Interpretable high-resolution dimension reduction of spatial transcriptomics data by DeepFuseNMF

Junjie Tang[1,2,8], Zihao Chen[1,8], Kun Qian[1,8], Siyuan Huang[3], Shenyi Yin[4], Yang He[4], Xinyu He[4], Buqing Ye[4], Yan Zhuang[5,6], Hongxue Meng[7,*], Jianzhong Jeff Xi[4,*], Ruibin Xi[1,9,*]

[1]School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing 100871, China

[2]Current address: School of Computer Science, Department of Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[3]Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China

[4]State Key Laboratory of Natural and Biomimetic Drugs, Department of Biomedical Engineering, College of Future Technology, Peking University, Beijing 100871, China

[5]Department of Colorectal, Tianjin Medical University Cancer Institute & Hospital, Tianjin 300060, China

[6]Tianjin's National Clinical Research Center for Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin 300060, China

[7]Department of Pathology, Harbin Medical University Cancer Hospital, Harbin 150081, China

[8]These authors contributed equally

[9]Lead Contact

*Correspondence: ruibinxi@math.pku.edu.cn (R.X.), jzxi@pku.edu.cn (J.J.X), menghongxue15@163.com (H.M.)

Abstract

Spatial transcriptomics (ST) technologies have transformed our ability to study tissue architecture by capturing gene expression profiles along with their spatial context. However, the high-dimensional ST data often have limited spatial resolution and exhibit considerable noise and sparsity, thus posing significant challenges for deciphering subtle spatial patterns. To address these challenges, we introduce DeepFuseNMF, a novel multi-modal dimensionality reduction framework that enhances spatial resolution by integrating low-resolution ST data with high-resolution histology images.

DeepFuseNMF incorporates nonnegative matrix factorization into a neural network architecture for interpretable high-resolution embedding learning and spatial domain detection. Furthermore, DeepFuseNMF seamlessly handles multiple samples simultaneously and is compatible with various types of histology images. Extensive evaluations using synthetic datasets and real ST datasets from various technologies and tissue types demonstrate DeepFuseNMF's ability to produce highly interpretable high-resolution embeddings and to detect refined spatial structures. DeepFuseNMF represents a powerful multi-modal data integration approach to fully leverage the rich information in ST data and associated high-resolution image data, paving the way for understanding the organization and function of complex tissue structures.

Introduction

The revolutionary spatial transcriptomic (ST) technologies[1-4] have enabled the simultaneous capture of gene expression profiles while preserving the spatial locations of cells or spots within a tissue sample[5, 6]. These cutting-edge techniques have been widely applied in various fields to uncover the complex tissue structures, organizations, and microenvironments[7-9] . However, the spatial resolution of commonly used platforms such as 10X Visium remains limited. Each ST spot typically contain multiple cells, and large tissue areas are not covered by ST spots. The newer technologies such as Stereo-seq[10] and Xenium[11] can achieve much higher resolution, but their data are much sparser and noisier. Aggregation of individual spots into larger meta-spots is often needed to increase the signal-noise ratio, which unfortunately significantly reduces spatial resolution[12].

The ST data are high-dimensional and often exhibit considerable noise and sparsity, necessitating dimension reduction as a crucial step in their analysis. Since traditional dimension reduction methods cannot consider spatial information, specialized methods[13-19] tailored for ST data have recently been developed. By leveraging spatial information, these methods can generate spatially-aware low-dimensional representations that preserve spatial relationships between cells or spots and allow for more accurate identification of spatial patterns. A number of these methods, such as NSFH[16] and SpiceMix[17], are based on Nonnegative Matrix Factorization (NMF), and thus can generate more interpretable parts-based representations of the high dimensional ST data. However, the available

dimension reduction techniques for ST data are constrained by the resolution limitation, facing significant challenges in detecting subtle, localized expression patterns and in deciphering complex spatial architectures. Further, when ST data contain multiple samples, spots between samples may not have a spatial relationship. Methods that heavily rely on spatial information are not directly applicable to these multi-sample datasets.

ST data are often accompanied by histology images, such as Hematoxylin-Eosin (H&E) stained images, or immunofluorescence images, which provide valuable information about the tissue morphology, cellular composition, and spatial organization at a higher resolution. Several methods that can integrate ST gene expression and histology image patches around sequenced spots have been developed for dimension reduction and region detection[13, 20, 21]. To enhance spatial resolution in ST data, novel computational methods have been developed to enable high-resolution recovery of gene expression and improved identification of tissue architecture [22-25]. Nevertheless, these multi-modal resolution-enhanced methods are designed to address the high-resolution gene expression recovery rather than the high-resolution dimension reduction problem. Furthermore, they primarily focus on integrating H&E images and often lack support for other types of histology images, and do not support the simultaneous analysis of multiple samples.

In this paper, we present DeepFuseNMF (deep learning fused with NMF), a multi-modal dimension reduction framework to generate interpretable high-resolution representations of the ST data by leveraging histology images. In DeepFuseNMF, a two-modal encoder is developed to identify the interpretable high-resolution representations by integrating the low-resolution spatial gene expression from ST data with the high-resolution histological feature from histology images. Then, a two-modal decoder uses the representations to recover the spatial gene expression and the histology image. Similar to NMF, the learnable loading matrix in the expression's decoder induces the interpretability to the high-resolution representation. This versatile framework can seamlessly handle multiple samples simultaneously and is compatible with various types of histology images. We benchmarked DeepFuseNMF using two synthetic datasets and 12 real ST datasets from various technologies and tissue types, The real datasets included 5 mouse brain datasets, 4 medulloblastoma datasets from patient-derived orthotopic xenograft (PDOX) mice, and 3 newly sequenced datasets from colorectal cancer (CRC) patients. These extensive benchmarking analyses demonstrated that DeepFuseNMF

generated highly interpretable high-resolution embeddings and helped uncover complex and fine spatial structures that were difficult for available methods.

Results

DeepFuseNMF fuses expression and image data for high-resolution dimension reduction

DeepFuseNMF is based on a multi-modal neural network that takes advantage of the high-dimensionality of transcriptomics data and the high-resolution of image data to achieve interpretable high-resolution dimension reduction (Fig.1; Methods). The high-dimensional expression data enable refined functional annotations and the high-resolution image data help to enhance the spatial resolution. After the high-resolution embedding learning, DeepFuseNMF can then perform high-resolution downstream analyses, such as identification of embedding-associated genes, high-resolution spatial domain detection, and gene expression recovery.

The DeepFuseNMF learning framework has two major components, feature representation and high-resolution embedding learning. (1) In the feature representation component, DeepFuseNMF independently learns a low-resolution embedding from the ST data and a high-resolution pixel-wise visual feature from the image data (Fig.S1; Methods). The low-resolution embedding is learned by NMF coupled with a graph convolutional network (GCN). For the image feature representation, DeepFuseNMF pretrains a multi-channel U-Net that recovers the image and then transfers its backbone to get pixel-wise image features. (2) DeepFuseNMF then leverages a multi-modal fusion auto-encoder to learn a high-resolution embedding via simultaneously reconstructing the ST expression and image data. The encoder (feature fusion module) fuses and maps the low-resolution embedding and the pixel-wise image features to a high-resolution embedding space. Then, the decoder (data reconstruction module) reconstructs expression and image data based on the learned high-resolution embedding. Similar to NMF, the entries of the embedding and the learnable loading matrix in the expression's recovery module are all nonnegative, thus allowing an interpretable parts-based high-resolution representation of the data. The reconstruction loss is a weighted sum of the losses for the expression and image data reconstruction. Details of the DeepFuseNMF framework are provided in Methods.

DeepFuseNMF gives more accurate embedding and spatial clustering in simulation

We evaluated the performance of DeepFuseNMF in comparison with available NMF-based methods, including NMF[26], NSFH[16], and SpiceMix[17], as well as Giotto[18] using two simulations. The available methods can only give low resolution embedding. We thus further included a high-resolution embedding method TESLA+NMF in our comparison. In this method, we employed TESLA[24] to recovery high-resolution expression data from the image data, and then applied NMF to the imputed high-resolution expression data to obtain the high-resolution embedding.

The first simulation utilized H&E image of mouse cerebellum for data generation (Fig.2a; Methods). The image was segmented to 3 distinct regions based on the established mouse brain annotation and intensities for the embedding dimension of each region were generated. Gene expression data at individual spots were simulated using zero-inflated Poisson distributions, where the distribution parameters were set to be dependent on the spot embedding intensities. To simulate different resolutions of the ST data, we varied the number of spots through random spot exclusion or adjusted the spot radius. Additionally, we introduced additive noises or dropouts to simulate different levels of noise in the ST data (Methods).

In terms of the recovery of embedding intensities, DeepFuseNMF consistently demonstrated superior performance, often by a large margin, than other methods as measured by the mean absolute error (MAE) between the true and inferred embedding intensities (Fig. 2b; Fig. S2a, c-e; Methods). After embedding learning, we proceeded to identify spatial domains by applying the k-means algorithm to the learned low-dimensional representations and compared clustering accuracies using the Adjusted Rand Index (ARI) (Methods). Again, we found that DeepFuseNMF gave substantially higher ARIs than other methods in all simulation scenarios (Fig. 2c; Fig. S2b), demonstrating the effectiveness of DeepFuseNMF's integration of ST expression data and high-resolution image data. Fig. 2d and 2e show the local embedding and clustering, in one of the simulation datasets (spot radius 107), given by various methods in a region of interest (ROI) that contained fine spatial structures (Fig. 2a). Even though the resolution of ST expression data was low, the spatial fine structures could still be faithfully captured by the embedding given by DeepFuseNMF (Fig. 2d), leading to its significantly more accurate clustering than other methods (Fig. 2e), especially at the sub-regions with very fine

spatial structures (the light blue boxes in the figure).

The second simulation generated data from a human breast cancer Xenium dataset[11], which was high-resolution single-cell expression data for 313 genes. We aggregated expressions from multiple neighboring single-cells to create low-resolution spot expression data (Fig. 2f; Fig. S3a, b). Similar to the first simulation, we generated simulation data with varying spatial resolutions (Methods). Comparison of the embeddings given by different methods with the high-resolution embedding inferred by NMF from the original Xenium dataset revealed that, DeepFuseNMF again consistently produced more accurate embeddings (Fig. S3c). To further assess the DeepFuseNMF's capability in recovering fine spatial structures, we conducted a highly challenging simulation scenario. In this scenario, we excluded all spots in three selected ROIs that contained local spatial structures (Fig. 2g; Fig. S3d). Although many methods could recover these local patterns when the spots in these ROIs were not excluded, only DeepFuseNMF and TESLA+NMF were able to provide embeddings in these three ROIs in this challenging simulation (Fig. 2g; Fig. S3e). Notably, only DeepFuseNMF could recover the local pattern for ROI 3.

DeepFuseNMF robustly recovered interpretable high-resolution embeddings in mouse brain datasets

We applied DeepFuseNMF to the 10X Visium ST dataset MBC-01 sequenced from an adult mouse brain coronal section, comprising three immunohistochemistry (IHC) stained images (Fig. S4a; Methods). The embeddings identified by DeepFuseNMF were enriched in known specific brain regions (Fig. S4). For example, E15 was mainly enriched in isocortex layers 1, 2/3 and 4, E18 in isocortex layers 4 and 5, and E9 in isocortex layers 6a and 6b, E10 in the dentate gyrus (DG), E14 in the pyramidal cell layers CA1 and CA2, E16 in the pyramidal cell layers CA2 and CA3 (Fig. 3a, b). Many of these embeddings were also identified by other methods (Fig. S5). However, most embeddings given by other methods were of low resolution. TESLA+NMF could give high resolution embeddings, but several its embeddings did not correspond well to the known brain regions (e.g. TESLA+NMF's embedding E5, E8 and E11 in Fig. S5). Zoomed views of embedding E1 and E6 from

DeepFuseNMF in four selected ROIs showed that local fine structures unveiled by the embeddings of DeepFuseNMF were fully supported by neuron-specific nuclear binding protein (NeuN) or glial fibrillary acidic protein (GFAP) stained image (Fig. 3c,d). In comparison, TESLA+NMF could only identify the local fine structure in ROI 1, and other methods could hardly detect any due to their low resolution (Fig. S6).

To further investigate the enrichment of the embeddings in specific brain regions, we obtained genes that were upregulated in different brain regions and investigated their rankings in the embeddings from DeepFuseNMF (Methods; Supplementary Table 2). The marker genes of different brain regions were only ranked high in a few embeddings, implying that the embeddings were associated with specific brain regions (Fig. 3e). For example, marker genes of olfactory areas were only enriched in embedding E20, those of cerebral nuclei in E13, and those of thalamus in E3. Marker genes of isocortex ranked high in three embeddings from DeepFuseNMF (E15, E18 and E9). We thus further obtained marker genes of sublayers of isocortex and found that isocortex layer 1, 2&3 and 4 were enriched in embedding E15, layer 5 in E18 and layer 6a/6b in E9 (Fig. 3f), demonstrating that E15, E18 and E9 were associated with the sublayers of isocortex. Similarly, marker genes of hippocampal formation were enriched in three embeddings (E10, E14 and E16) and these embeddings were associated with subregions of hippocampal formation (Fig. 3g).

We considered two more mouse coronal section datasets MBC-02 and MBC-03, where MBC-02 contained two IHC images and MBC-03 contained one H&E image (Methods; Fig. S7a). Note that the image data of these two datasets were different from the image data of the previous MBC-01 dataset. We found that the embeddings learned from the MBC-02 and MBC-03 datasets aligned well with the embeddings learned from the MBC-01 dataset (Fig. S7b), and their gene ranks were also highly correlated with embeddings learned from the MBC-01 dataset (Fig.S8a,b; Supplementary Table 2). Gene enrichment analysis also confirmed that the embeddings learned from the MBC-02 and MBC-03 datasets were associated with specific brain regions (Fig.S8c-f).

In addition, we further analyzed two adjacent mouse posterior brain sagittal slices (dataset MPBS-01 and MPBS-02) sequenced on the 10X Visium platform with H&E images (Methods; Fig.S9a). Again, both slices consistently exhibited enrichment of their embeddings in specific brain regions

(Fig.S9b,c), marker genes of the brain regions consistently ranked high in their respective embeddings (Fig.S9d; Supplementary Table 2), and the embeddings learned from both slices had strong correspondence (Fig.S9e). These demonstrated that DeepFuseNMF could robustly identify high-resolution interpretable embeddings from ST data with various high resolution imaging data.

DeepFuseNMF identified fine-grained spatial structure in mouse brain data

Using the embeddings, we further performed cluster analysis and detected spatial domains from the mouse brain dataset MBC-01 (Fig.4; Fig.S10; Methods). For comparison, we further included three state-of-the-art spatial domain detection methods, iStar, SpaGCN and stLearn, that also utilized image data. DeepFuseNMF, TESLA+NMF, and iStar achieved highest spatial resolution in spatial domain detection. Although SpaGCN and stLearn also utilized image data, their spatial resolutions were still at the spot level. Overall, the spatial domains from DeepFuseNMF showed better concordance with the known mouse brain domains[27] compared to other methods (Fig. 3a), and this concordance was robust to the resolution parameter of the Louvain clustering algorithm (Fig.S10).

The advantage of DeepFuseNMF was particularly evident in the hippocampal formation region that contained fine-grained spatial structures (Fig. 4a, b). Only DeepFuseNMF successfully identified and clearly delineated the subregions of the hippocampal formation and their complex topological relationships (Fig. 4a, b). Among the low-resolution methods, stLearn and SpiceMix performed the best in detecting these local regions. However, due to their lower resolution, the complex topological relationships between the subregions identified by these methods were much less clear. The high-resolution methods iStar and TESLA+NMF largely recovered pyramidal cell layers CA1, CA2 & CA3, and granule cell layer of the dentate gyrus (DG). However, iStar incorrectly separated each of pyramidal cell layer CA1 and granule cell layer of DG into two clusters, TESLA+NMF misclassified many points in pyramidal cell layers to another cluster, and both methods merged stratum layers CA1, CA2 & CA3 and molecular layer DG into a single cluster (Fig. 4a). Known marker genes[28, 29] of the hippocampal pyramidal cell layers CA1, pyramidal cell layers CA2 and CA3 and granule cell layer DG were highly expressed in these three subregions identified by DeepFuseNMF (Fig. 4c; Fig. S11a),

and the spatial expression patterns of these genes were clearer for the recovered high-resolution expression by DeepFuseNMF than the raw expression (Fig. 4c; Fig. S11).

DeepFuseNMF detected a thin cluster C9 located at the meninges (Fig. 4d; Fig. S12a). The meninges are important protective membranes that envelop the brain and consist of three layers: the dura, arachnoid, and pia mater[30, 31]. The arachnoid and pia mater together form the leptomeninges. Beneath the pia mater lies the glia limitans, a thin membrane layer primarily composed of astrocyte end-feet[32]. Within the meninges, a network of cerebral arteries and veins traverses inward to the brain[33, 34]. Cluster C9 highly expressed marker genes of astrocytes and major cell types in leptomeninges[32], including meningeal fibroblasts subtype 1 (FB1) in the pia mater, FB2, FB3 in the arachnoid mater, as well as endothelial cells and vascular smooth muscle cells (VSMCs) within the blood vessels, but not marker genes of cell types in the dura mater (Fig. 4e; Fig. S12b,c). GFAP, the protein marker of astrocytes, showed a higher IHC intensity in cluster C9 than in other clusters (Fig. 4f). These data suggested that cluster C9 was composed of the leptomeninges and the glia limitans. Several other methods also detected clusters located near the meninges (Fig. S12d). However, cluster C9 was much thinner and exhibited higher IHC intensity of GFAP (Fig.4g) as well as higher expression level of FB1, FB2 and endothelial cell marker genes (Fig. 4h; Fig. S12e), demonstrating that C9 matched better to the thin structure of the leptomeninges and glia limitans. These results highlight DeepFuseNMF's capability to enhance resolution and capture fine-scale structures.

DeepFuseNMF recovered condition-specific embeddings across multiple tumor samples

We applied DeepFuseNMF to a 10X Visium dataset of sonic hedgehog (SHH) medulloblastoma from PDOX mouse models[35]. The dataset consisted of two Palbociclib-treated samples (Palbociclib A and B) and two untreated control (Control C and D) samples (Fig. 5a). Joint analysis of these four samples by DeepFuseNMF yielded 20 embedding dimensions, ten of which were enriched in normal mouse brains, five in the implanted tumors and five in the tumor-normal interfaces, as indicated by their respective high intensity at mouse, human and human-mouse mix spots, as well as their rankings of mouse and human genes (Fig. 5a, b; Fig. S13; Supplementary Table 2). The normal embeddings

aligned well with known mouse brain structures, such as the Purkinje layer (E4), granular layer (E8), and molecular layer (E9) in the cerebellar cortex (CBX) (Fig. 5b; Fig. S13). It was found that astrocytes and microglia (resident macrophages in brains) were enriched in the tumor-mouse brain interface[35]. Consistently, the interface embedding dimensions E11 and E15 were enriched with marker genes of astrocytes and microglia, respectively (Fig. 5b; Fig. S13).

Hovestadt et al.[36] identified three important transcriptional signatures, cell cycle activity (SHH-A), undifferentiated progenitors (SHH-B) and differentiated neuronal-like programs (SHH-C), in SSH medulloblastoma. We found that tumor-enriched embedding E13 was enriched with SSH-A and E2F target genes, interface-enriched E5 with SSH-B genes and tumor-enriched E6 with SSH-C and neuronal differentiation genes (Fig. 5b; Fig. S14). E6 had increased intensities in the Palbociclib-treated samples compared with the control samples. In contrast, E13 and E5 had reduced intensities in the treatment samples, and the reduction largely only occurred in the interior regions of the tumors but not in the tumor boundaries (Fig. 5c). These results were consistent with the observation that Palbociclib could increase neuronal differentiation and decrease cell proliferation, but have less impact at the tumor boundaries[35], and consistent with the fact that the a CDK4/6 inhibitor Palbociclib can downregulate E2F-mediated transcription of cell cycle genes, thereby induce a G1 arrest and inhibit cancer cell proliferation[37].

Using the DeepFuseNMF's embeddings, we performed joint cluster analysis of the four samples (Fig. S15a). As expected, the normal brain regions shared highly similar clusters across the four samples, while the tumor regions were heterogeneous. In the tumor regions, samples with the same treatments had more shared clusters than samples with different treatments. We also applied available methods to cluster the spatial locations (Fig. S15a, b). DeepFuseNMF and SpiceMix allowed joint analysis of multiple samples, resulting in naturally aligned clusters between samples. For other methods, aligning clusters across samples was more challenging. Furthermore, DeepFuseNMF revealed refined spatial structures. For example, DeepFuseNMF clearly characterized the ordered layers of CBX (Fig. 5d; Fig. S15c, d), including the molecular layer (C17, C9), Purkinje layer (C13) and granular layer (C5). In comparison, other methods could not clearly delineate the CBX layers, especially the fine Purkinje layer (black dotted area in Fig. 5d). Among the low-resolution methods, SpiceMix performed the best in detecting the CBX layers, possibly due to its ability to integrate

multiple samples. Interestingly, the molecular layer cluster C17 was adjacent to the tumor regions, and thus more susceptible to tumor cell infiltration (Fig. S15e). Consistently, C17 was significantly enriched with human genes and exhibited higher invasion and proliferation activity compared to C9 and other normal clusters (Fig. 5e). Additionally, the C17 areas decreased in the Palbociclib-treated samples compared to the control samples, likely due to Palbociclib's effect on reducing cell proliferation and invasion (Fig. S15e). These analyses demonstrate that joint high-resolution embedding learning by DeepFuseNMF can successfully capture multi-sample fine-scale structures and enable comparative analysis between multiple samples.

DeepFuseNMF depicts tumor heterogeneity and immune activities in CRC samples

We applied DeepFuseNMF to the three newly profiled ST CRC samples (CRC-01, 02, 03) for spatial domain detection. According to the pathologist annotation (Fig. 6a), sample CRC-01 comprised four major spatial clusters: tumor, interstitial, necrotic, and muscle clusters. DeepFuseNMF identified 16 more refined spatial subclusters within these broader categories, including 6 tumor subclusters, 7 interstitial subclusters, 1 necrotic subclusters, and 2 muscle subclusters (Fig. 6b; Fig. S16a). Compared with spatial clusters given by available methods, clusters given by DeepFuseNMF agreed better with the H&E image (Fig. 6a-c; Fig. S16b-d). For example, only DeepFuseNMF accurately distinguished necrosis from tumor in ROI 1, and tumor from interstitial tissues in ROI 2 and ROI 3 (Fig. 6d; Fig. S16c).

Among the tumor clusters, T1, T2 and T3 resembled the CMS2 subtype of colorectal cancer, T6 was similar to the CMS4 subtype, and T4 and T5 were intermediate states between CMS2 and CMS4[38, 39] (Fig. S17a; Supplementary Note 1), suggesting that these subclusters were distinct tumor spatial subclusters. Interestingly, in ROI 1, the tumor clusters T1 and T3 had a clear boundary on the H&E image, which were only detected by DeepFuseNMF (Fig. 6d; Fig.S16c). The H&E image showed that T1 was more enriched with fibroblast than T3 (Fig. S17b). Consistently, T1 showed significantly higher expressions of fibroblast and extracellular matrix (ECM) genes than T3 (Fig. 6e). Deconvolution analysis by RCTD[40] (Methods) also showed that tumor purity in T1 (0.56) was lower

than that in T3 (0.70). Further, T1 had significantly higher cancer-associated ECM[41] (C-ECM) and invasion scores (Fig. 6e), indicating that T1 might be more aggressive and invasive. In fact, genes upregulated in T1 relative to T3 were predictive of the overall worse survival of colorectal patients in the TCGA dataset[42] (Fig. S17c). Taken together, these results supported that T1 and T3 were distinct tumor clusters with important clinical implications. Similarly, DeepFuseNMF accurately detected fine structures in sample CRC-02 (Fig. S18; Supplementary Note 1).

Sample CRC-03 was sequenced by Stereo-seq without a paired high-resolution image. To apply DeepFuseNMF and other methods that require image data, we first generated a pseudo-image using the noisy subcellular Stereo-seq expression data (Fig. S19a; Methods). DeepFuseNMF identified 5 tumor subclusters and 2 normal clusters (Fig.6f, g; Fig. S19b; Supplementary Note 1). Comparing with other high-resolution methods (Fig. 6h; Fig. S19c), clusters given by DeepFuseNMF agreed better with the H&E image of a nearby slice (Fig. 6f). For example, in an ROI (the black-box region in Fig. 6f-h; Fig. S19c, d), DeepFuseNMF and spot-level methods correctly distinguished normal and tumor regions. In contrast, iStar and TESLA+NMF misclassified large tumor regions as normal, possible because they were designed for integrating H&E images and may not be suitable for analyzing the pseudo-image generated from Stereo-seq data.

Moreover, DeepFuseNMF identified a refined cluster I1 around the tumor boundary (Fig. 6g). Compared with the nearby interstitial cluster I2, cluster I1 was more enriched with *C1QC*+ macrophages[43], but not macrophages (Fig. 6i; Fig. S19e). *C1QC*+ macrophages were known to have increased antigen-presenting activities[43]. Consistently, genes highly related to I1 were significantly enriched with antigen-presenting genes (Fig. 6j). These implied that I1 might be involved in tumor-associated inflammatory responses. Moreover, genes highly expressed in I1 were associated with overall better survival of colorectal patients in the TCGA dataset (Fig. S19f), indicating the potential clinical relevance of this refined cluster. These analyses demonstrate the superior performance of DeepFuseNMF in accurately identifying fine spatial structures of tumor samples at high resolution, even in the absence of a paired high-resolution image.

Discussion

In this paper, we developed DeepFuseNMF that can generate high-resolution, interpretable embeddings by effectively integrating ST gene expression data and high-resolution histology image data. DeepFuseNMF can jointly analyze multiple samples and support various types of histology images. This novel approach opens up new possibilities for uncovering fine-grained spatial patterns and biological insights that were previously unattainable due to the resolution limitations of ST technologies. Extensive benchmarking analyses using both simulated and real ST datasets demonstrated that DeepFuseNMF consistently outperformed currently available state-of-the-art methods in embedding learning and spatial domain detection, particularly in challenging scenarios with complex spatial structures. In computational speed, DeepFuseNMF is comparable to the available methods, but it requires slightly more memory usage (Fig. S20).

Deep learning methods have been widely used for dimension reduction or clustering analysis of single-cell and ST data due to their performance advantages over traditional statistics and machine learning methods[44-47]. However, the black-box nature of these methods often hinders the interpretability of the learned features, which is crucial for understanding and unveiling the important structures in the high-dimensional single-cell or ST data. To address this limitation, DeepFuseNMF introduces a framework that fuses NMF with a neural network, enabling more interpretable dimension reduction, while leveraging the power of deep learning and rendering higher resolution. As demonstrated in applications to mouse brain and PDOX datasets, the learned embeddings often represented unique spatial expression patterns or known tissue structures, and could even reveal important transcriptional signatures related with tumor treatment, thus facilitating a more straightforward interpretation of the data. The NMF component can also be replaced by other dimension reduction methods, such as the spatially-aware methods NSFH[16] or SpatialPCA[14]. For example, incorporating the spatial NMF method NSFH might enable more accurate estimation of embeddings with spatial patterns.

Based on the high-resolution embedding, DeepFuseNMF enabled identification of refined spatial structures. In the mouse brain data, DeepFuseNMF identified fine-grained structures in the hippocampal formation and the thin structure of the leptomeninges and glia limitans (cluster C9).

Beyond physical protection, meninges play important roles in central nervous system homeostasis, development, immunity and pathology[30, 33]. Accurate meningeal detection is important for understanding their diverse functions and involvement in brain health and disease. DeepFuseNMF's ability to resolve the intricate leptomeningeal and glia limitans anatomy highlights its potential for enhancing the study of these important delicate brain structures. One potential limitation is that DeepFuseNMF could not separate the two sublayers of leptomeninges and the glia limitans, as the pia mater and the glia limitans are extremely thin, comprising only one or two layers of cells, making their delineation highly challenging.

In CRC-01, we identified two spatially adjacent tumor clusters, T1 and T3, with significantly different levels of invasiveness and fibroblast infiltration, despite having similar copy number profiles (Fig. S17e). This suggests that the difference in invasiveness is more likely associated with the varying levels of fibroblast infiltration rather than genetic differences. Cancer-associated fibroblasts (CAF) can modify ECM proteins to produce stiff and intricated fiber organization, thus increasing invasion potential and may even hinder drug delivery[48]. The presence of adjacent tumor clusters like T1 and T3, with major differences in fibroblast infiltration, highlights the potential impact of CAFs on the local tumor microenvironment and the resulting heterogeneity in invasiveness. By analyzing of a larger cohort of samples using DeepFuseNMF, we can explore the prevalence and significance of adjacent tumor clusters with distinct fibroblast infiltration levels and their association with various tumor phenotypes.

In CRC-03, we identified a normal cluster I1 near the tumor boundary that were enriched with *C1QC*+ macrophage. Previous studies have shown that *C1QC*+ macrophages are pro-inflammatory. They primarily interact with other immune cells, especially T cells, and are thought to be involved in recruiting or activating T cells. However, we did not observe elevated expression of T cell genes in cluster I1, suggesting that the T cell recruiting or activating role of *C1QC*+ macrophages might depend on other factors within the tumor microenvironment. Consistently, *C1QC*+ macrophages could be identified in healthy individuals, again indicating that their presence alone may not be sufficient to trigger a robust T cell response[49].

DeepFuseNMF could be further improved in several directions. First, DeepFuseNMF currently requires matched ST expression and histology image data. However, certain ST technologies, such as Stereo-seq, do not have paired image data from the same tissue slice. Although we demonstrated an application of DeepFuseNMF to Stereo-seq data using generated pseudo-image data, extending DeepFuseNMF to handle ST data with image data from adjacent tissue slices will expand the applicability of DeepFuseNMF to a wider range of ST datasets. This extension would involve simultaneous embedding learning and alignment of adjacent slices. Second, with the fast development of ST technologies, ST datasets are becoming increasingly larger. There is an urgent need to develop computational methods that can effectively handle large scale ST datasets. By leveraging information from a large number of samples, such methods could enable identification of important spatial structures or patterns that are hard to be detected using only a few samples. Third, the interpretability of DeepFuseNMF is currently achieved by extracting the associated gene signature for each embedding dimension, which may represent complex biological processes. Incorporating prior knowledge of biological pathways into DeepFuseNMF would enhance interpretability by breaking down these multifaceted processes into specific pathway-related gene signatures.

Methods

DeepFuseNMF algorithm

Data preprocessing. DeepFuseNMF takes spatial transcriptomics (ST) data and high-resolution image data as input. Let $\Omega = \{1,2,\dots,h\} \times \{1,2,\dots,w\}$ be the sequenced tissue with $h \times w$ pixels (height h, width w), $I \in \mathbb{R}^{h \times w \times c}$ be the normalized image (divided by the maximum value, with a range in [0,1]) with height h, width w and c channels within the sequenced tissue, and $\mathrm{X}_0 \in \mathbb{R}^{N \times G_0}$ be the unique molecular identifier (UMI) count data with G_0 genes at N sequenced spots. For spot $n\,(n = 1,2,\dots,N)$, let $S_n = \{(x,y) \in \Omega \mid (x,y) \text{ belongs to spot } n\}$ stand for pixel locations within this spot. For each spot, the count expression data are normalized by dividing the total UMI count in the spot, multiplying 10,000, and then transforming to the log scale. After normalization, we select

top G (default $G = 3000$) spatially varying genes with the highest Moran's I[50, 51] for downstream analyses (denoted as $X \in \mathbb{R}^{N \times G}$).

Image feature representation learning. DeepFuseNMF pretrains a neural network (named U-Net-DeepFuseNMF) based on U-Net, a popular architecture for semantic segmentation[52] and image restoration[53], for ST image feature representation learning. Specifically, U-Net-DeepFuseNMF first encodes 256×256 image patches $I \in \mathbb{R}^{256 \times 256 \times c}$ centered at given pixels from the ST images with a series of convolutions and down-sampling operations to extract multi-scale image features, resulting in a semantically rich image feature representation $\hat{Z} \in \mathbb{R}^{16 \times 16 \times 128}$. Next, U-Net-DeepFuseNMF uses a series of convolution operations and up-sampling operations to gradually merge the multi-scale image features into a high-resolution image feature representation $Z \in \mathbb{R}^{256 \times 256 \times 32}$. Finally, a convolution operation and an element-wise Sigmoid function are applied to reconstruct the image patches $\tilde{I} \in \mathbb{R}^{256 \times 256 \times c}$. The training data is taken as sliding overlapping 256×256 image patches from the ST images (15% overlap between neighboring patches). The reconstruction loss is the mean squared error (MSE) loss between the input image and the reconstructed image, and is minimized by the Adam optimizer with weight decay of 1×10^{-5}. We set an initial learning rate to 4×10^{-4} and decreased to 1×10^{-6} with cosine annealing strategy[54]. The mini-batch size is set to 32 image patches, and the number of iterations is set to 5000 by default.

Low-resolution embedding learning. DeepFuseNMF first learns a low-resolution embedding from the ST expression and image data. Let $M > 0$ be the dimension of the embedding space. Applying NMF to the normalized gene expression data $X \in \mathbb{R}^{N \times G}$, we first obtain an interpretable representation of the expression data. That is, we have a factor matrix (i.e. embedding) $F \in \mathbb{R}^{N \times M}$ and a loading matrix $W \in \mathbb{R}^{M \times G}$ $(F \geq 0, W \geq 0)$ such that

$$(W, F) = \underset{W \geq 0, F \geq 0}{\operatorname{Argmin}} \frac{1}{2} ||X - FW||_F^2 \ . \tag{1}$$

However, standard NMF can only provide embeddings for the sequenced spots, which are spatially very sparse, and cannot provide any information for other locations. In addition, this simple NMF embedding is only based on the noisy expression data and ignores the spatial and image information, and thus tends to be noisy and difficult to capture spatial fine structures. DeepFuseNMF first integrates expression data, image data and spatial information and uses the following procedure to learn a denoised low-resolution embedding that can cover the entire spatial region (Fig. S1).

In the NMF, each column F_j $(j = 1, \cdots, M)$ of the factor matrix F represents a noisy estimate of the spatial expression pattern j at the N sequenced spots. Note that the spatial expression pattern j is the embedding dimension j. We aim to denoise F_j and recover the low-resolution expression pattern j at all locations of the sequenced tissue. This is achieved using a two-step procedure. The first step is to denoise and recover the expression patterns for locations that are much denser than the sequenced spots (denoise-and-recovery step), and the second step is to extend the expression patterns to the whole sequenced tissue by Voronoi tessellations[55] and rasterization (extension step).

The denoise-and-recovery step is based on a GCN. We first construct an attributed graph $G(V, E, H)$ for the GCN, where V represents the vertices of the graph, E represents the graph edges and H represents the attributes of the vertices. The vertex set V consists of all N sequenced spots and R randomly sampled pseudo spots (default: $R = 5 * N$). The attribute $H_v \in \mathbb{R}^M$ represents the expression patterns at the spot v. For a sequenced spot v, H_v is the corresponding row of F. For a pseudo spot v, H_v is unknown and need to be learnt. The graph edges are constructed by considering the spatial distances between spots and the image similarities. More specifically, for each sequenced spot $v \in V$, we first take K_0 sequenced spots that are spatially closest to v (not include v itself), and then connect v to the top K spots whose nearby images are most similar to that of v among the selected K_0 sequenced spots (default: $K = 2$, $K_0 = 50$). Then, we apply the same procedure to connect v with K neighboring pseudo spots. Thus, each sequenced spot is connected to K sequenced spots and K pseudo spots. To measure the similarity between images near spots, we extract the semantically rich image representation $\hat{Z}_{\mathrm{v}} \in \mathbb{R}^{16\times16\times128}$ by passing the square images (256×256 pixels) centered on every spot into the pretrained U-Net-DeepFuseNMF model.

Then, by averaging $\hat{Z}_{\mathrm{v}}$ through $\frac{1}{256}\sum_{x=1}^{16}\sum_{y=1}^{16}\hat{Z}_{v,x,y,:}$, we obtain a 128-dimensional vector representation for each spot and use the Pearson correlations as the image similarities.

Let A be the adjacency matrix of the graph $G(V,E,H)$, and $\tilde{A} = D^{-1}A$ be the normalized adjacency matrix, where D is the out-degree matrix of A with $D_{ii} = \sum_{j=1}^{(N+R)} A_{ij}$. Denote $h_i^{(0)} = H_i$ $(i = 1, \dots, N+R)$ and $H^{(0)} = \left(h_1^{(0)}, \cdots, h_{N+R}^{(0)}\right)^T$. The l-th layer of the GCN is updated using

$$H^{(l)} = f^{(l)}\left(\tilde{A}H^{(l-1)}W_l\right),$$

where $f^{(l)}$ is the nonlinear activation function and W_l is the trainable weight matrix. The encoder of this GCN network is set to a two-layer neural network (dimension: $M \to 64 \to 256$), and the decoder is set to a two-layer neural network (dimension: $256 \to 64 \to M$). The activation function of the first three layers is the rectified linear unit (ReLU), and the activation function of the final layer is the Sigmoid function. The output of the GCN, $H^{(4)} \in R^{(N+R)\times M}$, is the predicted embedding for observed spots and pseudo spots. The loss function is the squared loss between the output layer and the input layer on the observed spots:

$$L = \left\| H_{obs}^{(4)} - H_{obs}^{(0)} \right\|_F^2,$$

where $\|\cdot\|_F$ represents the Frobenius norm and $H_{obs}^{(j)}$ is the submatrix of $H^{(j)}$ $(j = 0,4)$ corresponding to the observed spots. The Adam optimizer with weight decay of 1×10^{-5} is used to minimize the reconstruction loss. We set the initial learning rate to 5×10^{-3} and decreased to 1×10^{-3} with cosine annealing strategy[54]. The number of iterations is set to 5000 by default.

Finally, we use Voronoi tessellations[55] and rasterization to extend the embedding $H^{(4)}$ to all pixels of the sequenced tissue and get the tissue-wide low-resolution embedding $F^{\mathrm{low}} \in R^{h\times w\times M}$. Specifically, the entire tissue Ω is segmented into multiple multi-pixel tiles by expanding the central spatial coordinates of all spots using Voronoi tessellations. Then, for the m-th embedding dimension ($m = 1,2,\dots,M$), the corresponding tissue-wide low-resolution embedding intensity $F_{:,:,m}^{\mathrm{low}}$ is constructed by associating spot-wise embedding $H_{:,m}^{(4)}$ to their respective tiles.

High-resolution embedding learning. DeepfuseNMF combines the pretrained U-Net-DeepfuseNMF model with two modules, the feature fusion module (encoder) and the data reconstruction module (decoder), to learn the high-resolution embedding. ST image data are extremely high-resolution (usually with hundreds of millions of pixels), so directly learning embedding and reconstructing gene expression and image intensity for every pixel are computationally too expensive. DeepFuseNMF instead first divides the spatial transcriptomics data to small patches, learns the embeddings for the small patches and finally gets the pixel-wise high-resolution embedding by assembling all patches. To get the data patches, the sequenced tissue Ω is first segmented into tissue patches $\Omega^{(j)} \subset \Omega\ (j = 1, \dots, J)$ using sliding overlapping 256×256 rectangular with 15% overlapping between neighboring patches. The data patch $\mathcal{D}^{(j)}$ of j-th tissue patch $\Omega^{(j)}\ (j = 1, \dots, J)$ includes the normalized expression data $X^{(j)}$ at the spots within $\Omega^{(j)}$, the image data $I^{(j)} \in \mathbb{R}^{256\times256\times c}$ and the pre-estimated low-resolution embedding $\mathrm{F}^{(\mathrm{j,low})} \in \mathbb{R}^{256\times256\times M}$ within $\Omega^{(j)}$, i.e., $\mathcal{D}^{(j)} = \{\Omega^{(j)}, X^{(j)}, I^{(j)}, \mathrm{F}^{(\mathrm{j,low})}\}$.

For each data patch $\mathcal{D}^{(j)}$, the feature fusion module fuses the low-resolution embedding $\mathrm{F}^{(\mathrm{j,low})} \in \mathbb{R}^{256\times256\times M}$ and the high-resolution image features $Z^{(j)} \in \mathbb{R}^{256\times256\times32}$ from $I^{(j)}$ through U-Net-DeepFuseNMF to generate the patch embedding $\tilde{\mathrm{F}}^{(\mathrm{j})}$ of the same shape as $\mathrm{F}^{(\mathrm{j,low})}$ as follows:

$$\tau^{(j)} = f_{\mathrm{CNN}}\left(Z^{(j)}\right) \in \mathbb{R}^{256\times256\times M}, \tag{5}$$

$$\tilde{\mathrm{F}}^{(\mathrm{j})} = \mathrm{Sigmoid}[\tau^{(j)} + \mathrm{Logit}(\mathrm{F}^{(\mathrm{j,low})})], \tag{6}$$

where $f_{\mathrm{CNN}}(\cdot)$ is a two-layer convolutional neural network, $\mathrm{Sigmoid}(\cdot)$ and $\mathrm{Logit}(\cdot)$ are the element-wise Sigmoid function and Logit function, respectively.

The data reconstruction module takes the high-resolution patch embeddings $\tilde{\mathrm{F}}^{(\mathrm{j})}\ (j = 1, \dots, J)$ to reconstruct both the gene expressions and the image data simultaneously. To reconstruct the normalized gene expression profile of the n-th sequenced spot within the j-th data patch $\mathcal{D}^{(j)}$ ($S_n \cap \Omega^{(j)} \neq \emptyset$), DeepFuseNMF first aggregates the patch embeddings $\tilde{\mathrm{F}}^{(\mathrm{j})}$ at the pixels within the n-th spot to get the spot-level embeddings,

$$\hat{F}_{n,m}^{(j)} = \frac{1}{|S_n \cap \Omega^{(j)}|} \sum_{(x,y) \in S_n \cap \Omega^{(j)}} \tilde{F}_{x,y,m}^{(j)}, \quad m = 1,2,\dots,M, \tag{7}$$

where $|S_n \cap \Omega^{(j)}|$ is the number of pixels within the intersecting area between the n-th spot and the j-th tissue patch. If the j-th data patch $\mathcal{D}^{(j)}$ does not contain any sequenced spot, we simply ignore this patch in the reconstruction of the gene expression profiles. Then, DeepFuseNMF reconstructs the normalized gene expression profile of the n-th spot in the j-th tissue patch using a nonnegative linear combination of the spot embedding $\hat{F}_{n,m}^{(j)}$ $(m = 1,2,\dots,M)$ as follows:

$$\hat{X}_{n,g}^{(j)} = \sum_{m=1}^{M} \hat{F}_{n,m}^{(j)} \, \hat{W}_{m,g}, \quad g = 1,2,\dots,G, \tag{8}$$

where $\hat{X}_{n,g}^{(j)}$ is the reconstructed normalized gene expression profile of the g-th gene at the n-th spot from the j-th tissue patch, and $\hat{W} \in R^{M \times G}$ is a non-negative loading matrix that needs to be learned. The loss function for reconstructing the expression is taken as the following negative log-likelihood of the Poisson distribution:

$$\mathcal{L}_{\text{exp}} = \sum_{j=1}^{J} \frac{1}{|A_j| \cdot G} \sum_{n \in A_j} \sum_{g=1}^{G} \left(\hat{X}_{n,g}^{(j)} - X_{n,g}^{(j)} \log\left(\hat{X}_{n,g}^{(j)} + \epsilon \right) \right), \tag{9}$$

where $A_j = \{n: S_n \cap \Omega^{(j)} \neq \emptyset\}$ stands for spots belonging to the j-th tissue patch and $\epsilon = 1 \times 10^{-8}$.

For the reconstruction of the images $I^{(j)}$ of the j-th patch, DeepFuseNMF uses a two-layer convolution neural network and an element-wise Sigmoid function to get the reconstructed images $\hat{I}^{(j)}$ from the j-th patch embeddings $\tilde{F}^{(j)}$.The image reconstruction loss is taken as the mean squared error loss

$$\mathcal{L}_{\text{img}} = \sum_{j=1}^{J} \frac{1}{256 \cdot 256 \cdot c} \sum_{x=1}^{256} \sum_{y=1}^{256} \sum_{k=1}^{c} \left(\hat{I}_{x,y,k}^{(j)} - I_{x,y,k}^{(j)} \right)^2. \tag{10}$$

The overall reconstruction loss function is a weighted sum of the expression reconstruction loss and the image reconstruction loss

$$\mathcal{L} = \lambda \mathcal{L}_{exp} + (1 - \lambda)\mathcal{L}_{img}, \tag{11}$$

where $\lambda \in [0,1]$ is the weight of the two loss functions. This weight λ is set to $1/3$ as default. The Adam optimizer with a weight decay of 1×10^{-5} is used to minimize the overall reconstruction loss. The mini-batch size is set to 32 image patches, and the number of iterations is set to 2000 by default. In the first 1000 iterations, we fix the parameter of U-Net-DeepfuseNMF, only optimize the feature fusion module and the image reconstruction module with an initial learning rate of 5×10^{-3} and decreased to 1×10^{-6} with cosine annealing strategy. Then, in the last 1000 iterations, we fine-tune all parameters of DeepfuseNMF with a lower initial learning rate of 4×10^{-4} and decreased to 1×10^{-6} with cosine annealing strategy.

Finally, DeepFuseNMF outputs pixel-wise high-resolution embeddings $F^{high} \in \mathbb{R}^{h \times w \times M}$ by averaging the patch embeddings $\tilde{\mathrm{F}}^{(j)} (j = 1, \dots, J)$ that contain the pixel (x, y)

$$F_{x,y,m}^{high} = \frac{1}{\sum_{j=1}^{J} \mathbf{1}\left((x,y) \in \Omega^{(j)}\right)} \sum_{j=1}^{J} \left[\tilde{F}_{x,y,m}^{(j)} \cdot \mathbf{1}\left((x,y) \in \Omega^{(j)}\right)\right], \tag{12}$$

where $\mathbf{1}(\cdot)$ is the indicator function.

Ranking and identification of genes associated with embedding dimensions. DeepFuseNMF ranks genes' contribution to each high-resolution embedding dimensions $F_{:,:,m}^{\mathrm{high}}$ based on the nonnegative loading matrix $\widehat{W} \in R^{M \times G}$ $(\widehat{W} \geq 0)$ obtained from the expression reconstruction module. Specifically, the loading matrix $\widehat{W}$ is first column-normalized as $\widetilde{W} \geq 0$ with $\widetilde{W}_{m,g} = \widehat{W}_{m,g} / \left(\sum_{m=1}^{M} \widehat{W}_{m,g}\right)$. Then, for a given embedding dimension m, DeepFuseNMF ranks the genes by the decreasing order of $\widetilde{W}_{m,g}$ and takes top genes as the genes associated with this embedding dimension.

High-resolution spatial domain detection. DeepFuseNMF clusters pixels (x, y) based on the high-resolution embeddings $F^{high} \in \mathbb{R}^{h \times w \times M}$ to get the high-resolution spatial domains. Considering

that the number of pixels in tissue is extremely large, directly clustering the pixels is computational too expensive. Therefore, DeepFuseNMF first divides the whole tissue into many equal-sized squares, called meta-pixels. Then, DeepFuseNMF averages the embeddings of the pixels in the meta-pixels to get the meta-pixel embeddings and applies the k-means algorithm to cluster the meta-pixels. The initial value of the k-means algorithm is taken as the clustering results of the spot embeddings given by the Louvain clustering algorithm. Here, the spot embeddings are obtained using a procedure similar to the meta-pixel embeddings. The size of the meta-pixel is usually much smaller than sequenced spots and can be adjusted by users. Specifically, for the 10X Visium mouse brain coronal slices with IHC images, the size of the meta-pixel is 50 pixels, and the spot diameter is 178 pixels. For the four 10X Visium SHH medulloblastoma slices stained with H&E images, the meta-pixel size is 10 pixels, and the spot diameter is 118 pixels. For the two 10X Visium human CRC tumor slices with H&E images, the meta-pixel size is 20 pixels, and the spot diameter is 150 pixels. Lastly, for the human CRC slice from Stereo-seq, the meta-pixel size is 10 pixels, and the spot diameter is 100 pixels.

Ranking and identification of genes associated with detected high-resolution spatial domain. Based on the detected D high-resolution spatial domains, we first calculate the mean embedding intensity for each spatial domain $\bar{E} \in \mathbb{R}^{D \times M}$, and extract the nonnegative loading matrix for spatial domain $\widehat{W}^D \in \mathbb{R}^{D \times G}$ by $\widehat{W}^D = \bar{E}\,\widehat{W}$. The column-normalization and gene ranking way are similar to those in Section Ranking and Identification of Genes Associated with Embedding Dimensions.

High-resolution gene expression recovery. For the g-th gene, DeepFuseNMF recovers its high-resolution normalized expression $X_g^{\text{high}} \in \mathbb{R}^{h \times \text{w}}$ by weighted sum of the high-resolution embeddings F^{high},

$$X_g^{\text{high}} = \sum_{m=1}^{M} W_{m,g} F_{:,:,m}^{high} \ . \tag{13}$$

Note that X_g^{high} quantifies expression of the g-th gene for every pixel in the tissue and thus essentially imputes the sparse expression quantification given by the original ST expression data.

Joint analysis of multi-sample ST data. When there are multiple ST samples, the DeepFuseNMF framework is similar as before, but we introduce batch parameters to account for the potential batch effect across different samples. Specifically, let Ω_b be the sequenced tissue of b-th sample ($b = 1,2,\dots,B$). Similar to Equation (8), DeepFuseNMF reconstructs the normalized gene expression profile $\hat{X}_{n,:}^{(j,b)}$ of the n-th spot in the jth tissue patch from the b-th sample by employing the nonnegative linear combination component alongside an additional batch effect term as follows:

$$\hat{X}_{n,g}^{(j,b)} = \mathrm{ReLu}\left(\sum_{m=1}^{M} \hat{F}_{n,m}^{(j,b)}\ \widehat{W}_{m,g}\ + \gamma_{b,g}\right),\ \ g = 1,2,\dots,G. \tag{14}$$

Here, $\widehat{W} \in R^{M\times G}$ is a non-negative loading matrix to be learnt, $\gamma_{b,g} \in \mathbb{R}$ denotes the mean batch effect parameter of batch b on gene g. For identifiability, we set the first batch as the reference batch and $\gamma_{1,g} = 0$. Further, the low-resolution factor matrix $F^{(b)}$ of batch b derived from equation (1) is replaced by

$$\breve{F}_{n,m}^{(b)} = \ \mathrm{ReLu}\left(F_{n,m}^{(b)} - \frac{1}{N_b}\sum_{i}^{N_b} F_{i,m}^{(1)}\right), m = 1,\dots,M,$$

where N_b is the number of spots in the reference batch.

If the samples are from C biologically different experimental conditions (e.g. the Palbociclib-treated and control conditions in the SHH medulloblastoma dataset). Suppose for condition c, there are B_c batches. Then, we replace $\gamma_{b,g}$ in equation (14) with $\gamma_{c,b,g}$ ($c = 1,\dots,C;\ b = 1,\dots,B_c$), that is

$$\hat{X}_{n,g}^{(j,c,b)} = \mathrm{ReLu}\left(\sum_{m=1}^{M} \hat{F}_{n,m}^{(j,c,b)}\ \widehat{W}_{m,g}\ + \gamma_{c,b,g}\right),\ \ g = 1,2,\dots,G.$$

For identifiability, we set the first batch in each condition as the reference batch and let $\gamma_{c,1,g} = 0$. All parameters are trained using the Adam optimizer to minimize the overall reconstruction loss (11).

Generation and analysis of simulated datasets

Two methods are used to generate the simulation datasets. The first method utilizes an H&E image of a mouse cerebellum and generates ST expression data using Zero-inflated Poisson distributions, where the distribution parameters are dependent on the given H&E image. The second method uses a real single-cell resolution ST dataset sequenced by the 10X Xenium platform to generate simulation datasets.

Simulation data generation using an H&E image of mouse cerebellum. We choose the H&E image, with a size of 4000×4000 pixels, of a 10X Visium mouse posterior brain sagittal slice (https://www.10xgenomics.com/resources/datasets/mouse-brain-serial-section-1-sagittal-posterior-1-standard-1-1-0) for simulating ST expression data. The image is first segmented into four regions: the fiber tracts, granular layer of the cerebellum, molecular layer of the cerebellum, and the background area. We introduce $M = 3$ true high-resolution embedding dimensions, denoted as $F^* \in \mathbb{R}^{h \times w \times M}$, where $h = 4000$ and $w = 4000$ are the height and width of the image, respectively. Each of the dimensions corresponds to and is only enriched in one segmented region, excluding the background area.

We generate N spots with radius r using a 10X Visium-style meshgrid. Denote the pixel location set of the nth spot as S_n $(n = 1, \dots, N)$. Using the high-resolution embedding F^*, we induce the spot-level embedding dimension intensity $F^{*,\text{spot}}_{n,m}(n = 1, \dots, N;\ m = 1,2,3)$ as follows,

$$F^{*,\text{spot}}_{n,m} = \frac{F^{*,1}_{n,m}}{\sum_{m'=1}^{3} F^{*,1}_{n,m'}},$$

$$F^{*,1}_{n,m} = F^{*,0}_{n,m} + \varepsilon_{n,m}\ , F^{*,0}_{n,m} = \frac{1}{|S_n|} \sum_{(x,y) \in S_n} F^*_{x,y,m},$$

where $\varepsilon_{n,m} \sim \mathcal{N}(0, 0.05^2)$ is a Gaussian random variable that controls the spot-specific variation of the embedding intensity. Then, we generate the gene contribution matrix $W^* \in \mathbb{R}_+^{G \times M}$ as follows. We divide the G genes into 3 gene sets, and assign each gene set to one embedding as its associated gene set. The gene contribution value $W^*_{g,m}$ is generated from the Gamma distribution with a scale parameter of 0.05. The shape parameter of the Gamma distribution is set to 15 if the gene is associated with the dimension, and 5 otherwise.

Given the generated spot-level embedding dimension intensity $F^{*,\text{spot}}$ and the gene contribution matrix W^*, we first generate the true underlying gene expression $X_{n,g}$ for spot n and gene g as $X_{n,g} = \sum_{m=1}^{M} F_{n,m}^{*,\text{spot}} W_{g,m}^* + \epsilon$ where ϵ is generated from $\mathcal{N}(0, \sigma^2)$. Then, we generate the observed count expression $X_{n,g}^0$ for spot n and gene g from a Zero-inflated Poisson distribution. The zero-inflation parameter of the Zero-inflated Poisson distribution is π and the mean parameter is $\lambda = l_n \{exp\left(X_{n,g}\right) - 1\}$, where l_n represents spot-specific library size and is generated using $\log(l_n) \sim \mathcal{N}(10, 0.05^2)$. After generating the count expression data, we also randomly exclude $\rho\%$ spots to simulate different spot densities.

In the simulation, we set the number of genes as $G = 3000$. We vary the spot exclusion percentage ρ, the spot radius r, the spot expression variance σ, and the zero-inflation parameter inflation π to generate different simulation scenarios. For each scenario, we generate 30 replicates.

Simulation based on a real single-cell-level ST dataset. We use a single-cell-level Xenium ST dataset from a human breast carcinoma tissue[11] to generate the simulated spot-level ST expression profiles. Specifically, this Xenium ST dataset contained a ST expression profile for 313 genes in 167,780 sequenced cells, and a high-resolution H&E image with 5478×7526 pixels. To generate the simulated spot-level ST expression data, we first generate the spots S_n $(n = 1, \ldots, N)$ with radius r with a 10X Visium-style meshgrid and sum the expression counts of cells within each spot as the spot-level expression. Then, a proportion ρ of the generated spots are dropped along with their spatial coordinates and gene expression.

Considering that the number of annotated cell type for this Xenium dataset is 19, we set the number of embedding dimensions as $M = 20$ in this simulation. The underlying true high-resolution embedding is set as the embedding given by NMF on the original single-cell-level Xenium ST dataset. We vary the spot exclusion proportion ρ and the spot radius r to generate different simulation scenarios, and generate 30 replicates for each scenario.

Evaluation criteria of the embedding learning methods. We use the MAE between the true embedding intensity F^* and the inferred embedding intensity $\hat{F}$ to evaluate each method's performance on embedding learning. The MAE for one simulated dataset is defined as

$$\mathrm{MAE}(F^*, \hat{F}) = \frac{1}{h \times w \times M} \sum_{x=1}^{h} \sum_{y=1}^{w} \sum_{m=1}^{M} \left| F^*_{x,y,m} - \hat{F}_{\mathrm{x,y,m}} \right|,$$

and the MAE for a simulation scenario is defined as the mean of the above value over all 30 replicates of the simulation scenario. For methods that only give spot-level embeddings, we extend their spot-level embeddings to whole-tissue-wise embeddings by Voronoi tessellations and rasterization and then calculate their MAEs. To evaluate the spatial domain assignment in simulation data generation from mouse cerebellum, we cluster the pixels into three regions by applying K-means algorithm to the inferred embedding of each method and compare the inferred region label with the three true regions using the ARI metric.

Comparison with available methods

We compare DeepFuseNMF with embedding learning methods and spatial domain detection methods. The embedding learning methods includes NMF ("decomposition.NMF" function in python scikit-learn package (v1.5.1)), NSFH , SpiceMix, Giotto (v4.0.0) and TESLA (v1.2.4) +NMF and the spatial domain detection methods include iStar, stLearn (v0.4.12) and SpaGCN (v1.2.7). NMF, NSFH, SpiceMix and Giotto only need ST expression data, while iStar, stLearn, TESLA and SpaGCN need both ST expression data and histology image data. The available methods that utilize histology images all require RGB images as input. So, for datasets with H&E images, we directly provide the H&E images to these methods, while for datasets with IHC images, we convert the IHC images to RGB images and provide the obtained RGB images to these methods.

In terms of embedding learning, for TESLA+NMF, we first use TESLA[24] to enhance the spatial resolution of gene expression based on the high-resolution image and spatial expression profile. The default meta-pixel size of 50×50 pixels is used for TESLA. Then, we learn an embedding using

standard NMF based on the resolution-enhanced expression. For other embedding learning methods, their default parameters are used.

In terms of the spatial domain detection, for the embedding learning methods, we first apply these methods to obtain the embeddings. Then, for the spot-level embedding methods, we apply the Louvain algorithm[56] to their embeddings for spot clustering. For the high-resolution embedding methods including DeepFuseNMF and TESLA+NMF, we first apply the Louvain algorithm to cluster the spots and then apply the Kmeans algorithm to cluster the meta-pixels with the clustering results given by Louvain as the initial values. Across all embedding methods, we use the same resolution parameter for the Louvain algorithm and the same cluster number of the Kmeans algorithm for clustering. For the spatial domain detection methods, we employ their default parameters and specify the same number of clusters as that of the embedding methods for spatial domain identification.

Sample collection and library preparation of colorectal cancer

CRC-01 were sampled from the abdominal wall of a metastatic colorectal cancer patient and CRC-02 were sampled form the ovarian metastasis site of a colorectal cancer patient Formalin-fixed paraffin-embedded (FFPE) samples of CRC-01 and CRC-02 were sectioned into 5 μm by Leica HistoCore BIOCUT and were affixed to adhesive slides. Subsequently, the FFPE samples were used for Visium CytAssist Spatial Gene Expression Reagent Kits (10X Genomics, USA) following the manufacturers protocols and recommended third party reagents.

CRC-03 was sampled from a primary colorectal cancer patient. Fresh tissue samples of CRC-03 were obtained and embedded in Tissue-Teck OCT within 30 minutes of resection, with extra fluid being wiped out. The samples were then frozen in dry ice and transferred to a -80°C freezer for storage. Freezing and OCT-embedding of the target tissue samples were sectioned to 10μm by Leica CM1950 cryostat, and immediately attached to Stereo-seq chips. Then, the samples underwent formaldehyde fixation and permeabilization, cDNA synthesis, fluorescence imaging, and screening of the optimal permeabilization time based on fluorescence imaging intensity. The target tissue sections were fixed with formaldehyde, and stained using nucleic acid dye (Thermo fisher, Q10212). The sections were

permeabilized for RNA release at the selected time to capture the mRNA molecules from the target tissue cells on the microarray for cDNA synthesis. The indexed cDNA libraries were constructed and sequenced in paired-end.

Generation of the pseudo image for the colorectal cancer Stereo-seq dataset

The Stereo-seq dataset CRC-03 lacks paired histology images. To apply DeepFuseNMF and other high-resolution methods, we first generated a high-resolution pseudo image based on the subcellular expression profile of Stereo-seq data. Specifically, we first extracted the high-resolution bin-level expression counts using the "io.read_gef" function in Python Stereopy package with a bin size of 10. This bin size is significantly smaller than the bin size commonly used for Stereo-seq data analysis (i.e., bin size = 100). Next, the high-resolution bin-level expression counts were log transformed and normalized using total sum count as the library size. PCA was then performed on the high-resolution bin-level normalized expression. Finally, we summarized the top 30 PCs into three UMAP components and visualized the three-dimensional components using red, green, and blue (RGB) colors through an RGB plot. This process resulted in a pseudo image that captures the high-resolution spatial structure of the tissue (Fig. S19a).

Analysis of colorectal cancer samples

Cell type proportion deconvolution was performed using RCTD[40] in R package spacexr (v2.2.1). The reference single-cell RNA-seq data of RCTD was downloaded from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) under accession code: GSE183202[57]. To associate the spot-level deconvolution with our high-resolution clusters, we first computed the proportion of high-resolution cluster c in spot n, denoted as $\alpha_{c,n}$. Suppose the RCTD-inferred proportion of cell-type b in spot n was $\beta_{b,n}$. We estimated the proportion of cell-type b in high-resolution cluster c by $\sum_{n=1}^{N} \alpha_{c,n}\beta_{b,n}$.

To characterize different biological activities, we computed aggregated expression aggregated expression using the *AddModuleScore* function in the R package *Seurat* (v4.0.0), with the gene lists provided in Supplementary Table 3. Given the embeddings from DeepFuseNMF, we calculated the mean ranks of given gene lists in each high-resolution cluster. The gene lists include genes related to

CMS subtypes[38, 39], functional states of cancer cells from CancerSEA[58] and cell-type markers[43] in Supplementary Table 3. Copy number variation detection was carried out using the R package infercnv (v1.20.0, https://github.com/broadinstitute/inferCNV).

For survival analysis, we downloaded TCGA[42] expression and survival data using R package *cBioPortalData*[59-61]. For each gene, we standardized its expression to have a mean 0 and standard deviation 1 across samples. Given a list of genes, we first calculated their mean expression for each patient and the upper and lower quantiles of its mean expressions across the patients. Patients with a mean expression larger than the upper quantile were assigned to the high expression group, while patients with a mean expression smaller than the lower quartile were assigned to the low expression group. Kaplan-Meier curves and log rank tests were carried out with the *survfit* and *ggsurvplot* functions in R packages *survival* and *survminer,* respectively.

CMS classification was performed using the R package *CMSclassifier* (v1.0.0)[38]. Briefly, raw count matrix of RNA-seq was transformed by variance stabilizing transformation[62] (SCTransform). Then, gene identifiers were converted from gene Symbol to Entrez gene ids using the R package org.Hs.eg.db (genome-wide annotation for Human, v3.19.1). The processed matrix was used as input for the *classifyCMS* function with the option method "SSP".

Data description

The first simulation dataset was generated based on the H&E image of MPBS-01 from 10X Visium, while the second simulation dataset was generated from a human breast cancer ST dataset with an H&E image sequenced on 10X Xenium. The mouse brain datasets MBC-01, MBC-02, MBC-03, MPBS-01 and MPBS-02 were sequenced on the 10X Visium platform, with IHC images for MBC-01 and MBC-02 and H&E images for the other mouse brain datasets. The four SHH medulloblastoma PDOX datasets were sequenced on the 10X Visium platform and accompanied with H&E images, with their tumor/normal/interface annotations from the original paper[35]. CRC-01 and CRC-02 were sequenced on the 10X Visium CytAssist platform. CRC-03 was sequenced with Stereo-seq with a nearby slice stained by H&E. See Supplementary Table 1 for details.

The mouse brain annotations for mouse brain coronal and sagittal sections were downloaded from

the aligned slices (76 coronal section and 9th sagittal section) in the Allen Reference Atlas – Mouse Brain[63] (atlas.brain-map.org). The mouse brain region marker genes were also downloaded from the Allen Mouse Brain Atlas[27] (https://mouse.brain-map.org/). The marker genes of major cell types in the pia mater were collected from Pietila et al.[32]. The marker genes of mouse microglial cells, astrocytes, and endothelial cells were downloaded from the CellMarker 2.0 database (http://117.50.127.228/CellMarker/index.html). The gene signatures of SHH-A, SHH-B, SHH-C, E2F target genes, and neuronal differentiation were collected from Vo et al.[35]. Reference scRNA-seq of metastasized colorectal cancer samples were downloaded from the NCBI GEO database under accession code GSE183916.

Data availability

The ST datasets in the paper are publicly available at:

(1) 10X Visium mouse brain coronal slice data.

https://www.10xgenomics.com/resources/datasets/adult-mouse-brain-section-2-coronal-stains-dapi-anti-gfap-anti-neu-n-1-standard-1-1-0

https://www.10xgenomics.com/resources/datasets/adult-mouse-brain-section-1-coronal-stains-dapi-anti-neu-n-1-standard-1-1-0

https://www.10xgenomics.com/resources/datasets/mouse-brain-section-coronal-1-standard-1-0-0

(2) 10X Visium mouse posterior brain sagittal slice data.

https://www.10xgenomics.com/resources/datasets/mouse-brain-serial-section-1-sagittal-posterior-1-standard-1-1-0

https://www.10xgenomics.com/resources/datasets/mouse-brain-serial-section-2-sagittal-posterior-1-standard-1-1-0

(3) 10X Xenium human breast cancer data, replicate 1.

https://www.10xgenomics.com/products/xenium-in-situ/preview-dataset-human-breast

(4) 10X Visium sonic hedgehog medulloblastoma data.

https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-023-01185-4#availability-of-data-and-materials

Raw count matrices, images and spatial data from the colorectal cancer samples will be available at the time of publication. Details of the datasets analyzed in this paper are described in Table S1.

Code availability

An open-source implementation of the DeepFuseNMF algorithm can be accessed at https://github.com/XiDsLab/DeepFuseNMF. Codes to reproduce the analyses in this manuscript are available at Zenodo: https://doi.org/10.5281/zenodo.13690848.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (12425110, 12371286 and 11971039 to R.X.), the National Key R&D Program of China (2020YFE0204200 to R.X.), Science and Technology Project of Tianjin Binhai New Area Health Commission (2022BWKY016 to Y.Z.), and the Sino-Russian Mathematics Center. Part of the analysis was performed on the high-performance computing platform of the Center for Life Sciences (Peking University). We thank Quan Nguyen and Laura A. Genovesi for sharing the 10X Visium mouse brain sonic hedgehog (SHH) medulloblastoma data.

Author contributions

R.X. conceived the study. R.X. and J.J.X. supervised the study. J.T. formulated the method. J.T. and K.Q. developed the software. J.T. and Z.C. performed bioinformatic and statistical analyses and prepared figures and tables in the manuscript. S.H. helped develop biological interpretations. S.Y., B.Y. and H.M. conducted pathological annotations. Y.Z. prepared the samples. Y.H. performed the spatial transcriptomic sequencing. J.T., Z.C., X.H. and R.X. wrote the manuscript.

Ethics declarations

All human tissue samples were obtained from Tianjin Medical University Cancer Institute & Hospital. Before surgery at the center, all patients provided written informed consent to allow any excess tissue to be used for research studies. The study was approved by the Ethics Committee (ID: bc2022189).

Competing interests

J.J.X. and R.X. holds stock in GeneX Health Co., Ltd. All other authors declare that they have no competing interests.

References

1. Vickovic, S. et al. High-definition spatial transcriptomics for in situ tissue profiling. *Nat. Methods* **16**, 987-990 (2019).
2. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
3. Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
4. Huang, S. et al. Spatial transcriptomics: a new frontier in cancer research. *Clin. Cancer Bull.* **3**, 13 (2024).
5. Moses, L. & Pachter, L. Museum of spatial transcriptomics. *Nat. Methods* **19**, 534-546 (2022).
6. Rao, A., Barkley, D., França, G.S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211-220 (2021).
7. Erickson, A. et al. Spatially resolved clonal copy number alterations in benign and malignant tissue. *Nature* **608**, 360-367 (2022).
8. Carmona-Fontaine, C. et al. Metabolic origins of spatial organization in the tumor microenvironment. *Proc. Natl. Acad. Sci. USA* **114**, 2934-2939 (2017).
9. Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. Methods and applications for single-cell and spatial multi-omics. *Nat. Rev. Genet.*, 1-22 (2023).
10. Chen, A. et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **185**, 1777-1792. e1721 (2022).

11. Janesick, A. et al. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nat. Commun.* **14**, 8353 (2023).
12. Li, H. et al. A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics. *Nat. Commun.* **14**, 1548 (2023).
13. Pham, D. et al. Robust mapping of spatiotemporal trajectories and cell–cell interactions in healthy and diseased tissues. *Nat. Commun.* **14**, 7739 (2023).
14. Shang, L. & Zhou, X. Spatially aware dimension reduction for spatial transcriptomics. *Nat. Commun.* **13**, 7203 (2022).
15. Velten, B. et al. Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. *Nat. Methods* **19**, 179-186 (2022).
16. Townes, F.W. & Engelhardt, B.E. Nonnegative spatial factorization applied to spatial genomics. *Nat. Methods* **20**, 229-238 (2023).
17. Chidester, B., Zhou, T., Alam, S. & Ma, J. SpiceMix enables integrative single-cell spatial modeling of cell identity. *Nat. Genet.*, 1-11 (2023).
18. Dries, R. et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol.* **22**, 1-31 (2021).
19. DeTomaso, D. & Yosef, N. Hotspot identifies informative gene modules across modalities of single-cell genomics. *Cell Syst.* **12**, 446-456. e449 (2021).
20. Bao, F. et al. Integrative spatial analysis of cell morphologies and transcriptional states with MUSE. *Nat. Biotechnol.* **40**, 1200-1209 (2022).
21. Hu, J. et al. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods* **18**, 1342-1351 (2021).
22. Bergenstråhle, L. et al. Super-resolved spatial transcriptomics by deep data fusion. *Nat. Biotechnol.* **40**, 476-479 (2022).
23. He, B. et al. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat. Biomed. Eng.* **4**, 827-834 (2020).
24. Hu, J. et al. Deciphering tumor ecosystems at super resolution from spatial transcriptomics with TESLA. *Cell Syst.* **14**, 404-417. e404 (2023).
25. Zhang, D. et al. Inferring super-resolution tissue architecture by integrating spatial transcriptomics

with histology. *Nat. Biotechnol.*, 1-6 (2024).

26. Lee, D.D. & Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788-791 (1999).
27. Lein, E.S. et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168-176 (2007).
28. Batiuk, M.Y. et al. Identification of region-specific astrocyte subtypes at single cell resolution. *Nat. Commun.* **11**, 1220 (2020).
29. Zeng, H. et al. Spatially resolved single-cell translatomics at molecular resolution. *Science* **380**, eadd3067 (2023).
30. Decimo, I., Fumagalli, G., Berton, V., Krampera, M. & Bifari, F. Meninges: from protective membrane to stem cell niche. *Am. J. Stem Cells* **1**, 92 (2012).
31. Mapunda, J.A. et al. VE-cadherin in arachnoid and pia mater cells serves as a suitable landmark for in vivo imaging of CNS immune surveillance and inflammation. *Nat. Commun.* **14**, 5837 (2023).
32. Pietilä, R. et al. Molecular anatomy of adult mouse leptomeninges. *Neuron* **111**, 3745-3764. e3747 (2023).
33. Engelhardt, B., Vajkoczy, P. & Weller, R.O. The movers and shapers in immune privilege of the CNS. *Nat. Immunol.* **18**, 123-131 (2017).
34. Mastorakos, P. & McGavern, D. The anatomy and immunology of vasculature in the central nervous system. *Sci. Immunol.* **4**, eaav0492 (2019).
35. Vo, T. et al. Spatial transcriptomic analysis of Sonic hedgehog medulloblastoma identifies that the loss of heterogeneity and promotion of differentiation underlies the response to CDK4/6 inhibition. *Genome Med.* **15**, 29 (2023).
36. Hovestadt, V. et al. Resolving medulloblastoma cellular architecture by single-cell genomics. *Nature* **572**, 74-79 (2019).
37. Finn, R.S., Aleshin, A. & Slamon, D.J. Targeting the cyclin-dependent kinases (CDK) 4/6 in estrogen receptor-positive breast cancers. *Breast Cancer Res.* **18**, 1-11 (2016).
38. Guinney, J. et al. The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350-1356 (2015).
39. Khaliq, A.M. et al. Refining colorectal cancer classification and clinical stratification through a single-cell atlas. *Genome Biol.* **23**, 113 (2022).

40. Cable, D.M. et al. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat. Biotechnol.* **40**, 517-526 (2022).

41. Chakravarthy, A., Khan, L., Bensler, N.P., Bose, P. & De Carvalho, D.D. TGF-β-associated extracellular matrix genes link cancer-associated fibroblasts to immune evasion and immunotherapy failure. *Nat. Commun.* **9**, 4692 (2018).

42. Hoadley, K.A. et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173**, 291-304. e296 (2018).

43. Zhang, L. et al. Single-cell analyses inform mechanisms of myeloid-targeted therapies in colon cancer. *Cell* **181**, 442-459. e429 (2020).

44. Dong, K. & Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat. Commun.* **13**, 1739 (2022).

45. Li, J., Chen, S., Pan, X., Yuan, Y. & Shen, H.-B. Cell clustering for spatial transcriptomics data with graph neural networks. *Nat. Comput. Sci.* **2**, 399-408 (2022).

46. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053-1058 (2018).

47. Tian, T., Wan, J., Song, Q. & Wei, Z. Clustering single-cell RNA-seq data with a model-based deep learning approach. *Nat. Mach. Intell.* **1**, 191-198 (2019).

48. Caligiuri, G. & Tuveson, D.A. Activated fibroblasts in cancer: Perspectives and challenges. *Cancer Cell* **41**, 434-449 (2023).

49. Smillie, C.S. et al. Intra-and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell* **178**, 714-730. e722 (2019).

50. Li, H., Calder, C.A. & Cressie, N. Beyond Moran's I: testing for spatial dependence based on the spatial autoregressive model. *Geogr Anal* **39**, 357-375 (2007).

51. Moran, P.A. Notes on continuous stochastic phenomena. *Biometrika* **37**, 17-23 (1950).

52. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234-241 (2015).

53. Chen, L., Lu, X., Zhang, J., Chu, X. & Chen, C. Hinet: Half instance normalization network for image restoration. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

Recognition, 182-192 (2021).

54. Loshchilov, I. & Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. *International Conference on Learning Representations*, 1-16 (2022).
55. Voronoï, G. Nouvelles applications des parametres continus a la theorie des forms quadratiques. Deuxieme memoire: Recheches sur les parallelloedres primitifs. *J. Reine Angew. Math.* **134**, 198-287 (1908).
56. Blondel, V.D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).
57. Lenos, K.J. et al. Molecular characterization of colorectal cancer related peritoneal metastatic disease. *Nat. Commun.* **13**, 4443 (2022).
58. Yuan, H. et al. CancerSEA: a cancer single-cell state atlas. *Nucleic Acids Res.* **47**, D900-D908 (2019).
59. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401-404 (2012).
60. de Bruijn, I. et al. Analysis and visualization of longitudinal genomic and clinical data from the AACR project GENIE biopharma collaborative in cBioPortal. *Cancer Res.* **83**, 3861-3867 (2023).
61. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013).
62. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
63. Allen Institute for Brain Science. Allen Reference Atlas – Mouse Brain [brain atlas] (2011). Available from atlas.brain-map.org.

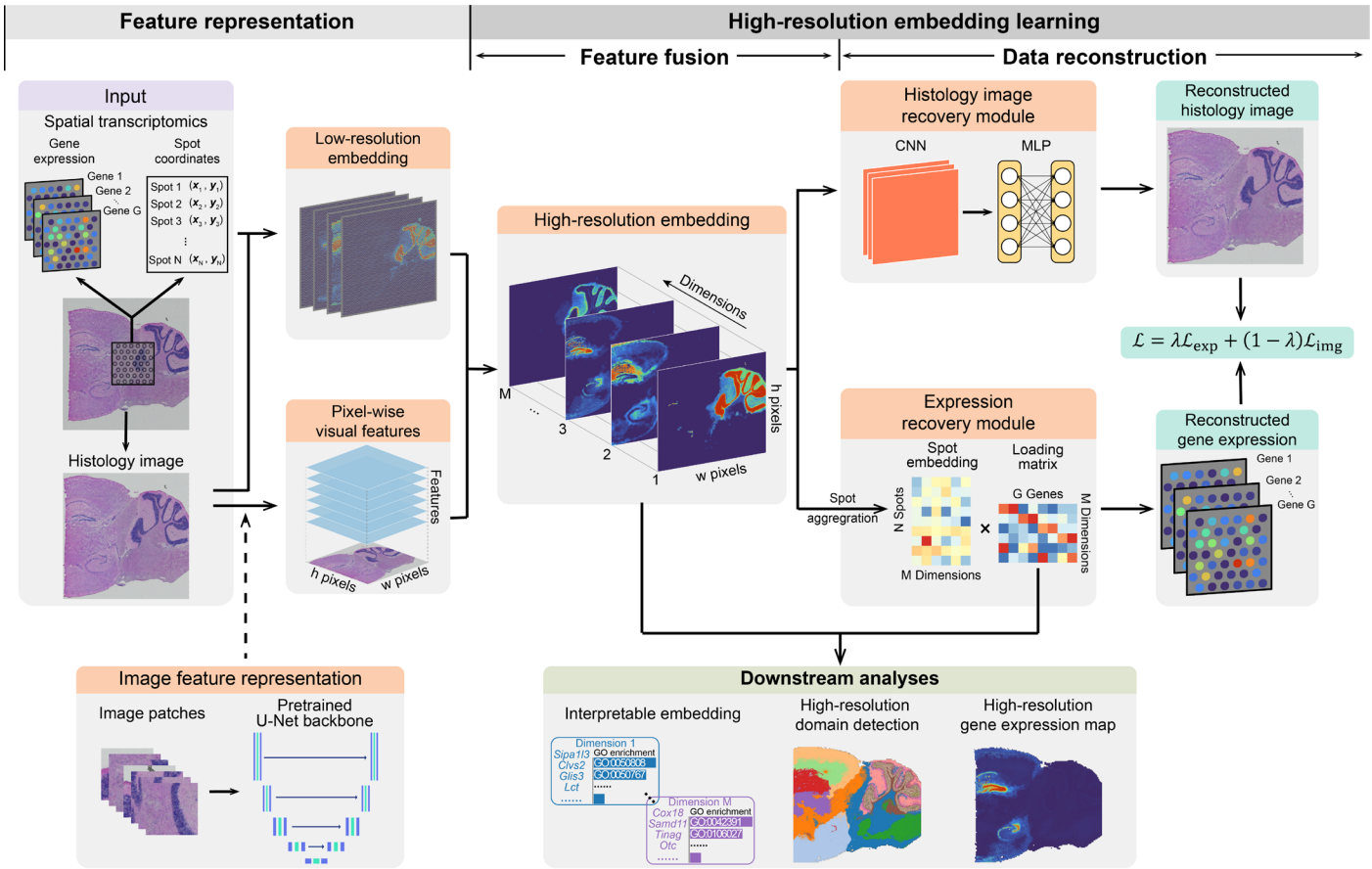

Fig. 1. Overview of DeepFuseNMF. DeepFuseNMF consists of two components: the feature representation and high-resolution embedding learning components. In the feature representation component, DeepFuseNMF independently learns a low-resolution embedding from the ST data and a high-resolution pixel-wise feature representation from the histology image data. In the high-resolution embedding learning component, DeepFuseNMF integrates the low-resolution embedding and the pixel-wise image features to learn a high-resolution embedding space. This is achieved by reconstructing both the gene expression data and the image data from the learned high-resolution embedding using an auto-encoder framework. With the learned high-resolution embedding, DeepFuseNMF can perform various downstream analyses, including identification of genes associated with the learned embeddings, high-resolution domain detection and gene expression recovery.

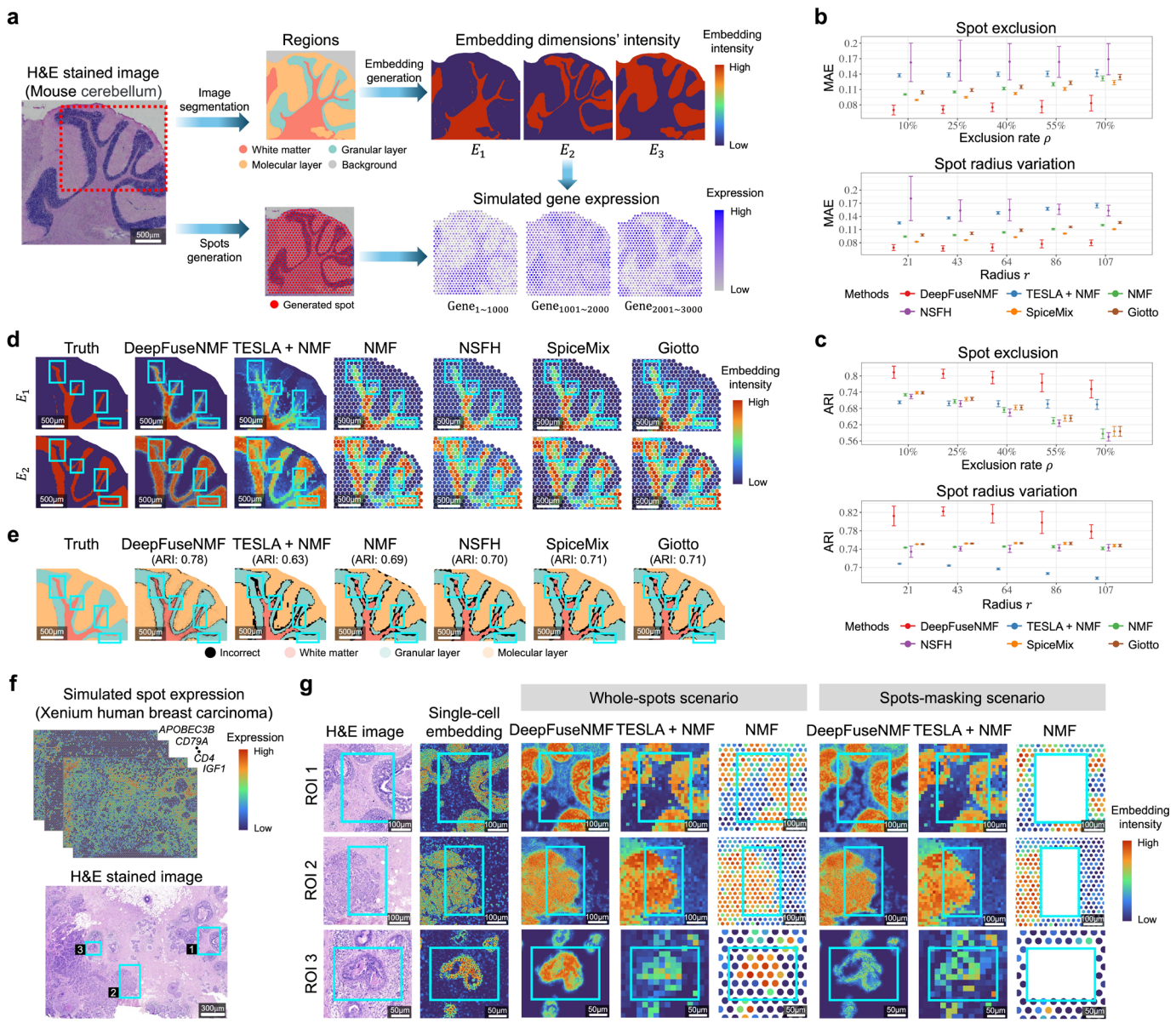

Fig. 2. Comparison of DeepFuseNMF with embedding learning and spatial clustering methods using simulation. a. Generation of the first simulation dataset. The H&E image of a mouse cerebellum sample is manually segmented into three brain regions and a background region. Spot locations are randomly generated on the tissue. Each brain region is then associated with a specific embedding dimension, and spot-wise ST data are generated based on the location of sequenced spots and the embedding dimensions. Scale bar: 500 μm. **b-c**. Error bar plots showing (b) the MAEs between the true and inferred embeddings given by different methods under various simulation scenarios, and (c) the ARIs between clustering results given by different methods and the true spatial domains. Error bars represent the mean $\pm$ standard deviation (SD). **d-e.** Zoom-in views of the red-box region highlighted in panel (a). Scale bar: 500 μm **d.** The local true embeddings E1 and E2 and corresponding embeddings estimated by different methods. Spatial intensities of each embedding dimension are linearly scaled to $[0,1]$ for visualization. **e.** The local clustering results given by different methods. The numbers on the top of the subfigures are the ARIs in this local region. Black color represents regions that are incorrectly clustered. **f.** Heatmaps showing examples of the second simulation data generated from a human breast carcinoma sample and the corresponding H&E histology image. Scale bar: 300 μm. **g.** Zoom-in views of the inferred embedding dimensions learned by various methods from one of the second simulation datasets (spot radius: 30 pixels) in the three selected ROIs in panel (f). Scale bar of ROI 1-3: 100, 100, and 50 μm. Whole-spot scenario means that expression of all spots of the simulation dataset are used for embedding learning, and the spot-masking scenario means that the ST expressions in the light blue boxes are masked for embedding learning.

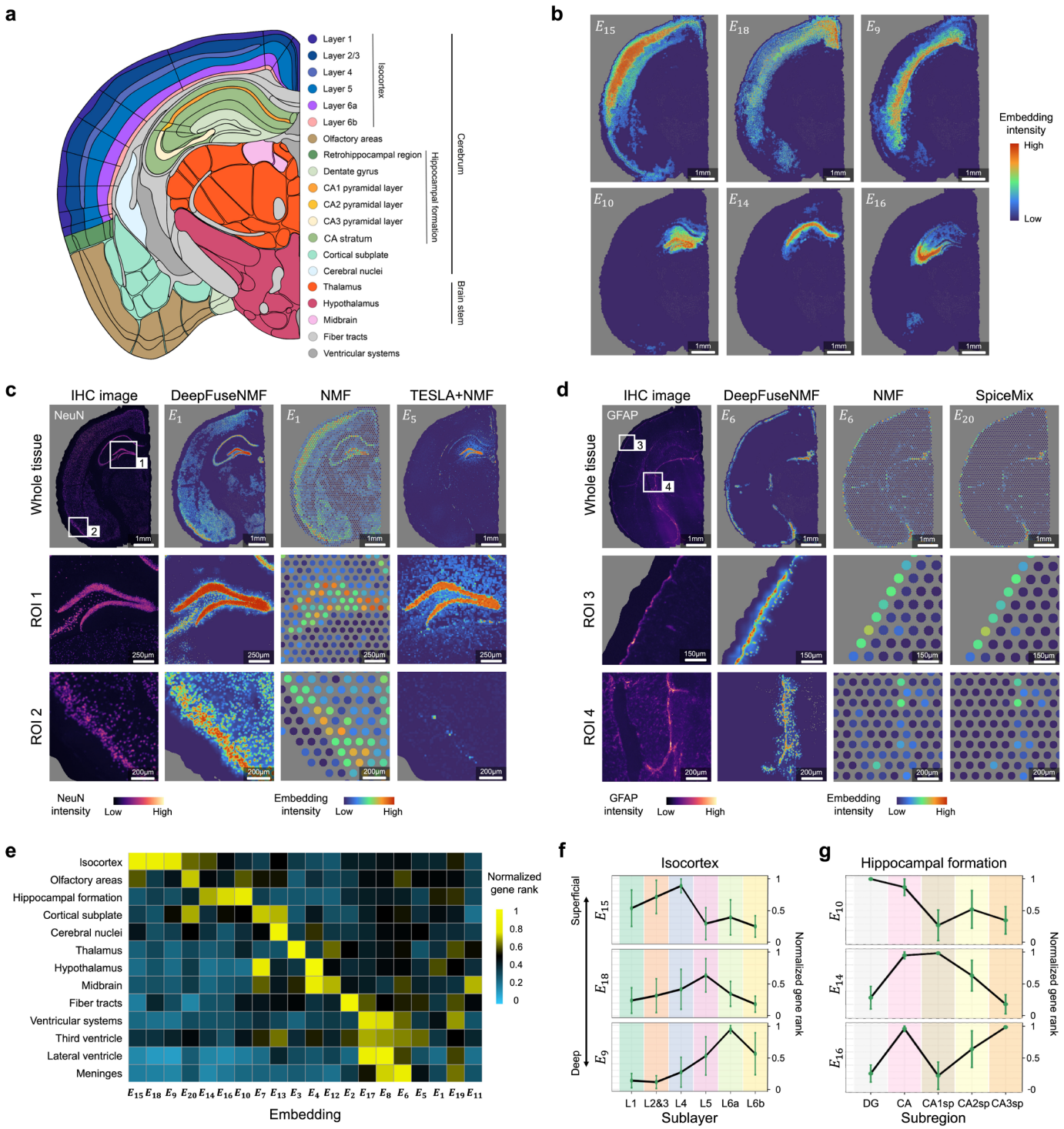

Fig. 3. Embeddings learned from the adult mouse brain coronal section data MBC-01. a. Anatomical annotations from the Allen Reference Atlas – Mouse Brain. **b**. Representative embedding dimensions identified by DeepFuseNMF including the isocortex-associated dimensions (E_{15}, E_{18}, E_9) and hippocampal-associated dimensions (E_{10}, E_{14}, E_{16}). **c-d**. Embedding dimensions associated with NeuN or GFAP protein expressions inferred by DeepFuseNMF, TESLA+NMF, NMF and SpiceMix and their zoom-in views in four selected ROIs. Scale bar of whole image: 1mm. Scale bar of ROI 1-4: 250, 200, 150 and 200 μm. **e**. A heatmap showing the mean rankings of brain-region marker genes in each embedding dimension. Gene rankings are normalized to [0,1]. In each embedding dimension, the genes are ranked increasingly by their corresponding values in the loading matrix. **f-g**. Error bar plots showing gene rankings of marker genes of subregions of isocortex and hippocampal formation in associated embedding dimensions. Error bars: median ± median absolute deviation (MAD, a robust SD). DG: dentate gyrus; CA: Cornu ammonis; CA1sp: pyramidal layer of field CA1; CA2sp: pyramidal layer of field CA2; CA3sp: pyramidal layer of field CA3.

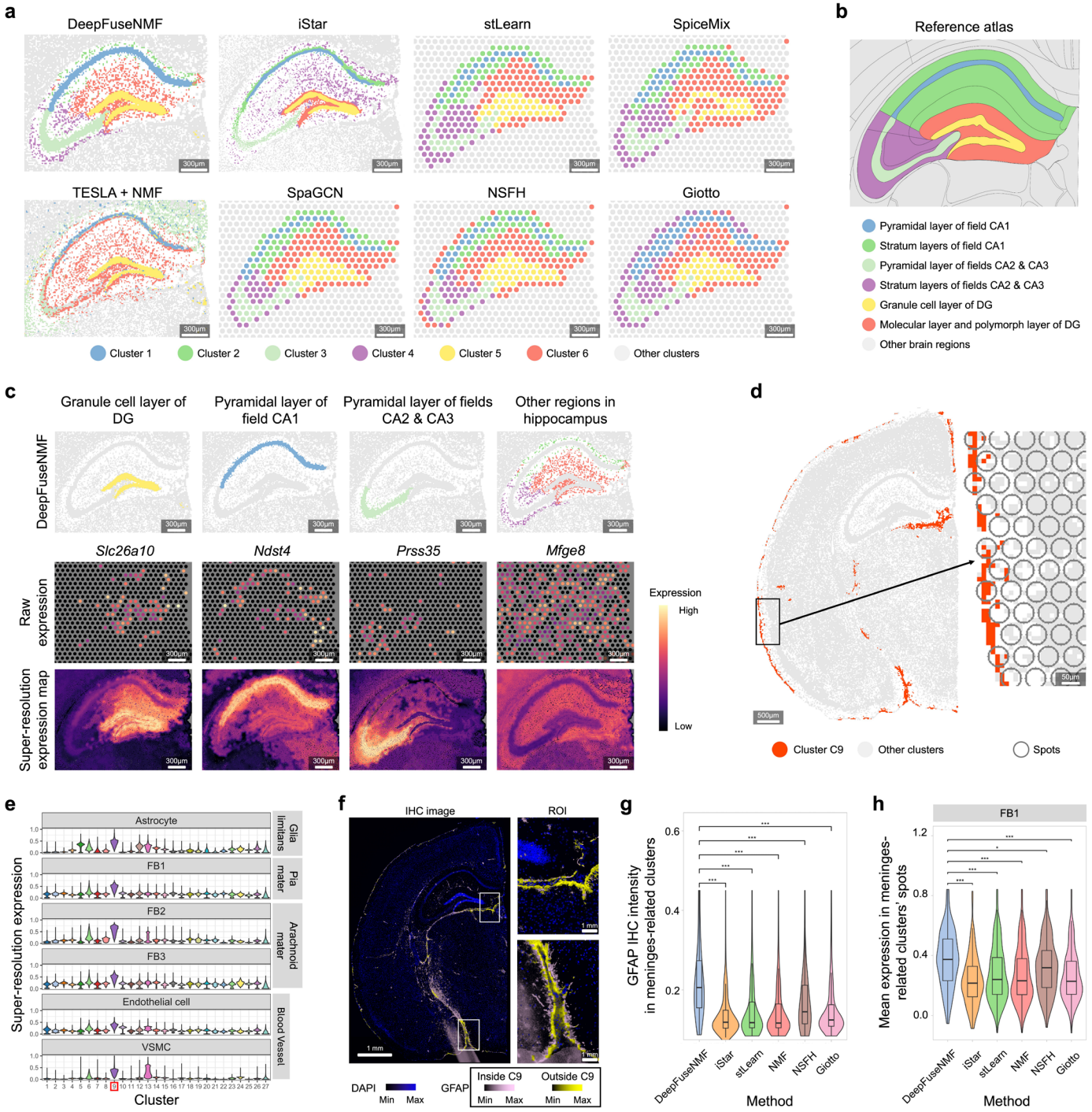

Fig. 4. Spatial domains detected from the adult mouse brain coronal section data MBC-01. a. Zoom-in views of clusters identified by different methods in the hippocampal formation. The spatial clusters were generated by Louvain clustering with resolution = 2.0 on the embeddings learned by different methods, except iStar which used the same number of clusters as DeepFuseNMF. Scale bar: 300 µm. **b**. Anatomical annotations from the Allen Reference Atlas – Mouse Brain hippocampal formation. DG: dentate gyrus; CA: Cornu ammonis. **c**. The raw expression and high-resolution expression recovered by DeepFuseNMF for the marker genes of hippocampal formation subregions. Scale bar: 300 µm. **d**. A refined cluster C9 located around the meninges. Scale bar: 500 µm. Right panel: The zoom-in view of an ROI showing that this cluster is often smaller than a spot. Scale bar: 50 µm. **e.** Violin plots showing the recovered high-resolution expression of marker genes of major cell types in glial limitans and leptomeninges across different spatial clusters. **f.** The IHC image of GFAP and 4',6-diamidino-2-phenylindole (DAPI). Blue: DAPI; Yellow: GFAP in cluster C9; Pink: GFAP in other clusters. Right panels: zoom-in views of two ROIs. Scale bar: 1 mm. **g.** Violin plots of the IHC intensity levels of GFAP in meninges-related clusters identified by different methods. **h.** Violin plots of mean expressions of FB1's marker genes in meninges-related clusters' spots identified by different methods. P-values are obtained using the two-sided Wilcoxon rank-sum test. *: p-value ≤ 0.05; ***: p-value ≤ 0.001.

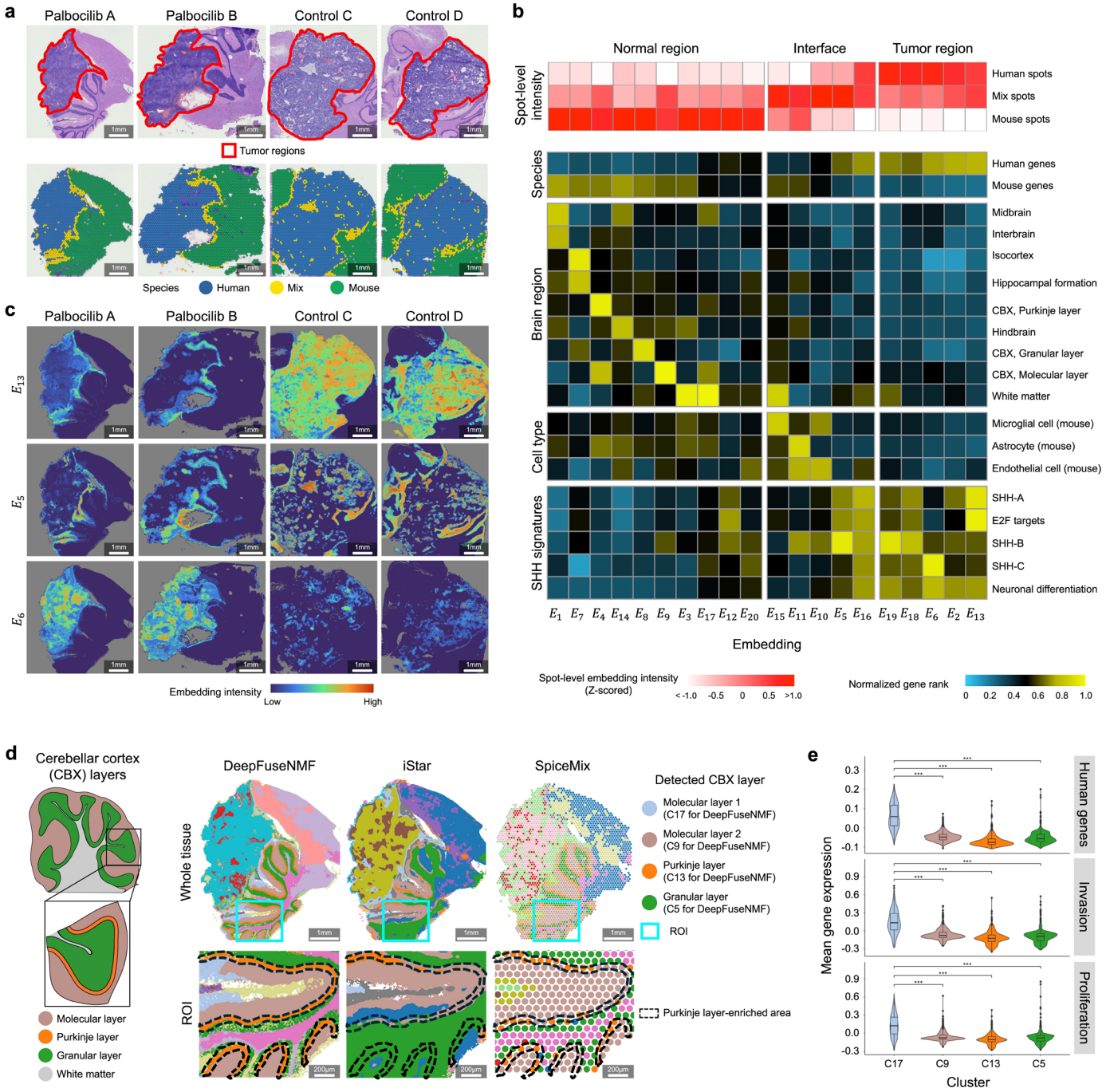

Fig. 5. Embedding and clustering results of the four PDOX samples. a. Upper panel: The H&E images with annotated tumor regions. Lower-panel: Spot locations along with their species classifications based on their expressions of human and mouse genes. Scale bar: 1mm. **b**. Rows 1-3 showing the 95th percentile of spot-level embedding intensity for different species and subsequent rows displaying the mean rankings of various gene signatures across each embedding dimension. Gene rankings were normalized to [0,1]. **c.** The embedding dimensions (E_6, E_{13}, E_5) associated with SHH transcriptional signatures. Scale bar: 1mm. **d**. Left panel: anatomical annotations from the Allen Reference Atlas – Mouse Cerebellum Cortex. Right panel: the spatial clusters identified by DeepFuseNMF, iStar, and SpiceMix in the Palbocliib A section. For SpiceMix, its filtered low-quality spots are colored in grey, and spot radius is enlarged 1.3-fold for better visualization. Scale bar: 1mm. Bottom panels: Zoom-in views of the ROI in the light blue box of the top panel. Scale bar: 200 μm. **e.** Violin plots of the mean normalized expressions of human genes, tumor invasion genes, and tumor proliferation genes across spots dominated by clusters identified by DeepFuseNMF. P-values are obtained using the two-sided Wilcoxon rank-sum test. ***: p-value ≤ 0.001.

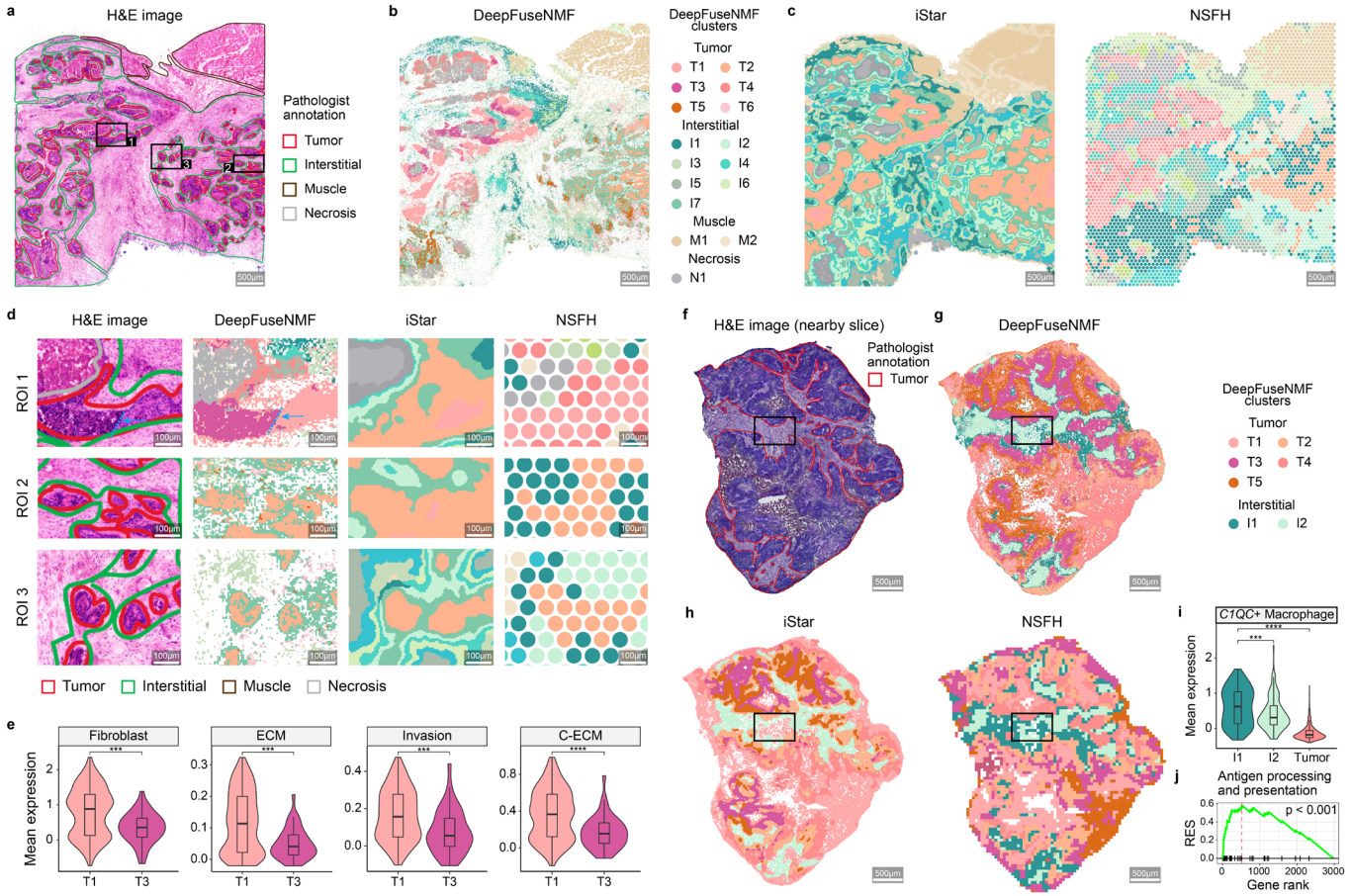

Fig. 6. Analysis of colorectal cancer samples. a-f. Analysis of sample CRC-01. **a.** The H&E image of sample CRC-01 with annotation by our pathologist. Annotated tumor regions, interstitial regions, muscle regions and necrosis regions are outlined in red, green, brown and grey, respectively. Scale bar: 500 μm. **b.** Spatial clusters identified by DeepFuseNMF. Spatial clusters are annotated by their marker genes (see Fig. S16). Scale bar: 500 μm. **C.** Spatial clusters identified by iStar and NSFH. For NSFH, its filtered low-quality spots are colored in grey, and spot radius is enlarged 1.2-fold for better visualization. Scale bar: 500 μm. **d.** Zoom-in views of H&E images and clusters identified by DeepFuseNMF, iStar and NSFH. ROIs correspond to black-box regions in (a). Scale bar: 100 μm. **e.** Violin plots of mean gene expressions of fibroblast, ECM, cancer-invasion and C-ECM signatures. **f-j.** Analysis of CRC-03. **f.** The H&E image of a nearby slice of sample CRC-03 with pathologist annotation. Annotated tumor regions are outlined in red. Other regions are all interstitial regions. Scale bar: 500 μm. **g.** Spatial clusters identified by DeepFuseNMF. Scale bar: 500 μm. **i.** Spatial clusters identified by iStar and NSFH. For NSFH, low-quality spots are colored in grey, and spot radius is enlarged 1.2-fold for better visualization. Scale bar: 500 μm. **i.** Violin plots of aggregated gene expressions of *C1QC+* macrophage markers. **j.** Gene set enrichment analysis (GSEA) on cluster I1-related genes for antigen processing and presenting. The enrichment plots contain profiles of the running enrichment scores (RES) and positions of gene set members on the rank ordered list in GSEA. P-values are obtained using the two-sided Wilcoxon rank-sum test. ***: p-value ≤ 0.001; ****: p-value ≤ 0.0001.