scatter plot of the proportions of the 17 clusters in sections BC1 and BC2, with the fitted smoothing line and confidence band determined by the linear regression. (f) Dot plot of the normalized expression levels aligned across two sections for the marker genes of the 17 clusters detected by FAST-P. (g) Dot plot of the significant KEGG pathways of the marker genes for the 17 clusters identified by FAST-P. (h) Violin plot of the activity score for the regulatory protein *CCNH* in the 17 clusters. (i) Percentage of different cell types in each domain detected by FAST-P with scaling to the summation of all cell types across all domains equal to 100%. CAFs: cancer-associated fibroblasts; PVL: perivascular-like cells. (j) Spatial heatmap of deconvoluted cell proportions in cancer cells. (k) Boxplots of the Moran's I for the 15-dimensional embeddings obtained by FAST and other methods. In the boxplot, the center line, box lines and whiskers represent the median, upper, and lower quartiles, and the 1.5-times interquartile range, respectively.

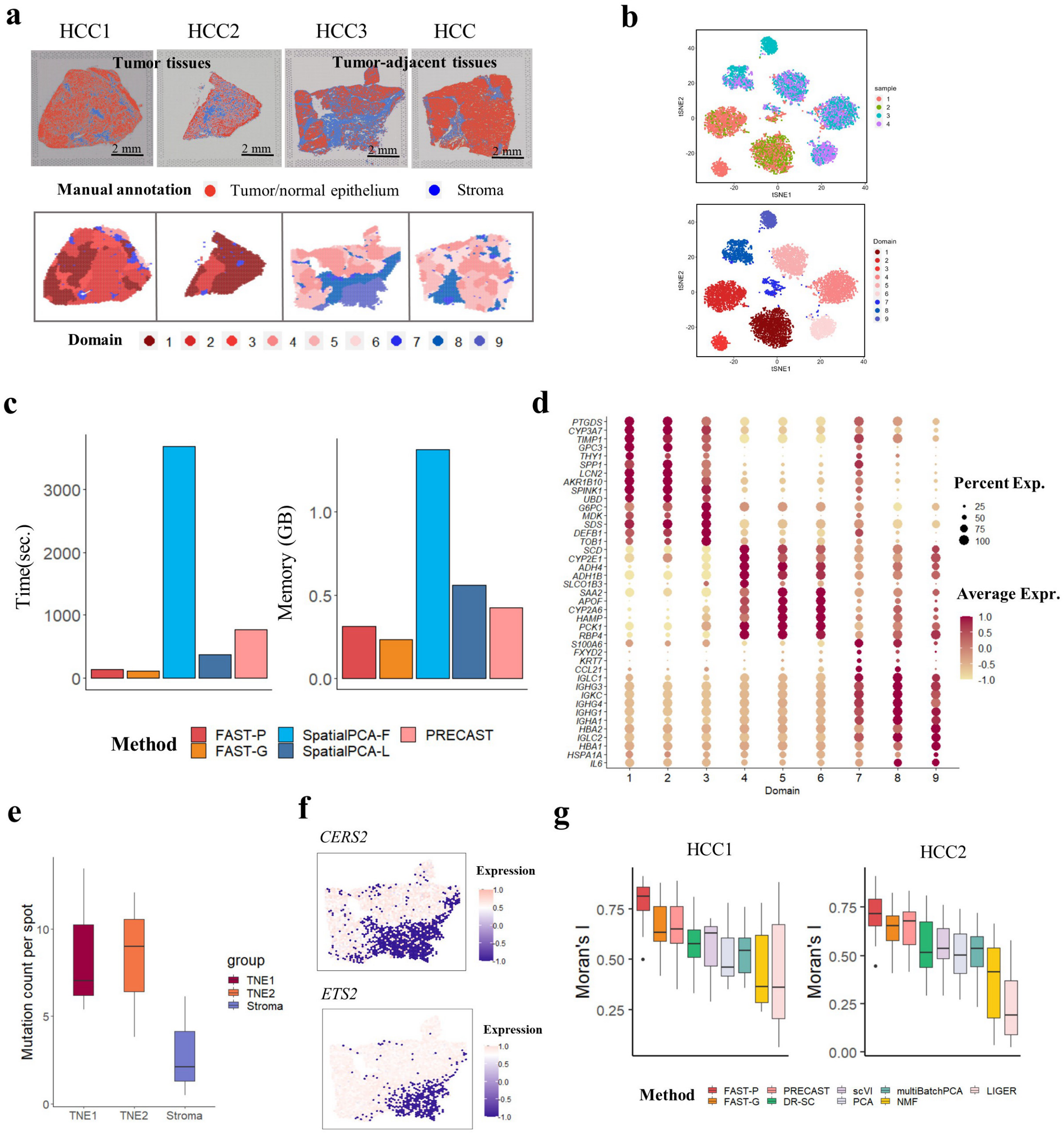

a
HCC1
HCC2
HCC3
HCC
Tumor tissues
Tumor-adjacent tissues
2 mm
Manual annotation
Tumor/normal epithelium
Stroma
Domain
1 2 3 4 5 6 7 8 9
b
tSNE1
tSNE2
sample
1 2 3 4
Domain
1 2 3 4 5 6 7 8 9
c
Time(sec.)
0 1000 2000 3000
Memory (GB)
0.0 0.5 1.0
Method
FAST-P
FAST-G
SpatialPCA-F
SpatialPCA-L
PRECAST
d
PTGDS
CYP3A7
TIMP1
GPC3
THY1
SPP1
LCN2
AKR1B10
SPINK1
UBD
G6PC
MDK
SDS
DEFB1
TOB1
SCD
CYP2E1
ADH4
ADH1B
SLCO1B3
SAA2
APOF
CYP2A6
HAMP
PCK1
RBP4
S100A6
FXYD2
KRT7
CCL21
IGLC1
IGHG3
IGKC
IGHG4
IGHG1
IGHA1
HBA2
IGLC2
HBA1
HSPA1A
IL6
Domain
1 2 3 4 5 6 7 8 9
Percent Exp.
25 50 75 100
Average Expr.
1.0 0.5 0.0 -0.5 -1.0
e
Mutation count per spot
0 5 10
TNE1 TNE2 Stroma
group
TNE1
TNE2
Stroma
f
CERS2
ETS2
Expression
1.0 0.5 0.0 -0.5 -1.0
g
HCC1
HCC2
Moran's I
0.00 0.25 0.50 0.75
Method
FAST-P
FAST-G
PRECAST
DR-SC
scVI
PCA
multiBatchPCA
NMF
LIGER

Figure 4: Analysis of human hepatocellular carcinoma Visium data ($n = 9,813$ locations over four tissue sections). (a): Top panel: H&E images from four tissue sections; Bottom panel: Clustering assignment heatmaps for four tissue sections by FAST-P. (b): tSNE plots of the aligned embeddings of FAST-P. (c): Bar plots of running time (seconds) and Memory usage (GB) for FAST, SpatialPCA and PRECAST. (d): Dot plot of the normalized expression levels aligned across four sections for the marker genes of the nine clusters detected by FAST-P. (e): Box plot of mutation count per spot for TNE1 (Domains 1-3), TNE2 (Domains 4-6) and stroma (Doamins 7-9). (f): Spatial heatmap of the expression levels of *CERS2* and *ETS2* genes in HCC3, corresponding to the top two SNPs with somatic mutations. (g): Boxplots of Moran's I for the 15-dimensional embeddings obtained by FAST and other methods. We did not plot Moran's I for SpatialPCA because SpatialPCA produced almost identical embedding for each spot and had a Moran's I value of 1. In the boxplots, the center line, box lines and whiskers represent the median, upper, and lower quartiles, and 1.5-times interquartile range, respectively.

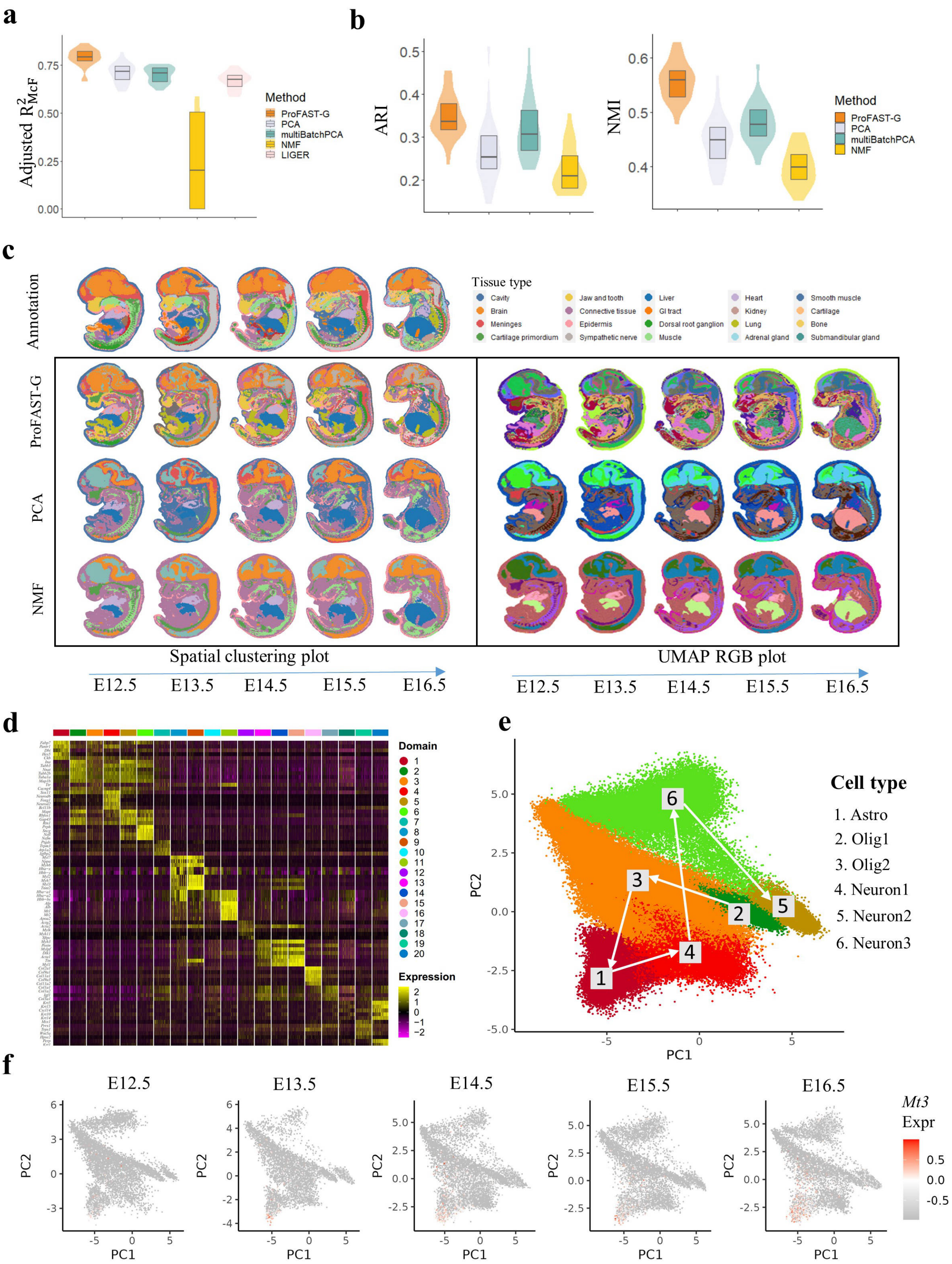

a
Adjusted R^2_{McF}
0.00
0.25
0.50
0.75
Method
ProFAST-G
PCA
multiBatchPCA
NMF
LIGER
b
ARI
0.2
0.3
0.4
0.5
NMI
0.4
0.5
0.6
Method
ProFAST-G
PCA
multiBatchPCA
NMF
c
Annotation
Tissue type
Cavity
Brain
Meninges
Cartilage primordium
Jaw and tooth
Connective tissue
Epidermis
Sympathetic nerve
Liver
GI tract
Dorsal root ganglion
Muscle
Heart
Kidney
Lung
Adrenal gland
Smooth muscle
Cartilage
Bone
Submandibular gland
ProFAST-G
PCA
NMF
Spatial clustering plot
E12.5
E13.5
E14.5
E15.5
E16.5
UMAP RGB plot
E12.5
E13.5
E14.5
E15.5
E16.5
d
Domain
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Expression
2
1
0
−1
−2
e
Cell type
1. Astro
2. Olig1
3. Olig2
4. Neuron1
5. Neuron2
6. Neuron3
1
2
3
4
5
6
PC2
PC1
5.0
2.5
0.0
-2.5
-5.0
-5
0
5
f
E12.5
E13.5
E14.5
E15.5
E16.5
PC2
PC1
Mt3
Expr
0.5
0.0
-0.5

Figure 5: Analysis of mouse embryo Stereo-seq data ($n = 2,323,044$ locations over 26 tissue sections). (a): Box/violin plot of adjusted $\mathrm{R}^2_{\mathrm{McF}}$ values for FAST-G and four other methods. (b): Box/violin plot of ARI/NMI values for iSC-MEB clustering based on the embeddings obtained by FAST-G and three other methods. (c): Left panel: Clustering assignment heatmaps for five sections (sample IDs: 1, 7, 11, 18, 23) from different embryo days (the first sample in E12.5, E13.5, E14.5, E15.5 and E16.5, respectively) by manual annotation, FAST-G, PCA and NMF. Right panel: UMAP RGB plots for these sections based on the aligned embeddings from FAST-G, PCA and NMF. (d): Heatmap of differentially expressed genes for the 20 domains identified by FAST-G. (e): Scatter plot of two-dimensional PCs from the embeddings of FAST-G within the brain region. The depicted path corresponds to the inferred differentiation pathway determined by Slingshot for the six distinct subregions within the brain. Astro: Astrocytes, Olig: Oligodendrocytes. (f): Scatter plot of two-dimensional PCs from the embeddings of FAST-G for the expression levels of gene *Mt3* during embryo days ranging from E12.5 to E16.5.