ARTICLE IN PRESS

Communications Biology

https://doi.org/10.1038/s42003-026-09841-2

Article in Press

HarveST uses a heterogeneous graph learning framework to reveal spatial transcriptomics patterns

Received: 3 August 2025

Accepted: 27 February 2026

Cite this article as: Feng, J., Yu, T., Zhang, Y. HarveST uses a heterogeneous graph learning framework to reveal spatial transcriptomics patterns. *Commun Biol* (2026). https://doi.org/10.1038/s42003-026-09841-2

Junning Feng, Tianwei Yu & Yanlin Zhang

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

ARTICLE IN PRESS

HarveST uses a Heterogeneous Graph Learning Framework to reveal Spatial Transcriptomics Patterns

Junning Feng[1], Tianwei Yu[2,*], and Yanlin Zhang[1,*]

[1]Data Science and Analytics Thrust, Hong Kong University of Science and Technology (Guangzhou), Guangzhou, Guangdong 511453, China
[2]School of Data Science, Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), Shenzhen, Guangdong 518172, China
[*]Correspondence: yutianwei@cuhk.edu.cn, yanlinzhang@hkust-gz.edu.cn

Abstract

Spatial transcriptomics enables in situ gene expression profiling, yet precise spatial domain identification and marker gene detection remain challenging. We present HarveST, a heterogeneous graph-based framework that integrates spatial, transcriptomic, and gene-gene interaction data through a unified computational model. HarveST employs dual learning strategies: self-supervised learning for feature extraction and partially supervised refinement for domain delineation. Additionally, it implements a Random Walk with Restart algorithm for identifying spatial domain-marker spatially variable genes (SVGs). Applied to human cortical tissue, mouse olfactory bulb, and tumor microenvironments across multiple platforms, HarveST demonstrates superior performance in detecting biologically meaningful spatial domains and associated marker genes. HarveST further supports joint analysis across consecutive spatial transcriptomics sections, enabling consistent reconstruction of functional domains across slices. By capturing both spatial topology and molecular relationships in a single graph-theoretical framework, HarveST advances spatial transcriptomics analysis beyond conventional clustering approaches, offering deeper insights into tissue architecture and cellular interactions in normal and pathological contexts.

KEYWORDS

Spatial transcriptomics, Heterogeneous Graph Learning, Random Walk, Spatially Specific Genes

INTRODUCTION

In multicellular organisms, cells are spatially organized into functional communities, forming intricate tissue architectures that govern physiological and pathological processes[1]. Understanding this spatial organization is fundamental for deciphering tissue function, disease progression, and cellular interactions[2,3]. The advent of spatial transcriptomics (ST) technologies has revolutionized this field by enabling the simultaneous measurement of gene expression and spatial coordinates, providing unprecedented insights into tissue organization in both healthy and diseased states[4–6]. ST data analysis involves two crucial steps: spatial domain identification, which involves delineating distinct regions where cells exhibit coherent gene expression patterns and spatial proximity, and spatial domain-marker spatially variable gene (SVG) detection, which aims to identify genes that define and characterize these spatial domains.

Accurate identification of spatial domains is crucial for understanding tissue development, disease progression, and microenvironmental interactions. Existing computational approaches

ARTICLE IN PRESS

for spatial domain identification can be broadly categorized into non-spatial clustering and spatial-aware clustering methods. Non-spatial clustering approaches, such as K-means, Louvain[7], and Seurat[8], rely exclusively on gene expression profiles, ignoring spatial information. While computationally efficient, these methods often yield fragmented clusters that lack biological consistency, particularly in tissues with complex spatial architectures. Spatial-aware clustering methods explicitly incorporate spatial dependencies, resulting in clusters that better reflect tissue morphology and improve biological interpretability. Among spatial-aware methods, probabilistic models such as BayesSpace[9] utilize hidden Markov random fields to enforce spatial smoothness, whereas deep learning-based approaches leverage graph neural networks to learn spatially informed embeddings. For example, SEDR[10] and Stagate[11] employ graph convolutional networks to integrate spatial and transcriptional information. More recently, contrastive learning frameworks such as GraphST[12] and SpaceFlow[13] have been introduced to enhance feature discrimination by contrasting spatially similar and dissimilar regions. Furthermore, methods like stDyer[14] utilize Gaussian Mixture Variational Autoencoders with dynamic graph learning for improved boundary detection, MNMST[15] employs multi-layer networks to enhance scalability, and PearlST[16] incorporates partial differential equation (PDE)-based adversarial learning to better characterize spatial structures. In addition, recent frameworks such as stKeep[17], Impeller[18], and CellCharter[19] further advance spatial modeling through multimodal integration and improved robustness and comparability across samples.

Despite these advancements, several critical limitations remain. Most existing methods rely on predefined spatial graphs, which assuming that neighboring spots share similar transcriptional profiles and thus belong to the same domain. This assumption can be problematic at domain boundaries, where transcriptional heterogeneity is high, potentially leading to misclassification and spatial discontinuity[20,21,21]. Additionally, these methods often struggle with accurately identify rare spot populations and small-scale tissue structures, as purely unsupervised learning approaches tend to favor dominant clusters, potentially overlooking minor but biologically relevant regions. This is particularly relevant in complex tissues such as tumor microenvironments, where immune-infiltrating cells or subtle pathological structures may be missed.

Beyond spatial domain identification, a second key challenge in ST analysis is linking these domains to biologically meaningful gene expression patterns by identifying spatial domain-marker SVGs, genes that exhibit substantially higher expression in specific spatial domains. These genes are crucial for characterizing functional tissue regions, identifying region-specific molecular programs, and facilitating tissue annotation across diverse biological contexts[22]. Methods such as Trendsceek[23], SpatialDE[24], and SPARK[25] detect overall SVGs, which are genes with spatially variable expression patterns. However, these methods do not explicitly associate SVGs with specific spatial domains, limiting their interpretability. More recent approaches, such as Scanpy[26], SpaGCN[27], and DESpace[28], attempt to address this limitation by identifying spatial domain-marker SVGs. However, these approaches still have substantial drawbacks. For example, Scanpy and SpaGCN use the Wilcoxon rank-sum test to compare gene expression between spatial domains, while DESpace employs a negative binomial regression model. These methods typically assume that spatial domain-marker SVGs can be inferred independently for each gene, disregarding gene-gene interactions and co-expression networks that drive tissue organization. Given that biological functions arise from coordinated gene activity, failing to account for these relationships may overlook key regulatory programs that define spatial domains and obscure the mechanisms underlying tissue architecture.

Here, we propose HarveST, a weighted heterogeneous graph-based learning framework that integrates spatial domain identification and spatial domain-marker SVGs detection within a unified approach. Unlike conventional spatial graphs that rely solely on spatial proximity, HarveST constructs a heterogeneous graph that simultaneously incorporates spatial location,

ARTICLE IN PRESS

gene expression profiles, and gene-gene mutual information, thereby capturing intricate spatial dependencies and transcriptional interactions. Unlike approaches based on curated ligand-receptor (L-R) pairs, HarveST models statistical dependencies among genes in a fully data-driven manner, thus encompassing both known and potentially novel axes of cell-cell communication.

For spatial domain identification, HarveST employs a dual learning strategy that combines self-supervised feature extraction with partially supervised refinement. Specifically, an autoencoder first learns a low-dimensional representation of the heterogeneous graph, preserving key biological features while reducing noise. This representation is subsequently refined through a pseudo-labeling mechanism, where high-confidence spatial spots are assigned model-generated intermediate labels to serve as anchor points for classifier training. This enhances clustering robustness, particularly in resolving ambiguous boundary regions and identifying underrepresented spotpopulations. For spatial domain-marker SVG detection, HarveST introduces a Random Walk with Restart (RWR) algorithm on the heterogeneous graph to quantify the association between genes and spatial domains. In contrast to existing methods that depend on independent statistical tests, RWR propagates signals from domain-specific seed spots, enabling the identification of genes that exhibit strong spatial specificity without complex statistical adjustments. Beyond single-section analysis, HarveST is further extended to jointly model multiple tissue slices within a unified framework, enabling the identification of recurrent spatial domains across adjacent sections.

To evaluate the effectiveness of HarveST, we benchmarked its performance against 14 alternative methods. HarveST consistently outperformed existing approaches in both spatial domain identification and spatial domain-marker SVGs detection, achieving superior results across a comprehensive suite of nine quantitative metrics that assess global clustering agreement and boundary fidelity, while simultaneously demonstrating enhanced biological interpretability. For example, in the dorsolateral prefrontal cortex (DLPFC), HarveST accurately delineated the six-layered cortical structure with improved boundary precision. In the mouse olfactory bulb (MOB), it successfully reconstructed the stratified laminar organization across different spatial resolutions. Furthermore, in breast and pancreatic cancer datasets, HarveST provided a refined characterization of tumor microenvironments, capturing fine-grained spatial variations that were overlooked by existing methods. Importantly, the genes identified by HarveST as spatially specific were strongly enriched in functionally relevant pathways, further validating the biological importance of its results. These findings demonstrate that HarveST provides a robust and accurate framework for spatial transcriptomics analysis, surpassing existing methods in both spatial domain identification and spatial domain-marker SVGs detection.

RESULTS

Overview of HarveST

To characterize the spatial organization of tissues and identify spatial domain-marker SVGs, we developed HarveST, a heterogeneous graph-based computational framework. The overall workflow of HarveST is illustrated in Fig.1, which consists of three major components: heterogeneous graph construction, spatial domain identification, and spatial domain-marker SVGs detection.

HarveST constructs a heterogeneous graph that integrates multi-scale spatial and transcriptional information (Fig.1A). Unlike conventional spatial graphs that rely solely on spatial proximity, the heterogeneous graph simultaneously models three types of biological

ARTICLE IN PRESS

relationships: spot-spot interactions based on spatial proximity, spot-gene interactions derived from gene expression profiles, and gene-gene interactions inferred from mutual information. This graph representation provides a richer and more biologically meaningful structure for spatial transcriptomics analysis. Interestingly, several high-score edges corresponded to canonical ligand-receptor pairs, suggesting that HarveST's MI-based graph implicitly reflects aspects of cell-cell communication.

To delineate spatial domains, HarveST employs a dual-learning strategy that combines self-supervised feature extraction with partially supervised refinement. Initially, a heterogeneous graph attention encoder-decoder learns latent feature representations for both spots and genes. This unsupervised feature extraction phase preserves key biological structures while mitigating noise (Fig.1B). The learned representations are then clustered using a Gaussian mixture model (GMM) to assign pseudo-labels to high-confidence spots. These pseudo-labeled spots are subsequently used to train a classifier, which refines the clustering assignments and enhances spatial domain delineation. This approach improves clustering robustness, particularly at domain boundaries where transcriptional heterogeneity is high, and enhances the detection of minor spot populations that may be overlooked by purely unsupervised methods.

To identify spatial domain-marker SVGs, HarveST applies a RWR algorithm on the heterogeneous graph (Fig.1C). This process begins by selecting seed spots from a given spatial domain and propagating information through the graph structure. Unlike traditional statistical approaches that evaluate each gene independently, RWR inherently integrates spatial and transcriptional relationships, allowing for the identification of genes that exhibit strong spatial and biological specificity. The steady-state probabilities obtained from RWR quantify the association between genes and spatial domains, providing a biologically meaningful ranking of spatially specific genes. This approach eliminates the need for complex statistical comparisons and ensures that identified genes are functionally relevant to the spatial organization of tissues.

By leveraging a heterogeneous graph representation, dual learning for spatial domain identification, and RWR-based spatial gene detection, HarveST provides a unified framework for analyzing spatial transcriptomics data. Benchmarking across multiple datasets demonstrates that HarveST outperforms existing methods in both spatial clustering and gene detection, offering improved biological interpretability. These results highlight the potential of HarveST as a powerful tool for studying spatially resolved gene expression in diverse biological contexts.

HarveST accurately delineates complex cortical structures with enhanced boundary precision

To evaluate the spatial clustering performance of HarveST, we applied it to the human dorsolateral prefrontal cortex (DLPFC) dataset[29]. This dataset comprises twelve consecutive DLPFC tissue sections from three human brain samples, which span the six-layered neuronal structure of the cortex, extending from the outermost cortical layers to the deep white matter regions. Maynard et al.[29] provided detailed manual annotations for these tissue sections based on cellular morphology and gene marker analysis, resulting in a comprehensive spatial gene expression map of the cortical and white matter layers. These annotations serve as a reference for evaluating the clustering capabilities of various spatial transcriptomics methods.

We compared HarveST's performance with 14 alternative methods: STMGCN[30], SpaGCN[27], Stagate[11], BayesSpace[9], SEDR[10], MNMST[15], stDyer[14], and GraphST[12], as well as recently developed approaches including CellCharter[19], Impeller[18], and stKeep[17]. Non-spatial clustering methods, specifically Scanpy[26] and Seurat[8], were also included for reference. All methods were implemented using default parameters. To provide a comprehensive assessment, we expanded our evaluation to include 9 distinct metrics: ARI, NMI, Boundary F1-score, Boundary

ARTICLE IN PRESS

Precision, Boundary Recall, Macro and Weighted Dice coefficients, Calinski-Harabasz index, and Silhouette score.

Across the 12 analyzed DLPFC sections, HarveST showed consistent superiority in recovering annotated domains. As depicted in Fig.2A, Supplementary Table 1 and Supplementary Figure 1, HarveST achieved the highest median values across global agreement metrics (ARI, NMI, weighted Dice), outperforming alternative methods. Statistical analysis confirmed that HarveST yielded significantly higher ARI scores compared to 13 out of 14 baseline methods (two-sided Wilcoxon signed-rank test, $P < 0.05$), including leading competitors such as CellCharter ($P = 0.00805$) and GraphST ($P = 0.017$). Regarding boundary delineation, HarveST reached the highest Boundary Precision, confirming its ability to define accurate and stable domain borders while avoiding over-segmentation. Although Scanpy achieved slightly higher Boundary Recall and Boundary F1, its boundaries were discontinuous and less anatomically coherent. For internal cluster separation, CellCharter (joint mode) achieved the highest Calinski-Harabasz and Silhouette scores, reflecting its strong emphasis on local niche separation. HarveST delivered comparable compactness while achieving superior alignment with anatomical structures, thereby offering a balanced trade-off between global agreement, boundary fidelity, and spatial smoothness.

A detailed analysis of section 151674 revealed that Seurat and Scanpy failed to accurately delineate the white matter layer, often merging cortical layers into non-contiguous clusters (Fig.2C,D and Supplementary Figures 2-4). In contrast, spatially aware methods such as stDyer, stKeep, Impeller, GraphST, and HarveST showed improved correspondence with manual annotations, particularly in discriminating white matter from adjacent cortical layers. Notably, HarveST and Stagate displayed the best performance in resolving the elongated Layer 2 and compact Layer 4, two anatomically challenging regions owing to their thin and continuous morphology. However, HarveST produced smoother and more contiguous cluster boundaries than the jagged edges observed in Stagate. Quantitatively, HarveST achieved the highest ARI for this section (0.68), exceeding GraphST (0.645) and stKeep (0.634), further supporting its superior precision in spatial domain delineation across DLPFC slices.

HarveST reveals spatial heterogeneity in the breast cancer microenvironment with improved domain resolution

To explore the spatial heterogeneity of the cancer microenvironment, we applied the HarveST to a human breast cancer dataset generated by the 10X Visium spatial transcriptomics platform. This dataset consists of 3798 spots and 36,601 genes. Histopathological annotations were provided by pathologists based on hematoxylin and eosin (H&E) staining, identifying four primary morphological types: ductal carcinoma in situ/lobular carcinoma in situ (DCIS/LCIS), healthy tissue, invasive ductal carcinoma (IDC), and tumor edge regions characterized by lower malignancy surrounding the tumor[10] (Fig.2E- Fig.2F). Given the complex and heterogeneous nature of tumor tissues, which often lack clear morphological boundaries, accurate identification of spatial domains in the breast cancer microenvironment poses a major challenge. We compared HarveST with several spatial clustering methods, including SEDR, MNMST, GraphST, Stagate, CellCharter, stKeep and Impeller using 9 quantative metrics to measure the concordance between the clustering results and the manual annotations (Supplementary Table 2). As shown in Fig.2, while all methods were effective in partitioning the dataset into continuous, non-overlapping clusters, the tumor edge regions posed a unique challenge due to their subtle morphological features, often leading to misclassifications between healthy and tumor regions. For example, MNMST incorrectly classified the healthy regions, Healthy 1 and Healthy 2, as part of the adjacent tumor edge.

ARTICLE IN PRESS

Other methods such as SEDR, GraphST, and Stagate, demonstrated varying degrees of proficiency in distinguishing tumor edge regions, including Tumor edge 2, Tumor edge 3, and Tumor edge 5. However, these methods tend to merge narrower tumor edge regions with adjacent tumor areas, illustrating their limitations in handling subtle heterogeneity. For instance, GraphST and Stagate misclassified Tumor edge 5 as part of the IDC 4, while SEDR merged portions of Tumor edge 3 with the neighboring IDC 1 region. CellCharter successfully identified the healthy regions but over-segmented the IDC areas, dividing them into several small fragmented clusters lacking biological coherence. HarveST, SEDR and Stagate demonstrated superior accuracy in identifying tumor edge regions, successfully identifying their boundaries. This level of precision is critical for accurate disease diagnosis and subsequent therapeutic interventions. Additionally, methods such as SEDR, Stagate, GraphST and stKeep fragmented the healthy region Healthy 1 into separate categories, HarveST was capable of identifying the entire healthy region without splitting it into multiple clusters. These results were further supported by the quantitative scores, with HarveST achieving the highest ARI score of 0.602, surpassing the maximum scores of 0.575 obtained by alternative methods. This performance underscores HarveST's enhanced accuracy and reliability in analyzing spatial transcriptomics data, particularly within the complex landscape of cancerous tissues.

To further validate the fidelity of the tumor domains identified by HarveST, we employed inferCNV to estimate copy number variations (CNVs) from the spatial transcriptomics data. The HarveST-identified normal regions were used as the diploid reference baseline to infer the genomic instability of the predicted tumor regions. As shown in Supplementary Figure 5A, the inferred CNV profiles revealed distinct chromosomal alterations in HarveST-defined tumor domains, in sharp contrast to the genomic stability of the normal domains.

Quantitatively, CNV scores in tumor regions were markedly elevated relative to normal regions (Supplementary Figure 5B), confirming that the HarveST-segmented domains align with the underlying genomic pathology of malignancy. When benchmarked against pathologist-annotated ground truth, HarveST achieved high precision (0.967) and a strong Pearson correlation coefficient (0.855), indicating accurate identification of malignant regions with minimal false positives. Notably, the model exhibited moderate recall (0.619), reflecting a conservative prediction strategy that prioritizes purity over coverage. This property is advantageous for downstream analyses requiring high-confidence tumor cell populations, ensuring that subsequent differential expression or survival assessments are not confounded by non-malignant tissue inclusion.

ARTICLE IN PRESS

HarveST reconstructs fine-grained laminar structures across diverse spatial resolutions

To validate the robustness of HarveST in identifying tissue structures across datasets with varying spatial resolutions, we applied it to mouse olfactory bulb (MOB) samples obtained from two distinct spatial transcriptomics platforms: Stereo-seq[31] and Slide-seqV2[21].

Stereo-seq Data The Stereo-seq dataset provides a high spatial resolution of 500 nanometers per spot, capturing 19,527 spots and 14,397 genes. According to Fu et al.[10], the MOB coronal sections are annotated into seven distinct layers: the Rostral Migratory Stream (RMS), Granule Cell Layer (GCL), Inner Plexiform Layer (IPL), Glomerular Layer (GL), Mitral Cell Layer (MCL), External Plexiform Layer (EPL), and Olfactory Nerve Layer (ONL).

We compared HarveST's performance with several methods, including Scanpy, SEDR, Stagate, MNMST, GraphST, CellCharter and Impeller. Non-spatial methods, such as Scanpy, performed poorly, exhibiting considerable inter-cluster mixing due to its inability to incorporate

ARTICLE IN PRESS

spatial information. In contrast, spatially aware methods, particularly HarveST, accurately captured the MOB's layered architecture, demonstrating strong concordance with manual annotations. Notably, HarveST and GraphST were the only methods capable of distinctly identifying the narrow MCL structure, as evidenced by the expression of the mitral cell marker *Gabra1*[32]. Additionally, HarveST outperformed other methods in resolving the elongated RMS structure, which is characterized by *Mbp* expression[11], a structure that posed substantial challenges for most approaches (Fig.3A- 3D, Supplementary Figure 6).

Furthermore, while both HarveST and GraphST effectively identified the elongated RMS, only HarveST demonstrated distinct delineation of adjacent layers, such as the RMS, IPL, and GCL. Other methods struggled with these closely apposed layers: Stagate erroneously merged the RMS with the IPL, while GraphST and MNMST failed to clearly separate the IPL from the GCL. Impeller produced fragmented and discontinuous regions, reflecting its limited ability to preserve spatial continuity, whereas CellCharter captured only the outer morphological contour without resolving the internal laminar architecture. In contrast, HarveST was uniquely capable of accurately distinguishing these structures, a result further supported by the distinct expression patterns of key marker genes, including *Mbp*, *Nrgn*, and *Pcp4*[33], demonstrating its superior spatial resolution and biological relevance.

Slide-seqV2 Data The Slide-seqV2 dataset provides near-cellular resolution, consisting of 20,139 spots and 21,220 genes[21]. According to structural annotations from the Allen Reference Atlas[34], the regions include ONL, GL, EPL, MCL, IPL, GCL, RMS, the accessory olfactory bulb (AOB), and the granular layer of the accessory olfactory bulb (AOBgr).

Consistent with the results obtained from the Stereo-seq data, methods such as SEDR, Scanpy, Stagate and CellCharter were only able to capture the general tissue outline within the MOB. In contrast, GraphST, Impeller, MNMST, and HarveST provided stratifications that closely aligned with annotated structures. These methods successfully identified spatial domains corresponding to the semicircular structure of the AOB and the crescent-shaped AOBgr. However, HarveST was the only method capable of precisely delineating the arcuate boundary between the AOB and AOBgr, as well as accurately defining the RMS within the inner medullary layer, which corresponds to the known architecture of the mouse olfactory bulb (Fig.3E - Fig.3H). The spatial domains identified by HarveST exhibited strong concordance with annotated structures and were further supported by known gene markers. For example, the *Doc2g* gene exhibited high expression in the EPL, while the *Gap43* showed strong expression in the AOB, both consistent with immunohistochemistry findings[33]. Additionally, the granule cell marker *Atp2b4*[35] displayed strong expression in the AOBgr area, and the mitral cell marker *Gabra1*[32] aligned with the MCL structure, all of which were accurately identified by HarveST (Supplementary Figure 7).

This accurate identification of spatial domains, supported by biologically relevant gene expression patterns, further underscores HarveST's robustness in resolving complex tissue architectures at near-cellular resolution.

Joint Analysis of Multiple Spatial Transcriptomics Slices

To assess whether HarveST can consistently integrate spatial information across tissue sections, we performed a joint analysis on two consecutive DLPFC slices 151673 and 151674 derived from the same donor. Each section was processed within a unified heterogeneous graph framework that preserves intra-slice topology while enabling cross-slice connections based on feature similarity and spatial continuity. This design allows HarveST to model both local spatial context and inter-slice correspondence within the same latent space.

ARTICLE IN PRESS

The resulting joint embedding revealed well-aligned molecular and morphological patterns across slices, indicating accurate structural correspondence between adjacent cortical layers. Quantitatively, batch effect correction markedly improved neighborhood mixing (LISI 1.9961 compared with 1.7502 before correction), demonstrating that HarveST effectively harmonizes multi-slice data without distorting underlying spatial organization. The reconstructed domains displayed continuous laminar organization, with cortical layers seamlessly extending across adjacent sections (Supplementary Figure 8). These results highlight HarveST's ability to identify shared functional spatial domains across tissue slices, achieving global architectural consistency while preserving local anatomical detail.

HarveST enhances the identification of spatial domain-marker SVGs and functional modules with biologically meaningful insights

To evaluate HarveST's performance in identifying both spatial domain-marker SVGs and spatially-associated gene modules, we benchmarked it against existing methods, including Scanpy, SpaGCN, and DESpace, across multiple datasets. The detailed evaluation framework is provided in Supplementary Note 1. Comprehensive statistical metrics for all identified marker genes,including exact p-values, adjusted p-values, and effect sizes (e.g., Log2 Fold Change) are listed in Supplementary Tables 3–31.

In the DLPFC dataset, HarveST identified *MOBP*, *MOG*, and *MAG* as the top domain-specific SVGs for white matter (WM) (Fig.4B). While Wilcoxon-based Scanpy and SpaGCN identified *MBP*, *PLP1*, and *GFAP* (Supplementary Figure 9 D-E), visualization revealed that HarveST's markers exhibited highly specific expression patterns restricted to the WM region. In contrast, *MBP* displayed relatively high expression across both WM and other regions. Although DESpace identified *BCAS1*[36] and *CLDN11*[37] as oligodendrocyte markers, *MOBP* stands out as the only marker strictly specific to oligodendrocytes in the human brain.

In Layer 5 of the cortex, HarveST identified *PCP4* as the gene most strongly associated with this region (Fig.4E), whereas Scanpy and DESpace identified *TMSB10*. Visualization demonstrated that *PCP4* exhibited high spatial specificity within Layer 5, with minimal background expression. Notably, SpaGCN also identified *PCP4* as the only region-specific gene, corroborating its role as a Layer 5 marker.

In the HBRC dataset analysis (Supplementary Figure 9; Tables 3-17), all four methods consistently identified *CRISP3* as highly specific to the IDC 2 region. While HarveST, Scanpy, and SpaGCN all identified *SLITRK6*, only HarveST and DESpace prioritized *C6orf141* (Fig.4H-I), a gene documented to be upregulated in tumor cells[38]. However, *EFHD1*, identified by DESpace, is generally recognized as a marker of ionocytes in lung tissue[39], highlighting the robustness of HarveST. Conversely, *H3F3A*, identified by Scanpy and SpaGCN, is a marker for NKT cells[40] but showed less pronounced spatial specificity compared to *C6orf141*. Among the top 20 genes, HarveST uniquely identified *ABCC11* (Fig.4M), which is associated with aggressive subtypes and poor survival[41].

In the IDC 3 region, HarveST identified *MS4A1*, *RASGRP2*, and *CD79A* as the top three spatially specific genes (Fig.4K). *CD79A* is a recognized marker for various cancer cells associated with invasiveness[42], while *MS4A1* is implicated in tumorigenesis and immune infiltration. Additionally, *RASGRP2* is abnormally expressed in cancer cells and influences immune cell infiltration[43]. Although all methods identified partially overlapping top genes, within the top 20, HarveST uniquely identified *IL7R*, *CD3E*, and *CD3D* (Fig.4N). Multiple studies confirm *IL7R* expression as a cancer cell marker in breast tissue[44,45], and *CD3D*/*CD3E* have been validated as breast cancer markers[46], demonstrating HarveST's enhanced sensitivity.

Beyond marker identification, we further validated the biological relevance by conducting

ARTICLE IN PRESS

functional enrichment analysis on the gene modules (Supplementary Figures 10-14; workflow in Supplementary Note 2). The analysis revealed meaningful enrichment in pathways related to invasive breast ductal carcinoma and recurrent tumors, suggesting that the modules detected by HarveST play crucial roles in invasion and recurrence.

To quantitatively evaluate spatial expression specificity, we calculated Moran's I index and Geary's C coefficient for the top 20 genes (Supplementary Figure 15)[47–49]. The median ratio of Moran's I index for HarveST consistently surpasses that of Wilcoxon, with a right-skewed distribution indicating more pronounced spatial clustering specificity. The adjusted Geary's C results further corroborate these findings, demonstrating lower dispersion and reinforcing HarveST's potential in identifying spatially specific genes.

In the Pancreatic ductal adenocarcinoma (PDAC) dataset (Supplementary Tables 18-31)[50], alternative methods identified *CTRB1*, *CTRB2*, and *REG3A* as SVGs (Fig.5C, E, F). Notably, HarveST also recognized *AC009078.2* alongside *CTRC*[51] and *PNLIPRP2*[52], both established markers for normal pancreatic cells. Unlike genes identified by alternatives, these exhibited distinct expression patterns specifically confined to the pancreatic region. Additionally, *PPY*, uniquely identified by HarveST among the top 20, displayed a distinct expression pattern (Fig.5D) and is a validated pancreatic cell marker[52]. Furthermore, HarveST uniquely identified *SPP1*, *MUC6*, and *CPA1* (Fig.5G), all experimentally validated markers for normal human pancreatic tissue[53,54].

In the cancer regions, HarveST, DESpace, and Scanpy consistently identified *KRT17* (Fig.5J-M), a well-documented marker for pancreatic cancer proliferation and invasion[46,55]. HarveST also highlighted *LAMC2*, which mediates metastasis[56], and *SFN*, which is essential for cell migration[57]. Notably, HarveST uniquely identified *IFI27* (Fig.5L), a gene highly upregulated in PDAC and linked to poor prognosis[58,59]. Further analysis of the top 20 genes revealed additional cancer-associated genes: *FXYD3*, *CHPF*, *C19ORF33*, *TAGLN*, and *AEBP1* (Fig.5O). For instance, *FXYD3* is markedly upregulated in PDAC[60], and *C19ORF33* correlates with patient prognosis[61]. *TAGLN* and *AEBP1* are recognized markers promoting proliferation and invasion[62,63].

Enrichment analysis (Supplementary Figures 16-18) revealed considerable associations with pathways implicated in malignant tumor progression. Fig.5P depicts the KEGG enrichment for the pancreatic region, where the pancreatic secretion pathway exhibited the highest enrichment. Furthermore, Fig.5Q highlights that the top 100 cancer region-specific genes are strongly involved in cell-matrix interactions, signal transduction, and oncogenic pathways. Collectively, these results underscore the efficacy and precision of HarveST in detecting functionally relevant region-specific genes within complex tissue architectures.

Technical validation and computational efficiency

To rigorously evaluate the reliability of the HarveST framework, we performed systematic ablation studies. These analyses confirmed that both the gene-gene interaction graph and the partially supervised refinement strategy are essential for accurate spatial domain identification (Supplementary Note 3, Supplementary table 32). We further assessed the model's stability, demonstrating its robustness against initialization bias and label perturbations (Supplementary Note 4), as well as its insensitivity to variations in the pseudo-label inclusion ratio (Supplementary Note 5). Finally, runtime benchmarking indicates that HarveST achieves competitive computational efficiency compared to existing deep learning-based baselines, supporting its scalability for large-scale datasets (Supplementary Note 6, Supplementary table 32).

ARTICLE IN PRESS

DISCUSSION

Spatial transcriptomics has revolutionized our ability to dissect tissue architecture at unprecedented resolution, but challenges persist in accurately delineating spatial domains and identifying biologically relevant, spatially variable genes. HarveST addresses these limitations by integrating a heterogeneous graph framework that integrates spatial proximity, gene expression similarity, and functional gene-gene interactions into a unified learning paradigm. A dual learning strategy, combining self-supervised graph embedding with pseudo-label-guided refinement, enhances the robustness of spatial domain identification, especially in regions with subtle transcriptional gradients. Furthermore, HarveST overcomes the limitations of gene-centric statistical tests by employing RWR on a heterogeneous graph. This approach propagates spatial signals in a biologically meaningful manner that reflects cell-cell communication, effectively leveraging biological context to identify spatially relevant features. This integrated approach refines spatial domain resolution and facilitates the discovery of biologically coherent gene modules, establishing HarveST as a powerful tool for spatial transcriptomic analysis.

Empirical validation across diverse datasets demonstrates the advantages of HarveST over existing methods. In the human dorsolateral prefrontal cortex (DLPFC), HarveST successfully reconstructs the six cortical layers with enhanced boundary precision, particularly in distinguishing the white matter from Layer 6, a region prone to misclassification due to transcriptomic similarity. HarveST achieves consistently higher quantitative scores than GraphST and stKeep, underscoring its ability to resolve complex tissue architectures with high fidelity. In the mouse olfactory bulb (MOB), HarveST reliably identifies laminar structures across platforms with different spatial resolutions, including Stereo-seq and Slide-seqV2, demonstrating robustness to resolution variability. Importantly, HarveST accurately delineates the mitral cell layer from the external plexiform layer, two transcriptionally similar regions that remain challenging for other graph-based frameworks. This fine-scale resolution capability indicates that HarveST captures subtle spatial transitions essential for dissecting cellular microenvironments. The ability to jointly integrate consecutive sections further enhances spatial domain inference, ensuring structural continuity and reproducibility across tissue slices.

Beyond spatial domain identification, HarveST excels in detecting functionally relevant, spatially specific genes. In the breast cancer microenvironment, genes identified by HarveST, including *IFI27*, are strongly enriched in tumor-immune interaction pathways, suggesting a role in tumor progression and immune modulation. Notably, many of these genes exhibit enrichment in pathways associated with epithelial-mesenchymal transition (EMT) and tumor microenvironment remodeling, validating that HarveST effectively captures biologically meaningful patterns that may be overlooked by conventional statistical approaches. Similarly, in pancreatic ductal adenocarcinoma (PDAC), HarveST identifies key markers differentiating malignant regions from normal pancreatic tissue, providing valuable insights into tumor heterogeneity. The ability to integrate spatial information with gene-gene functional interactions enables HarveST to uncover gene modules that are not only spatially distinct but also functionally coherent, facilitating a deeper understanding of tissue organization and disease progression.

Beyond identifying spatially specific genes, HarveST's construction of a mutual information-based gene-gene interaction graph provides a data-driven alternative to database-dependent cell-cell communication analysis. While traditional methods rely on curated ligand-receptor (L-R) databases, such resources often cover only a small fraction of the genes contributing to spatial variation (Supplementary Table 34). By leveraging mutual information, HarveST captures a broader range of dependencies, including co-expression modules, downstream signaling relationships, and indirect regulatory effects—while still recovering key L-R interactions. For example, HarveST autonomously assigned high interaction weights to

ARTICLE IN PRESS

NRXN1-NLGN3 in the DLPFC dataset and to the *SDC1-COL1A1* axis in the PDAC dataset, both consistent with known tissue-specific communication mechanisms. Moreover, our ablation studies demonstrate that HarveST's statistical graph yields higher clustering accuracy than architectures constrained by sparse L-R priors (Supplementary Table 35). Together, these findings indicate that HarveST provides a robust framework for dissecting tissue architecture without being limited by the incompleteness of current ligand-receptor interaction databases.

Although the inclusion of Mutual Information introduces a preprocessing overhead compared with nearest-neighbor approaches, our ablation studies confirm that this step is essential for resolving complex non-linear dependencies. From a workflow perspective, this initialization represents a one-time investment. In biological analyses that require frequent adjustments of parameters such as cluster number or resolution, HarveST allows reuse of the preconstructed graph. This separation between structural learning and inference minimizes repeated computation and makes the marginal cost of subsequent exploratory tasks negligible.

Furthermore, the selection of seed spots in the RWR algorithm, which influences the identification of spatially specific genes, remains a key methodological consideration. Future work could automate this process using spatial autocorrelation metrics or unsupervised domain adaptation techniques, broadening applicability across different tissue types and experimental conditions.

By integrating spatial domain identification and functional gene discovery within a unified heterogeneous graph framework, HarveST bridges a critical gap in spatial transcriptomics analysis. Unlike conventional approaches that treat these tasks independently, HarveST models them as interdependent components of a unified spatial landscape, capturing the reciprocal relationship between tissue architecture and gene regulation. As spatial omics technologies evolve toward single-cell resolution and multi-omic integration, the ability to jointly model spatial, molecular, and functional relationships will become increasingly important. HarveST provides a scalable and biologically interpretable framework that can be extended to incorporate multi-modal data, such as spatial proteomics and epigenomics, further enhancing our ability to decode spatially resolved gene regulation in health and disease.

ARTICLE IN PRESS

ARTICLE IN PRESS

METHODS

Data Preprocessing

The spatial gene expression data is represented as an $N \times M$ matrix $X = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_N}) \in \mathbb{R}^{N \times M}$, where N is the number of spots and M is the number of genes. Each spot s_u is associated with two-dimensional spatial coordinates (x, y), indicating its position within the tissue. To stabilize variance and minimize the influence of outliers, we first apply a logarithmic transformation to the raw gene expression data. Subsequently, we normalize the expression values for each spot and each gene independently. For each spot s_u, the normalized feature vector $\boldsymbol{x}_{s_u}$ is calculated as:

$$\boldsymbol{x}_{s_u} = \frac{\hat{\boldsymbol{x}}_{s_u} - \mu_{s_u}}{\sigma_{s_u}},$$

where $\hat{\boldsymbol{x}}_{s_u}$ represents the vector of raw expression values for all genes in spot s_u, μ_{s_u} is the mean expression value across all genes in spot s_u, and σ_{s_u} is the corresponding standard deviation.

Similarly, for each gene g_i, its normalized feature vector $\boldsymbol{x}_{g_i}$ across all spots is computed as:

$$\boldsymbol{x}_{g_i} = \frac{\hat{\boldsymbol{x}}_{g_i} - \mu_{g_i}}{\sigma_{g_i}},$$

where $\hat{\boldsymbol{x}}_{g_i}$ represents the vector of expression values for gene g_i across all spots, μ_{g_i} and σ_{g_i} are the mean and standard deviation of expression values of gene g_i across all spots, respectively.

Heterogeneous Graph Construction

To ensure terminological clarity, we explicitly distinguish between the physical spatial capture units and their graph-based representations. The term "spot" refers exclusively to the physical capture unit of a spatial transcriptomics platform (e.g., 10x Genomics Visium), characterized by its gene expression profile and spatial coordinates. In contrast, "node" denotes an abstract vertex in the heterogeneous graph, with spots mapped to "spot nodes" (V_s) and genes to "gene nodes" (V_g). This distinction separates biological observations from the topological framework used for representation learning.

Building upon these definitions, HarveST constructs a heterogeneous graph $\mathcal{G} = (V, E)$ to integrate spatial proximity, transcriptomic profiles, and gene regulatory priors. The vertex set is defined as $V = V_s \cup V_g$. The edge set $E = E_{ss} \cup E_{sg} \cup E_{gg}$ encapsulates three distinct interaction modalities: E_{ss} denotes spatial adjacency between spots, E_{sg} represents the expression abundance of genes within spots, and E_{gg} encodes gene-gene interaction networks. Each edge is assigned a weight reflecting the interaction strength, which enables the graph to capture multi-scale biological relationships within a unified topological framework.

Spot-Spot Subgraph The spot-spot subgraph $G_{ss} = (V_s, E_{ss})$ captures spatial proximity between spots. Edges in this subgraph are defined based on the Euclidean distance between spots. Specifically, the adjacency matrix A_{ss} is constructed as follows:

$$A_{ss}(u, v) = \begin{cases} 1 & \text{if } d(s_u, s_v) \leq r, \\ 0 & \text{otherwise}, \end{cases}$$

where $d(s_u, s_v)$ is the Euclidean distance between spots s_u and s_v, and r is a threshold optimized based on the data density and the spatial resolution of the sequencing platform.

ARTICLE IN PRESS

Spot-Gene subgraph The spot-gene subgraph $G_{sg} = (V_s \cup V_g, E_{sg})$ models the interaction between spots and genes. An edge exists between spot s_u and gene g_i if gene g_i is expressed at spot s_u. To quantify the strength of these interactions, we apply a k-means clustering to the normalized gene expression values and discretize them into quantiles (Q). The weight $A_{sg}(u,i)$ of the edge between spot s_u and gene g_i is assigned based on the quantile of gene g_i's expression at spot s_u as follows:

$$A_{sg}(u,i) = \begin{cases} 1.0 & \text{if } x_{u,i} \in Q_1, \\ 2.0 & \text{if } x_{u,i} \in Q_2, \\ 3.0 & \text{if } x_{u,i} \in Q_3, \\ 4.0 & \text{if } x_{u,i} \in Q_4, \\ 5.0 & \text{if } x_{u,i} \in Q_5. \end{cases}$$

The resulting matrix A_{sg} is then normalized by dividing each entry by the maximum weight across all spot-gene pairs.

Gene-Gene subgraph The gene-gene subgraph $G_{gg} = (V_g, E_{gg})$ captures the co-expression relationships between genes. The edge weights in this subgraph are determined using mutual information (MI) between the expression profiles of gene pairs. For two genes g_i and g_j, the mutual information $I(x_{g_i}; x_{g_j})$ is computed as:

$$I(x_{g_i}; x_{g_j}) = \int_{x_{g_i}} \int_{x_{g_j}} p(x_{g_i}, x_{g_j}) \log \frac{p(x_{g_i}, x_{g_j})}{p(x_{g_i})p(x_{g_j})} dx_{g_i} dx_{g_j},$$

where $p(x_{g_i}, x_{g_j})$ is the joint probability distribution of the expression levels of genes g_i and g_j, and $p(x_{g_i})$ and $p(x_{g_j})$ are their marginal distributions.

To construct a biologically interpretable gene interaction network, we calculated the Mutual Information (MI) for all gene pairs. MI was prioritized over linear correlation metrics to capture non-linear regulatory dependencies and was preferred to partial correlation, which can be computationally unstable under high-dimensional, sparse conditions typical of spatial transcriptomics data. To reduce noise and retain high-confidence interactions, a thresholding strategy was applied that preserves the top 20% of gene pairs with the highest MI scores. The adjacency matrix A_{gg} was then normalized to encode these weighted interactions.

Heterogeneous Graph Attention Model Following heterogeneous graph construction, we employ a graph attention mechanism to selectively aggregate information from neighboring nodes. The node features for each spot s_u and gene g_i at layer l are updated as follows:

For spot nodes, the contribution from gene neighbors is calculated as::

$$A_{s_u}^{g(l)} = \sum_{g_i \in V_{s_u}^g \subset G_{sg}} \frac{1}{|V_{s_u}^g|} \alpha_{ui} W_{sg}^{(l)} h_{g_i}^{(l-1)} A_{sg}$$

The contribution from spot neighbors and self-connection is calculated as:

$$A_{s_u}^{s(l)} = \sum_{s_v \in V_{s_u}^s \subset G_{ss}} \frac{1}{|V_{s_u}^s|} \alpha_{uv} W_{ss}^{(l)} h_{s_v}^{(l-1)} A_{ss} + W_{ss}^{(l)} h_{s_u}^{(l-1)} + b_s^{(l)}$$

These contributions are then combined and passed through an activation function:

$$h_{s_u}^{(l)} = \sigma(A_{s_u}^{g(l)} + A_{s_u}^{s(l)})$$

ARTICLE IN PRESS

Similarly, for gene nodes, the contribution from spot neighbors is calculated as:

$$A_{g_i}^{s(l)} = \sum_{s_u \in V_{g_i}^s \subset G_{sg}} \frac{1}{|V_{g_i}^s|} \alpha_{ui} W_{sg}^{(l)} h_{s_u}^{(l-1)} A_{sg}$$

The contribution from gene neighbors and self-connection is calculated as:

$$A_{g_i}^{g(l)} = \sum_{g_j \in V_{g_i}^g \subset G_{gg}} \frac{1}{|V_{g_i}^g|} \alpha_{ij} W_{gg}^{(l)} h_{g_j}^{(l-1)} A_{gg} + W_{gg}^{(l)} h_{g_i}^{(l-1)} + b_g^{(l)}$$

These contributions are then combined and passed through an activation function:

$$h_{g_i}^{(l)} = \sigma(A_{g_i}^{s(l)} + A_{g_i}^{g(l)})$$

where $V_{s_u}^g$ and $V_{s_u}^s$ represent the sets of gene and spot nodes connected to spot s_u, respectively. Similarly, $V_{g_i}^s$ and $V_{g_i}^g$ represent the sets of spot and gene nodes connected to gene g_i. The attention scores α_{ij} are computed as:

$$\alpha_{ij} = \mathsf{softmax}(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in N_i} \exp(e_{ik})}, \quad e_{ij} = \mathsf{LeakyReLU}\left(a^T \left[W_{\mathsf{type}}^{(L)} h_i^{(L)} \parallel W_{\mathsf{type}}^{(L)} h_j^{(L)}\right]\right) \cdot A_{ij},$$

where e_{ij} represents the attention logits based on the concatenated features of nodes i and j. This graph attention mechanism allows the model to focus on biologically relevant relationships, enhancing feature representation for downstream tasks.

Joint analysis For datasets comprising multiple spatial transcriptomics slices, HarveST performs joint analysis within a unified heterogeneous graph framework. All slices are first concatenated under a common expression matrix, followed by global selection of the 3,000 most variable genes. The combined matrix is normalized by total counts and log-transformed to harmonize sequencing depth and expression scale across slices.

Each slice independently constructs its spot–spot and spot–gene subgraphs to preserve local structure, while the gene–gene subgraph is computed once using all pooled spots to capture overall transcriptional co-expression. The resulting slice-wise graphs are concatenated into a global spatial graph, forming block-diagonal intra-slice connectivity. Cross-slice edges are then introduced according to feature similarity and spatial continuity, allowing the joint propagation of biological signals between adjacent tissue sections.

The integrated feature representation, combined adjacency structure, and global mutual information matrix are subsequently fed into the heterogeneous graph neural network for low-dimensional embedding and spatial domain inference. This joint framework effectively aligns molecular and spatial patterns across slices, enabling the identification of consistent tissue-wide functional domains.

Partially Supervised for Spatial Domain Identification

The identification of spatial domains in HarveST involves a two-step process: (1) feature extraction through an autoencoder and (2) refinement of clusters using partially supervised learning.

ARTICLE IN PRESS

Autoencoder for Latent representation HarveST begins by applying an autoencoder on the heterogeneous graph to learn latent representations of spots and genes. The autoencoder comprises an encoder that learns a latent feature representation Z and a decoder that reconstructs the original data. The objective is to minimize reconstruction loss, ensuring that the latent features retain essential biological information while reducing noise.

The overall loss function for the autoencoder is a weighted sum of mean squared errors for gene and spot reconstructions:

$$L = \lambda \sum_{i=1}^{M} \|h_{g_i}^{(0)} - h_{g_i}\|^2 + (1-\lambda) \sum_{u=1}^{N} \|h_{s_u}^{(0)} - h_{s_u}\|^2,$$

where h_{g_i} and h_{s_u} represent the reconstructed features for genes and spots, respectively, and λ balances the contributions of gene and spot reconstruction.

Pre-Clustering with Gaussian Mixture Model The latent spot features Z learned by the autoencoder are then clustered using a Gaussian mixture model (GMM). GMM models the distribution of the latent space as a mixture of K Gaussian distributions, where each Gaussian distribution represents a cluster. The probability density function $p(z_u)$ for spot z_u is defined as:

$$p(z_u) = \sum_{k=1}^{K} \pi_k \mathcal{N}(z_u | \mu_k, \Sigma_k),$$

where π_k is the mixing coefficient, μ_k is the mean, and Σ_k is the covariance matrix of the k-th Gaussian distribution. The cluster assignment for each spot is determined by the posterior probability:

$$\text{Cluster}(z_u) = \underset{k}{\text{argmax}}\, p(k|z_u) = \underset{k}{\text{argmax}} \left(\frac{\pi_k \mathcal{N}(z_u|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(z_u|\mu_j, \Sigma_j)} \right).$$

The top 50% of spots with the highest posterior confidence from the GMM clustering are selected, and their corresponding cluster assignments are designated as 'pseudo-labels.' These labels are model-generated provisional annotations derived solely from the GMM results, not from histological ground truth. They serve as supervisory signals to guide the classifier training during the subsequent refinement stage.

Partially Supervised Training of Classifier The pseudo-labeled high-confidence spots are then used to train a Support vector machine (SVM) classifier[64], which refines the cluster boundaries and improves the separability between different spatial domains. The SVM optimization is formally defined as:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{u} \xi_u,$$

subject to:

$$y_u(\mathbf{w} \cdot \phi(z_u) + b) \geq 1 - \xi_u, \quad \xi_u \geq 0,$$

where $\phi(z_u)$ maps the feature space of spot z_u into a higher-dimensional space to facilitate improved separation between clusters. The parameter $\mathbf{w}$ represents the weight vector, b is the bias term, and ξ_u are slack variables that allow for some degree of misclassification, enabling the model to handle non-linear boundaries.

ARTICLE IN PRESS

To ensure robust generalization, we employ a weighted cross-entropy loss function augmented with L_2 regularization:

$$L = -\sum_{u=1}^{N}\sum_{c=1}^{K} w_c y_{uc} \log(\hat{y}_{uc}) + \lambda\|\theta\|_2^2,$$

where w_c is the weight assigned to class c, $\hat{y}_{uk}$ is the predicted probability for calss c of spot s_u, y_{uc} is the ground truth for spot s_u, and λ controls the magnitude of the L_2 penalty .

By leveraging high-confidence spots with pseudo-labels, the SVM effectively learns from a partially labeled dataset, enhancing the robustness of clustering results. This partially supervised approach combines the strengths of unsupervised learning with the discriminative power of supervised learning, improving classification accuracy and the identification of spatial domains.

Random Walk with Restart for Spatially Specific Gene Detection

To identify spatial domain-marker SVGs, HarveST employs a Random Walk with Restart (RWR) algorithm on the heterogeneous graph. The process begins by initializing the random walk from a set of seed spots within a predefined spatial domain of interest. The transition probability matrix A' for the random walk is constructed as follows:

$$A' = D^{-1}\begin{pmatrix} D_{ss}^{-1}A_{ss} & D_{sg}^{-1}A_{sg} \\ D_{gs}^{-1}A_{gs} & D_{gg}^{-1}A_{gg} \end{pmatrix},$$

where $D_{ss}, D_{sg}, D_{gs}, D_{gg}$ are diagonal matrices representing the row-sum of the corresponding submatrices.

The RWR process is iteratively updated as:

$$p^{(t+1)} = (1-r)A'p^{(t)} + rp^{(0)},$$

where $p^{(0)}$ is the initial probability vector with 1 assigned to seed spots and 0 elsewhere, and r is the restart probability. The RWR converges when the change between successive iterations is below a predefined threshold $\|p^{(t+1)} - p^{(t)}\|_1 \leq \epsilon$, typically $\epsilon = 10^{-6}$.

spatial domain-marker SVGs Detection The RWR on the weighted heterogeneous graph naturally favors paths along higher-weight edges, which represent stronger interactions between nodes. As a result, genes identified through this process are likely to play crucial roles in the biological characteristics of the spatial domain. Once the RWR has converged, genes with the highest steady-state probabilities are considered to have crucial spatial specificity within the given domain:

$$\text{SVGs} = \{g \in V_g | p_g^{(\infty)} > \theta\}$$

where $p_g^{(\infty)}$ represents the steady-state probability of gene g fter RWR convergence, and θ is a threshold determined based on a permutation test. These genes are identified based on their final probability scores, highlighting their potential roles in domain-specific biological processes.

Identifying spatial-associated functional gene modules To identify spatial-associated functional gene modules, we employ a statistically approach leveraging the RWR-derived association scores. First, we establish a null distribution through several permutation iterations (500 in our test), where in each iteration, a set of spots equal in number to those in the region

ARTICLE IN PRESS

of interest is randomly selected from the entire tissue. For each gene, we calculate an empirical p-value by comparing its observed RWR score against this null distribution, with p $<$ 0.01 indicating statistically significant spatial specificity.

For these spatially significant genes, we quantify their region-specific enrichment using the fold change metric:

$$\text{FC}_i = \frac{\bar{x}_{i,R}}{\bar{x}_{i,\bar{R}}}$$

where $\bar{x}_{i,R}$ represents the mean expression of gene i in the region of interest, and $\bar{x}_{i,\bar{R}}$ is its mean expression in all other regions. Genes exhibiting fold change $>$ 1.5 are aggregated into functional modules, which represent coordinated gene sets with both spatial co-localization and functional relatedness. These modules undergo functional enrichment analysis using Gene Ontology (GO) and KEGG pathway databases to characterize their biological significance, enabling the linkage between spatial organization and molecular mechanisms underlying tissue function or pathology.

Robustness Analysis via Label Perturbation

We evaluated the robustness of the refinement stage by injecting synthetic noise into the initial GMM-derived pseudo-labels prior to classifier training. We employed two perturbation strategies on the DLPFC dataset. First, we applied random noise by randomly reassigning a fixed percentage (5%-40%) of spot labels to incorrect classes to simulate global initialization errors. Second, we applied boundary Noise by selectively flipping labels of spots located at the interfaces of spatial domains to simulate local ambiguity. We trained the refinement classifier using these corrupted labels and quantified the recovery of spatial structure using ARI, NMI, and Boundary F1 scores. We also assessed biological stability by calculating the overlap coefficient of identified marker genes between the models trained on perturbed labels and the reference annotations.

Sensitivity Analysis

We assessed the sensitivity of the refinement stage by sweeping the pseudo-label inclusion ratio k, selecting the top $k\%$ of spots based on their GMM posterior probabilities. This analysis, spanning a range from 30% to 100%, confirmed that the segmentation accuracy remains robust to variations in k within the 30-70 interval, thereby justifying the default selection of 50%.

Statistics and Reproducibility

Statistical analyses were performed using Python (version 3.8) and relevant libraries including SciPy and NumPy. To evaluate clustering performance, the two-sided Wilcoxon signed-rank test was employed to compare the Adjusted Rand Index (ARI) of HarveST against baseline methods across the 12 DLPFC sections, with a significance threshold of $P < 0.05$. For the identification of spatially variable genes (SVGs), empirical p-values were derived using a permutation test (500 permutations), and genes with $P < 0.01$ were considered statistically significant. No statistical methods were used to predetermine sample sizes; however, our sample sizes are consistent with those reported in previous spatial transcriptomics studies.

Reproducibility was ensured by fixing random seeds for all non-deterministic components of the algorithm, including neural network initialization and Gaussian Mixture Models (GMM). The experiments were conducted on biologically independent samples across multiple datasets,

ARTICLE IN PRESS

encompassing twelve consecutive tissue sections from three adult donors for the Human Dorsolateral Prefrontal Cortex (DLPFC), a single tissue section comprising 3,798 spots for Human Breast Cancer (HBRC), and one tissue section for Human Pancreatic Ductal Adenocarcinoma (PDAC). Additionally, two mouse olfactory bulb (MOB) tissue sections obtained from the Stereo-seq and Slide-seqV2 platforms, respectively, were analyzed. All data analysis code is open-source and available on GitHub to facilitate reproducibility; unless otherwise stated, error bars and box plots represent the distribution of data points as defined in the respective figure legends.

Animal Ethics Statement

We have complied with all relevant ethical regulations for animal use. This study involved the secondary analysis of two publicly available spatial transcriptomics datasets of the mouse olfactory bulb. The ethical approval information and animal metadata were reiterated from the original publications as follows:

Stereo-seq Dataset[31]: All procedures were compliant with ethical regulations and approved by the Animal Care and Use Committee of the Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences (License number: IACUC2021002). The mouse olfactory bulb was dissected from 12-week-old C57BL/6J female mice.

Slide-seqV2 Dataset[21]: All procedures were handled according to protocols approved by the Institutional Animal Care and Use Committee (IACUC) of Harvard University (Protocol number: 11-03) and followed the US National Institute of Health Guide for the Care and Use of Laboratory Animals. The study used C57BL/6J mice, both male and female, aged $\geq$ 60 days.

Data Availability

The spatial transcriptomics datasets analyzed in this study are publicly available from the following repositories. The Human Dorsolateral Prefrontal Cortex (DLPFC) dataset[29,65] is available at `http://spatial.libd.org/spatialLIBD/`. The 10x Visium Human Breast Cancer dataset is available from the 10x Genomics website at `https://www.10xgenomics.com/resources/datasets`. The Mouse Olfactory Bulb (Stereo-seq) dataset[31,66] is available from CNGBdb (MOSTA) at `https://db.cngb.org/data_resources/project/CNP0001543`. The Mouse Olfactory Bulb (Slide-seqV2) dataset[21,67] is available via the Broad Institute Single Cell Portal at `https://singlecell.broadinstitute.org/single_cell/study/SCP815`. The PDAC dataset is available from the Gene Expression Omnibus (GEO) under accession number GSE111672[50,68] at `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111672`.

Processed data supporting the findings of this study are available in the GitHub repository (`https://github.com/Seven595/HarveST`) and are archived on Zenodo (`https://doi.org/10.5281/zenodo.18532348`). Source data underlying the plots in the main figures are provided with this paper. Any remaining data are available from the corresponding author upon reasonable request.

Code Availability

The source code for HarveST, including the implementation of the heterogeneous graph learning framework and scripts for reproducing the analysis, is openly available on GitHub at `https://github.com/Seven595/HarveST`. A persistent version of the software code and processed data used to generate the results in this manuscript is archived in the Zenodo repository (DOI: https://doi.org/10.5281/zenodo.18532348)[69].

ARTICLE IN PRESS

ACKNOWLEDGMENTS

This work was partially supported by Shenzhen Science and Technology Program (Grant Nos. ZDSYS20230626091302006 and JCYJ20240813113536047). Y.Z. is partially supported by a Guangdong Provincial Project (2024QN11N085).

During the preparation of this work, the author(s) utilized OpenAI's ChatGPT-4 to enhance the readability and clarity of the text, given that the primary authors are non-native English speakers. The author(s) subsequently reviewed and edited the AI-generated content as necessary and take full responsibility for the final content of the publication.

AUTHOR CONTRIBUTIONS

Conceptualization: J.F., T.Y., and Y.Z.; Methodology: J.F. and Y.Z.; Software: J.F.; Validation: J.F. and T.Y.; Formal analysis: J.F.; Investigation: J.F. and Y.Z.; Resources: T.Y. and Y.Z.; Data curation: J.F.; Writing – original draft: J.F.; Writing – review & editing: J.F., T.Y., and Y.Z.; Visualization: J.F.; Supervision: T.Y. and Y.Z.; Project administration: Y.Z.; Funding acquisition: T.Y. and Y.Z.

Competing Interests

The authors declare no competing interests.

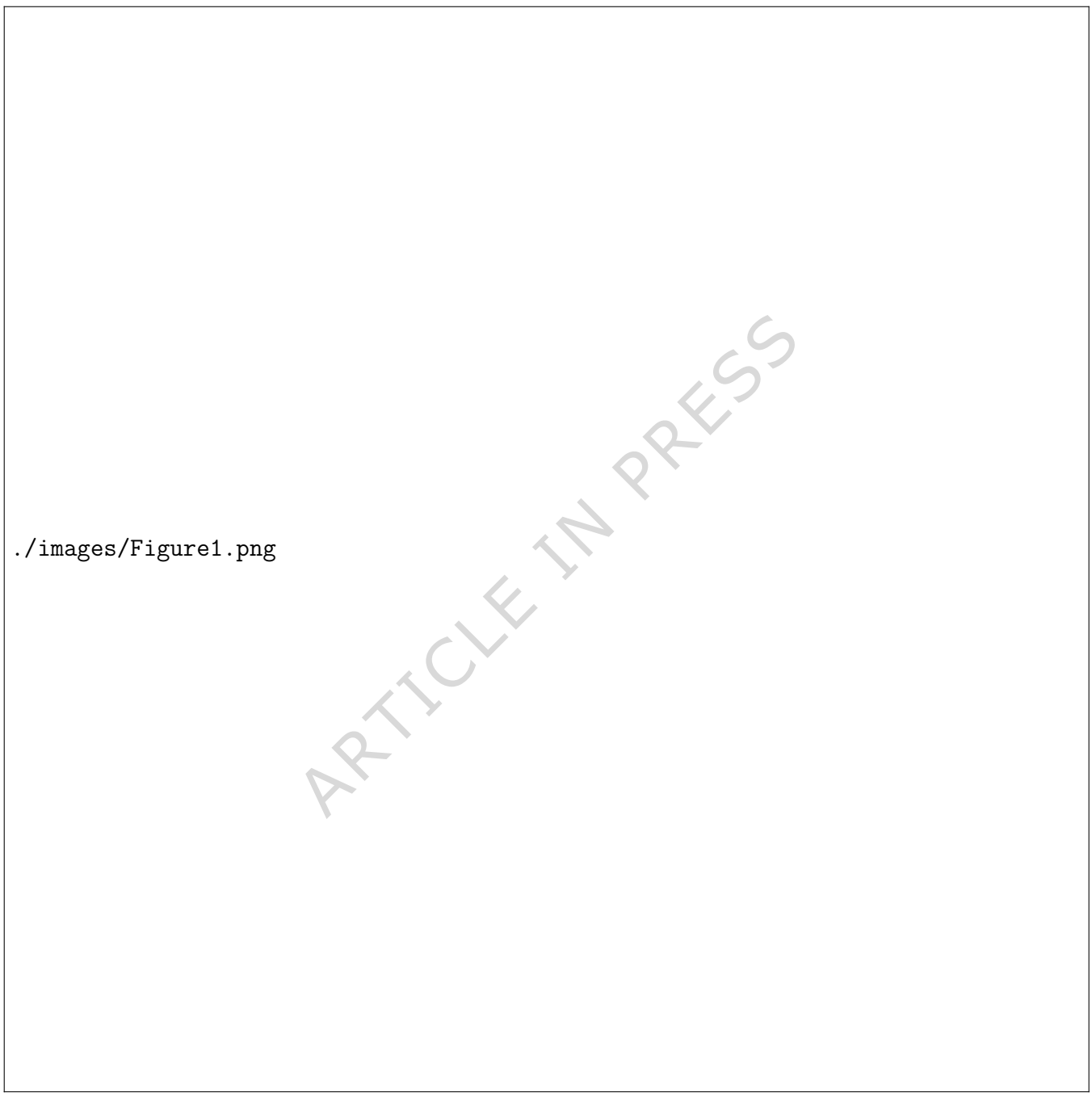
ARTICLE IN PRESS

ARTICLE IN PRESS

MAIN FIGURE TITLES AND LEGENDS

ARTICLE IN PRESS

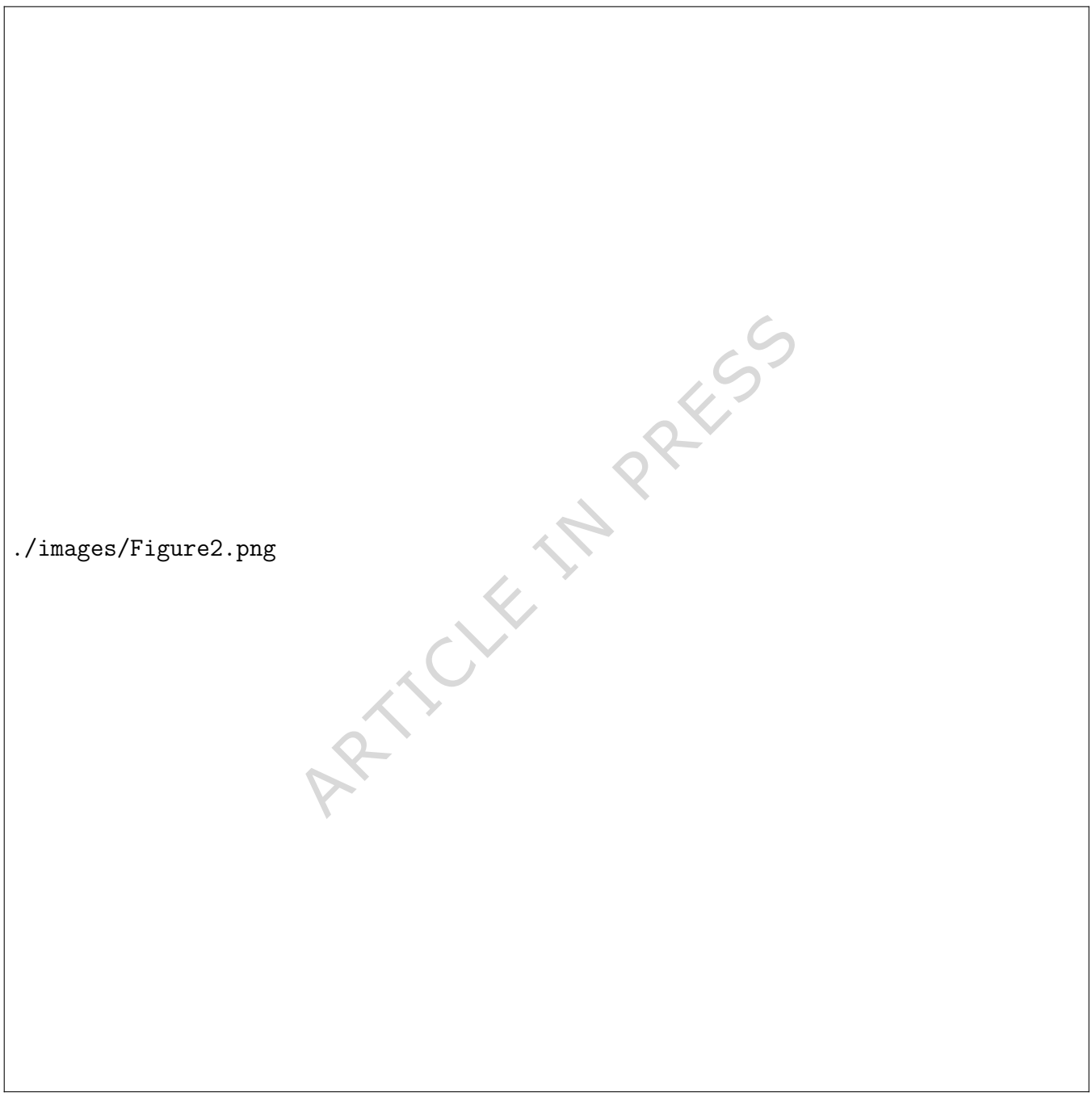
ARTICLE IN PRESS



ARTICLE IN PRESS

Fig 1: **Overall Framework of the HarveST Method**. **A** Heterogeneous graph construction involves three components: (i) a spot-spot subgraph based on spatial Euclidean distances, (ii) a spot-gene subgraph derived from gene expression quantiles within spots, and (iii) a gene-gene subgraph built on mutual information, normalized as weights. These subgraphs are integrated into a heterogeneous graph based on spots and genes. **B** The heterogeneous graph undergoes self-supervised learning with initial features for spots and genes derived from normalized gene expressions. GMM cluster the hidden features to calculate posterior probabilities for spots, with the top 50% spots assigned pseudo-labels for training an SVM classifier, which determines the final spatial domain identification. **C** Identification of spatially specific genes using the heterogeneous graph, where the association strength of each gene with cluster-specific seeds is quantified, and the steady-state probability values obtained from the RWR serve as association strength scores, with the top-ranking genes being considered highly related to the region. **D** Various downstream analyses and applications.

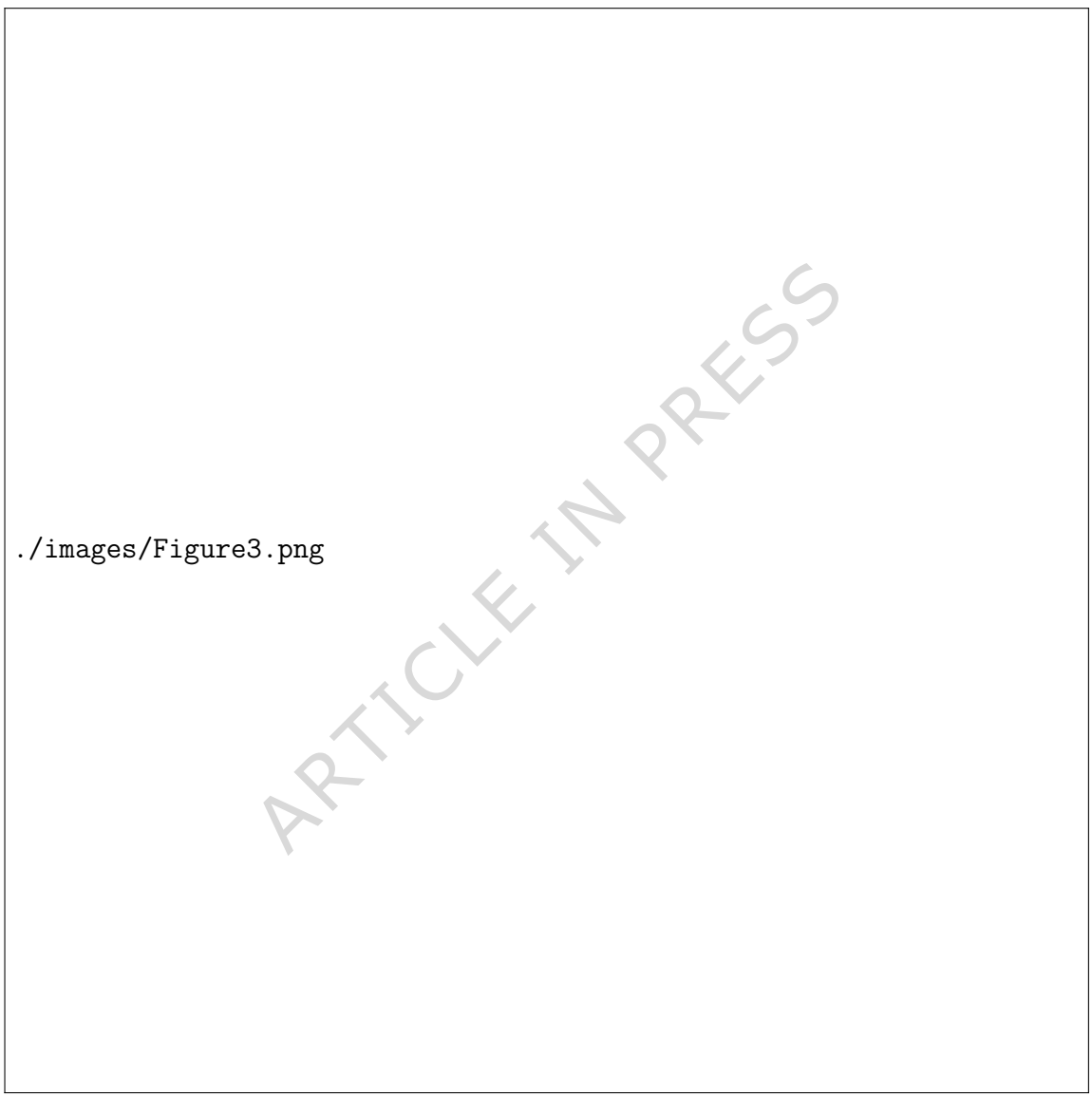
ARTICLE IN PRESS



ARTICLE IN PRESS

Fig 2: **Performance of HarveST and alternative methods in identifying spatial domains in the dorsolateral prefrontal cortex (DLPFC) and human breast cancer (HBRC) datasets. A** Box plots comparing the adjusted Rand index (ARI) scores of HarveST and baseline methods across 12 DLPFC slices. The center line of each box represents the median, box limits indicate the upper and lower quartiles, and whiskers extend to the 1.5$\times$ interquartile range. Statistics were derived from $n = 12$ biologically independent tissue sections. Statistical analysis confirms that HarveST yields significantly higher ARI scores compared to leading baseline methods (e.g., GraphST, $P = 0.017$; SpaGCN, $P < 0.001$; two-sided Wilcoxon signed-rank test). **B** Manually annotated layer structure for DLPFC slice 151674. **C–D** Clustering results obtained by HarveST and 13 alternative methods, including Scanpy, Seurat, SpaGCN, MNMST, SEDR, BayesSpace, Stagate, GraphST, stKeep, Impeller, and CellCharter, applied to slice 151674. **E–F** Corresponding histological image and manual pathological annotations based on hematoxylin and eosin (H&E) staining. **G–H** Comparison of clustering performance on the HBRC dataset by HarveST and alternative methods.

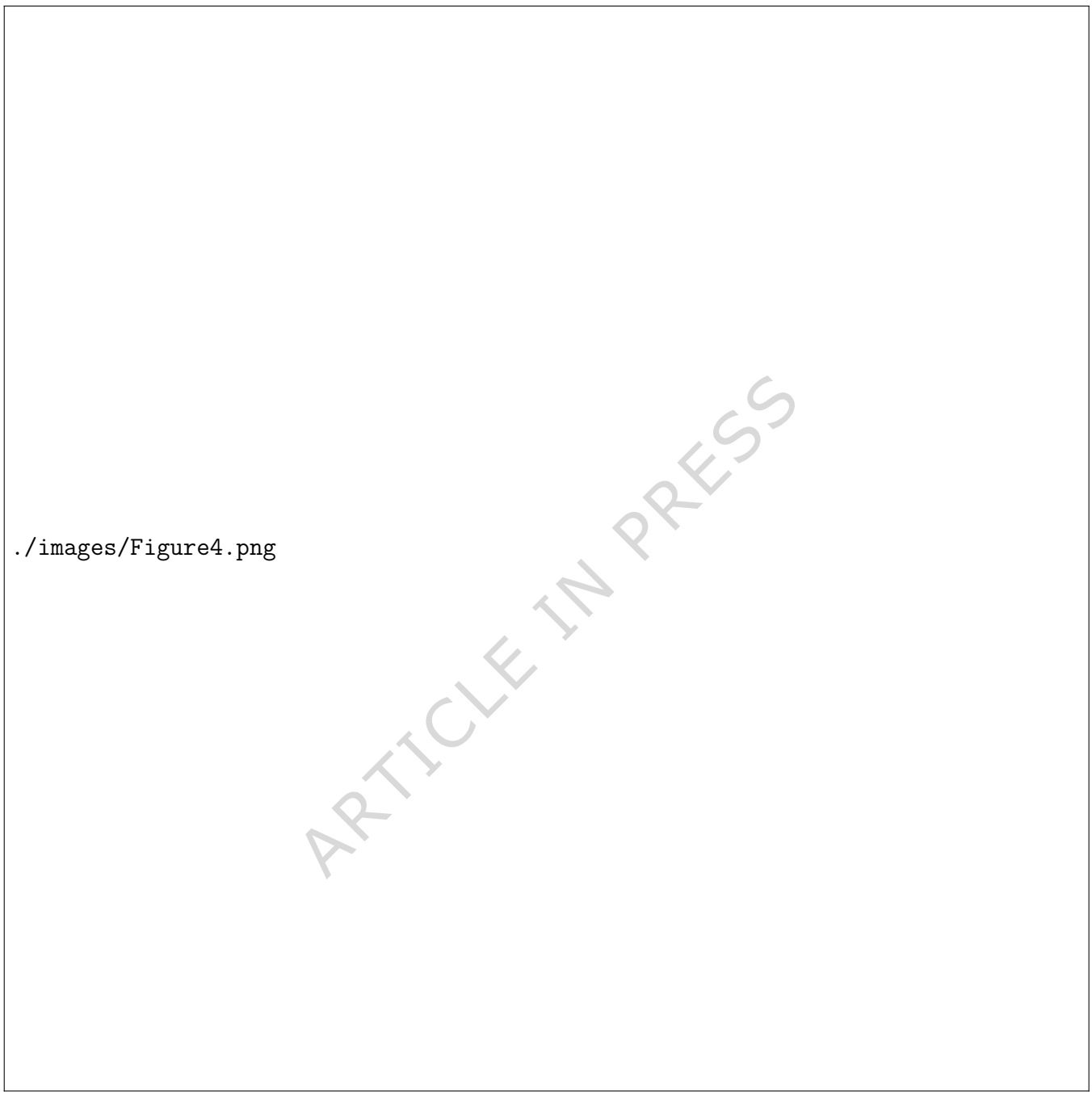
ARTICLE IN PRESS



ARTICLE IN PRESS

Fig 3: **Comparative analysis of laminar organization in the mouse olfactory bulb (MOB) tissue using Stereo-seq and Slide-seqV2 data.** (**A-D**) Visualization of Stereo-seq data. **A** Laminar organization of MOB annotated in the DAPI-stained image generated by Stereo-seq. **B** Spatial domains segmented by Scanpy, SEDR,Stagate, GraphST, MNMST, CellCharter-individual, Impeller and HarveST on the Stereo-seq data of the MOB tissue. **C** Visualization of each domain identified by HarveST in the Stereo-seq MOB tissue dataset. **D** Visualization of marker genes for each domain annotated by HarveST in the MOB tissue. (**E-H**) Visualization of Slide-seqV2 data. **E** Annotation of MOB tissue based on the Allen Reference Atlas. **F** Spatial domains segmented by Scanpy, SEDR, Stagate, GraphST, MNMST, CellCharter-individual, Impeller and HarveST on the Slide-seqV2 data of the MOB tissue. **G** Visualization of each domain identified by HarveST in the Slide-seqV2 of MOB tissue dataset. **H** Visualization of marker genes for each domain annotated by HarveST on the MOB tissue.

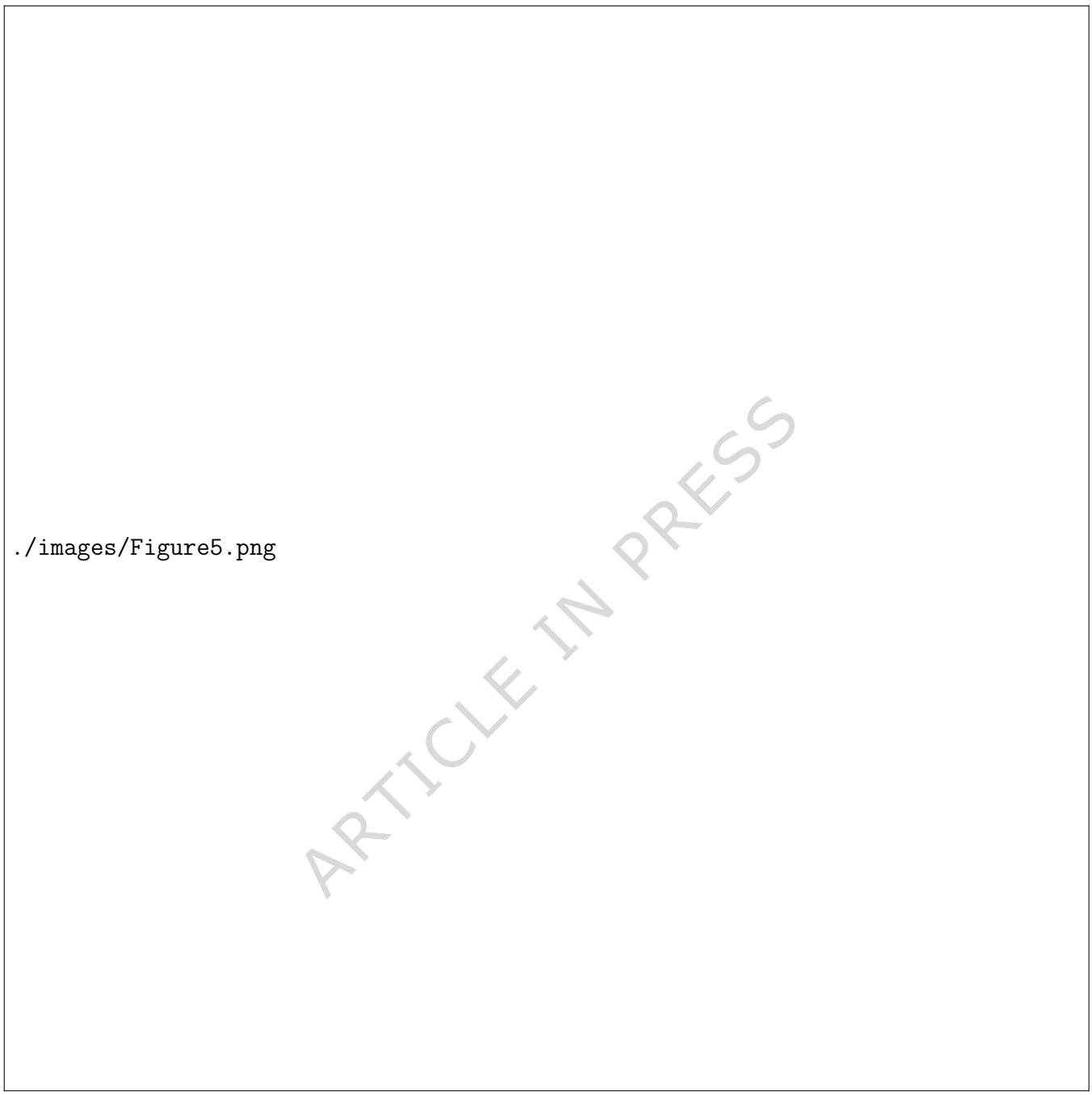
ARTICLE IN PRESS



ARTICLE IN PRESS

Fig 4: **Identification and verification of spatial domain-marker SVGs in DLPFC and HBRC datasets.** **A** Manually annotated white matter (WM) layer in DLPFC slice 151674. **B** Expression patterns of the top three genes (*MOBP*, *MOG*, and *MAG*) identified by HarveST as spatial domain–marker SVGs for the WM region. **C** Expression patterns of the top three genes (*BCAS1*, *CARNS1*, and *CLDN11*) identified by DESpace for the same region. **D** Manually annotated Layer 5 of slice 151674. **E** Expression patterns of the top three genes (*PCP4*, *HS3ST2*, and *SMYD2*) identified by HarveST as spatial domain–marker SVGs for Layer 5. **F** Expression patterns of the top three spatially variable genes identified by DESpace and Scanpy for Layer 5. **G** Manually annotated IDC 2 region of the HBRC dataset. **H** Expression patterns of the top three genes (*SLITRK6*, *C6orf141*, and *CRISP3*) identified by HarveST as spatial domain–marker SVGs for the IDC 2 region. **I** Expression patterns of the top three genes (*C6orf141*, *CRISP3*, and *EFHD1*) identified by DESpace for the IDC 2 region. **J** Manually annotated IDC 3 region of the HBRC dataset. **K** Expression patterns of the top three genes (*MS4A1*, *RASGRP2*, and *CD79A*) identified by HarveST as spatial domain–marker SVGs for the IDC 3 region. **L** Expression patterns of the top three genes (*CD52*, *TRBC2*, and *MS4A1*) identified by DESpace for the same region. **M–N** Visualization of domain-specific spatially variable genes (SVGs) for the IDC 2 (**M**) and IDC 3 (**N**) regions. Statistics were derived from $n = 383$ in IDC2 region and $n = 53$ in IDC3 region. Each dot represents a gene plotted according to its specificity score (x-axis; computed by HarveST) and fold change (y-axis) between the target region and all other regions. Colored markers (diamond, square, triangle) highlight genes also identified among the top 20 SVGs by comparison methods (DESpace, SpaGCN, and Scanpy), with overlapping symbols indicating consensus across different approaches. Statistical tests were performed on breast cancer tissue section comprising 3,798 spots.

ARTICLE IN PRESS



ARTICLE IN PRESS

Fig 5: **Comparative analysis of spatial domain-marker SVGs detected by various methods in human pancreatic cancer data. A** Annotation of the normal pancreatic region in the human pancreatic cancer dataset. **B-F, J-N**: Expression patterns of top three SVGs detected by comparison method. **B** Expression patterns of the top three SVGs (*AC009078.2, CTRC, PNLIPRP2*) identified by HarveST in the normal pancreatic region. **C** Expression patterns of the top three SVGs (*CTRB1, CTRB2, CPB1*) identified by DESpace in the normal pancreatic region. **D** Expression patterns of *PPY*, uniquely identified by HarveST among the top 20 genes detected by HarveST and alternative methods in the normal pancreatic region. **E** Expression patterns of the top three SVGs (*CTRB1, CTRB2, REG3A*) identified by Scanpy in the normal pancreatic region. **F** spatial domain-marker SVGs (*SPINK1, CTRB1*) identified by SpaGCN for the normal pancreatic region. **G** Visualization of domain-specific marker SVGs for the normal pancreatic region. Statistical tests were performed on breast cancer tissue section comprising 426 spots. **H** Gene enrichment analysis results for the Pancreatic region detected by HarveST. **I** Annotation of the cancer region in the human pancreatic cancer dataset. **J** Expression patterns of the top three SVGs (*LAMC2, SFN, KRT17*) identified by HarveST in the cancer region. **K** Expression patterns of the top three SVGs (*KRT17, KRT19, S100A6*) identified by DESpace in the cancer region. **L** Expression patterns of *IFI27*, uniquely identified by HarveST among the top 20 genes detected by HarveST and alternative methods in the cancer region. **M** Expression patterns of the top three SVGs (*KRT19, S100A6, KRT17*) identified by Scanpy in the cancer region. **N** Expression patterns of the top three SVGs (*KRT19, S100A6, PLEC*) identified by SpaGCN in the cancer region. **O** Visualization of domain-specific marker SVGs for the Cancer region. Statistical tests were performed on breast cancer tissue section comprising 426 spots. **P** KEGG pathway enrichment analysis of top 100 pancreatic region-specific genes identified by HarveST. **Q** KEGG pathway enrichment analysis of top 100 cancer region-specific genes identified by HarveST.

ARTICLE IN PRESS

References

[1] Asp, M. *et al.* A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. *Cell* **179**, 1647–1660 (2019).

[2] Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell–cell interactions and communication from gene expression. *Nature Reviews Genetics* **22**, 71–88 (2021).

[3] Dries, R. *et al.* Advances in spatial transcriptomic data analysis. *Genome research* **31**, 1706–1718 (2021).

[4] Ferreira, R. M. *et al.* Integration of spatial and single-cell transcriptomics localizes epithelial cell–immune cross-talk in kidney injury. *JCI insight* **6** (2021).

[5] Cheung, M. D. *et al.* Resident macrophage subpopulations occupy distinct microenvironments in the kidney. *JCI insight* **7** (2022).

[6] Zhou, R., Yang, G., Zhang, Y. & Wang, Y. Spatial transcriptomics in development and disease. *Molecular Biomedicine* **4**, 32 (2023).

[7] Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* **2008**, P10008 (2008).

[8] Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology* **33**, 495–502 (2015).

[9] Zhao, E. *et al.* Spatial transcriptomics at subspot resolution with bayesspace. *Nature biotechnology* **39**, 1375–1384 (2021).

[10] Xu, H. *et al.* Unsupervised spatially embedded deep representation of spatial transcriptomics. *Genome Medicine* **16**, 12 (2024).

[11] Dong, K. & Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature communications* **13**, 1739 (2022).

[12] Long, Y. *et al.* Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with graphst. *Nature Communications* **14**, 1155 (2023).

[13] Ren, H., Walker, B. L., Cang, Z. & Nie, Q. Identifying multicellular spatiotemporal organization of cells with spaceflow. *Nature communications* **13**, 4076 (2022).

[14] Xu, K., Xu, Y., Wang, Z., Zhou, X. M. & Zhang, L. stdyer enables spatial domain clustering with dynamic graph embedding. *Genome Biology* **26**, 1–25 (2025).

[15] Wang, Y., Liu, Z. & Ma, X. Mnmst: topology of cell networks leverages identification of spatial domains from spatial transcriptomics data. *Genome Biology* **25**, 133 (2024).

[16] Wang, H., Zhao, J., Nie, Q., Zheng, C. & Sun, X. Dissecting spatiotemporal structures in spatial transcriptomics via diffusion-based adversarial learning. *Research* **7**, 0390 (2024).

[17] Zuo, C., Xia, J. & Chen, L. Dissecting tumor microenvironment from spatially resolved transcriptomics data by heterogeneous graph learning. *Nature Communications* **15**, 5057 (2024).

ARTICLE IN PRESS

ARTICLE IN PRESS

[18] Duan, Z. *et al.* Impeller: a path-based heterogeneous graph learning method for spatial transcriptomic data imputation. *Bioinformatics* **40**, btae339 (2024).

[19] Varrone, M., Tavernari, D., Santamaria-Martínez, A., Walsh, L. A. & Ciriello, G. Cellcharter reveals spatial cell niches associated with tissue remodeling and cell plasticity. *Nature genetics* **56**, 74–84 (2024).

[20] Ji, A. L. *et al.* Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell* **182**, 497–514 (2020).

[21] Stickels, R. R. *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqv2. *Nature biotechnology* **39**, 313–319 (2021).

[22] Yan, G., Hua, S. H. & Li, J. J. Categorization of 34 computational methods to detect spatially variable genes from spatially resolved transcriptomics data. *Nature Communications* **16**, 1141 (2025).

[23] Edsgärd, D., Johnsson, P. & Sandberg, R. Identification of spatial expression trends in single-cell gene expression data. *Nature methods* **15**, 339–342 (2018).

[24] Svensson, V., Teichmann, S. A. & Stegle, O. Spatialde: identification of spatially variable genes. *Nature methods* **15**, 343–346 (2018).

[25] Sun, S., Zhu, J. & Zhou, X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature methods* **17**, 193–200 (2020).

[26] Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology* **19**, 1–5 (2018).

[27] Hu, J. *et al.* Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature methods* **18**, 1342–1351 (2021).

[28] Cai, P., Robinson, M. D. & Tiberi, S. Despace: spatially variable gene detection via differential expression testing of spatial clusters. *Bioinformatics* **40**, btae027 (2024).

[29] Maynard, K. R. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience* **24**, 425–436 (2021).

[30] Shi, X., Zhu, J., Long, Y. & Liang, C. Identifying spatial domains of spatially resolved transcriptomics via multi-view graph convolutional networks. *Briefings in Bioinformatics* **24**, bbad278 (2023).

[31] Chen, A. *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell* **185**, 1777–1792 (2022).

[32] Charych, E. I., Liu, F., Moss, S. J. & Brandon, N. J. Gabaa receptors and their associated proteins: Implications in the etiology and treatment of schizophrenia and related disorders. *Neuropharmacology* **57**, 481–495 (2009).

[33] Kadowaki, K. *et al.* Phosphohippolin expression in the rat central nervous system. *Molecular brain research* **125**, 105–112 (2004).

[34] Sunkin, S. M. *et al.* Allen brain atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic acids research* **41**, D996–D1008 (2012).

ARTICLE IN PRESS

ARTICLE IN PRESS

[35] Zacharias, D. A. & Kappen, C. Developmental expression of the four plasma membrane calcium atpase (pmca) genes in the mouse. *Biochimica et Biophysica Acta (BBA)-General Subjects* **1428**, 397–405 (1999).

[36] Castelli, L. M. *et al.* Srsf1-dependent inhibition of c9orf72-repeat rna nuclear export: genome-wide mechanisms for neuroprotection in amyotrophic lateral sclerosis. *Molecular neurodegeneration* **16**, 53 (2021).

[37] Neftel, C. *et al.* An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell* **178**, 835–849 (2019).

[38] Wei, Y., Yang, X., Gao, L., Xu, Y. & Yi, C. Differences in potential key genes and pathways between primary and radiation-associated angiosarcoma of the breast. *Translational oncology* **19**, 101385 (2022).

[39] Plasschaert, L. W. *et al.* A single-cell atlas of the airway epithelium reveals the cftr-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).

[40] Young, M. D. *et al.* Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *science* **361**, 594–599 (2018).

[41] Yamada, A. *et al.* High expression of atp-binding cassette transporter abcc11 in breast tumors is associated with aggressive subtypes and low disease-free survival. *Breast cancer research and treatment* **137**, 773–782 (2013).

[42] O'Flanagan, C. H. *et al.* Dissociation of solid tumor tissues with cold active protease for single-cell rna-seq minimizes conserved collagenase-associated stress responses. *Genome biology* **20**, 1–13 (2019).

[43] Shikang, Z., Xin, J. & Song, X. Expression of rasgrp2 in lung adenocarcinoma and its effect on immune microenvironment. *Zhongguo Fei Ai Za Zhi* **24** (2021).

[44] Wu, S. Z. *et al.* Stromal cell diversity associated with immune evasion in human triple-negative breast cancer. *The EMBO journal* **39**, e104063 (2020).

[45] Xydia, M. *et al.* Common clonal origin of conventional t cells and induced regulatory t cells in breast cancer patients. *Nature Communications* **12**, 1119 (2021).

[46] Kester, L. *et al.* Differential survival and therapy benefit of patients with breast cancer are characterized by distinct epithelial and immune cell microenvironments. *Clinical Cancer Research* **28**, 960–971 (2022).

[47] Li, H., Calder, C. A. & Cressie, N. Beyond moran's i: testing for spatial dependence based on the spatial autoregressive model. *Geographical analysis* **39**, 357–375 (2007).

[48] Abdelaal, T., Mourragui, S., Mahfouz, A. & Reinders, M. J. Spage: spatial gene enhancement using scrna-seq. *Nucleic acids research* **48**, e107–e107 (2020).

[49] Cable, D. M. *et al.* Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature biotechnology* **40**, 517–526 (2022).

[50] Moncada, R. *et al.* Integrating microarray-based spatial transcriptomics and single-cell rna-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nature biotechnology* **38**, 333–342 (2020).

ARTICLE IN PRESS

ARTICLE IN PRESS

[51] De Kanter, J. K., Lijnzaad, P., Candelli, T., Margaritis, T. & Holstege, F. C. Chetah: a selective, hierarchical cell type identification method for single-cell rna sequencing. *Nucleic acids research* **47**, e95–e95 (2019).

[52] Qadir, M. M. F. *et al.* Single-cell resolution analysis of the human pancreatic ductal progenitor cell niche. *Proceedings of the National Academy of Sciences* **117**, 10876–10887 (2020).

[53] Gonçalves, C. A. *et al.* A 3d system to model human pancreas development and its reference single-cell transcriptome atlas identify signaling pathways required for progenitor expansion. *Nature communications* **12**, 3144 (2021).

[54] Tosti, L. *et al.* Single-nucleus and in situ rna–sequencing reveal cell topographies in the human pancreas. *Gastroenterology* **160**, 1330–1344 (2021).

[55] Sun, H. *et al.* Dissecting the heterogeneity and tumorigenesis of brca1 deficient mammary tumors via single cell rna sequencing. *Theranostics* **11**, 9967 (2021).

[56] Li, C., Guo, L., Li, S. & Hua, K. Single-cell transcriptomics reveals the landscape of intra-tumoral heterogeneity and transcriptional activities of ecs in cc. *Molecular Therapy-Nucleic Acids* **24**, 682–694 (2021).

[57] Kim, J. Y. *et al.* Stratifin (sfn) regulates lung cancer progression via nucleating the vps34-becn1-traf6 complex for autophagy induction. *Clinical and translational medicine* **12** (2022).

[58] Egelston, C. A. *et al.* Tumor-infiltrating exhausted cd8+ t cells dictate reduced survival in premenopausal estrogen receptor–positive breast cancer. *JCI insight* **7** (2022).

[59] Merz, M. *et al.* Deciphering spatial genomic heterogeneity at a single cell resolution in multiple myeloma. *Nature communications* **13**, 807 (2022).

[60] Ge, P. *et al.* Identifying drug candidates for pancreatic ductal adenocarcinoma based on integrative multiomics analysis. *Journal of Gastrointestinal Oncology* **15**, 1265 (2024).

[61] Zhang, M. *et al.* Development and validation of cancer-associated fibroblasts-related gene landscape in prognosis and immune microenvironment of bladder cancer. *Frontiers in Oncology* **13**, 1174252 (2023).

[62] Han, J., DePinho, R. A. & Maitra, A. Single-cell rna sequencing in pancreatic cancer. *Nature reviews Gastroenterology & hepatology* **18**, 451–452 (2021).

[63] Majdalawieh, A. F., Massri, M. & Ro, H.-S. Aebp1 is a novel oncogene: mechanisms of action and signaling pathways. *Journal of oncology* **2020**, 8097872 (2020).

[64] Cortes, C. Support-vector networks. *Machine Learning* (1995).

[65] Maynard, K. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex [data set] (2021). URL `http://spatial.libd.org/spatialLIBD/`.

[66] Chen, A. *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays [data set] (2022). URL `https://db.cngb.org/stomics/mosta/`. Accession: CNP0001543.

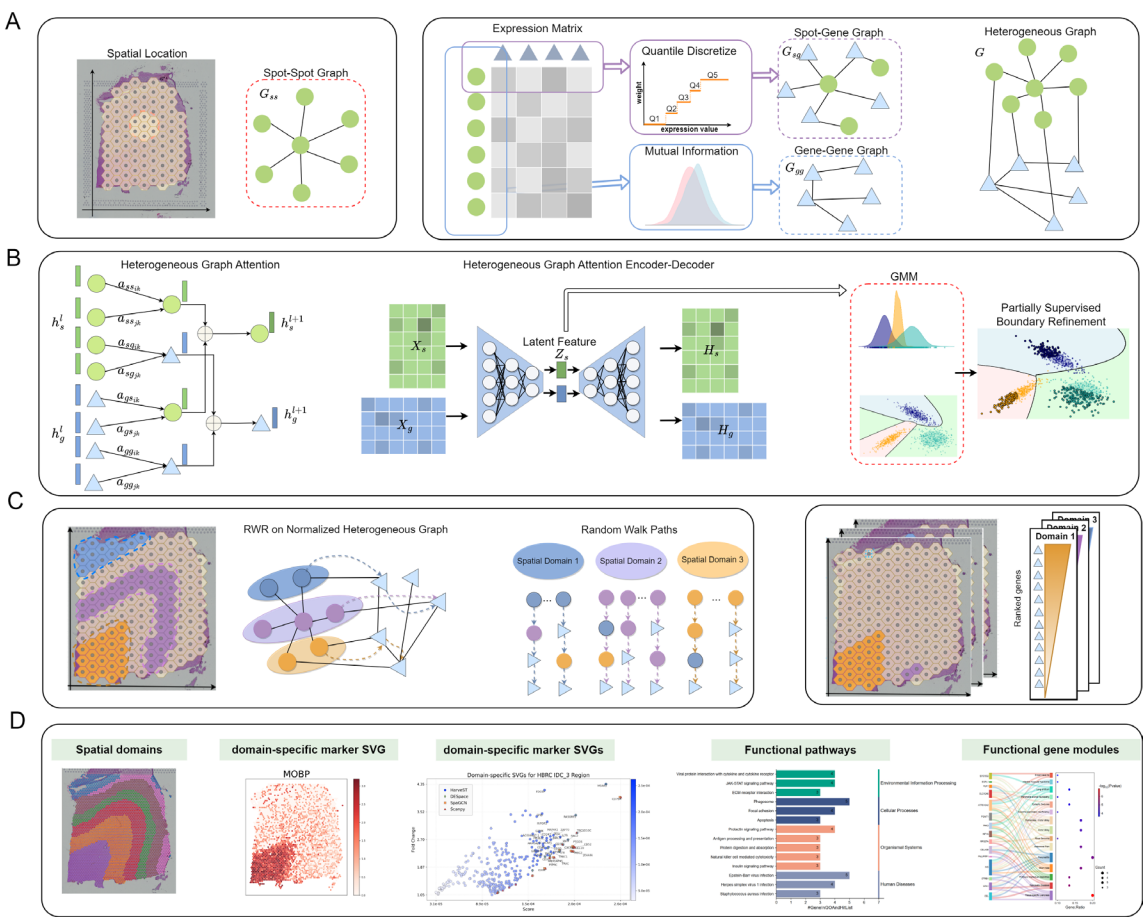
ARTICLE IN PRESS

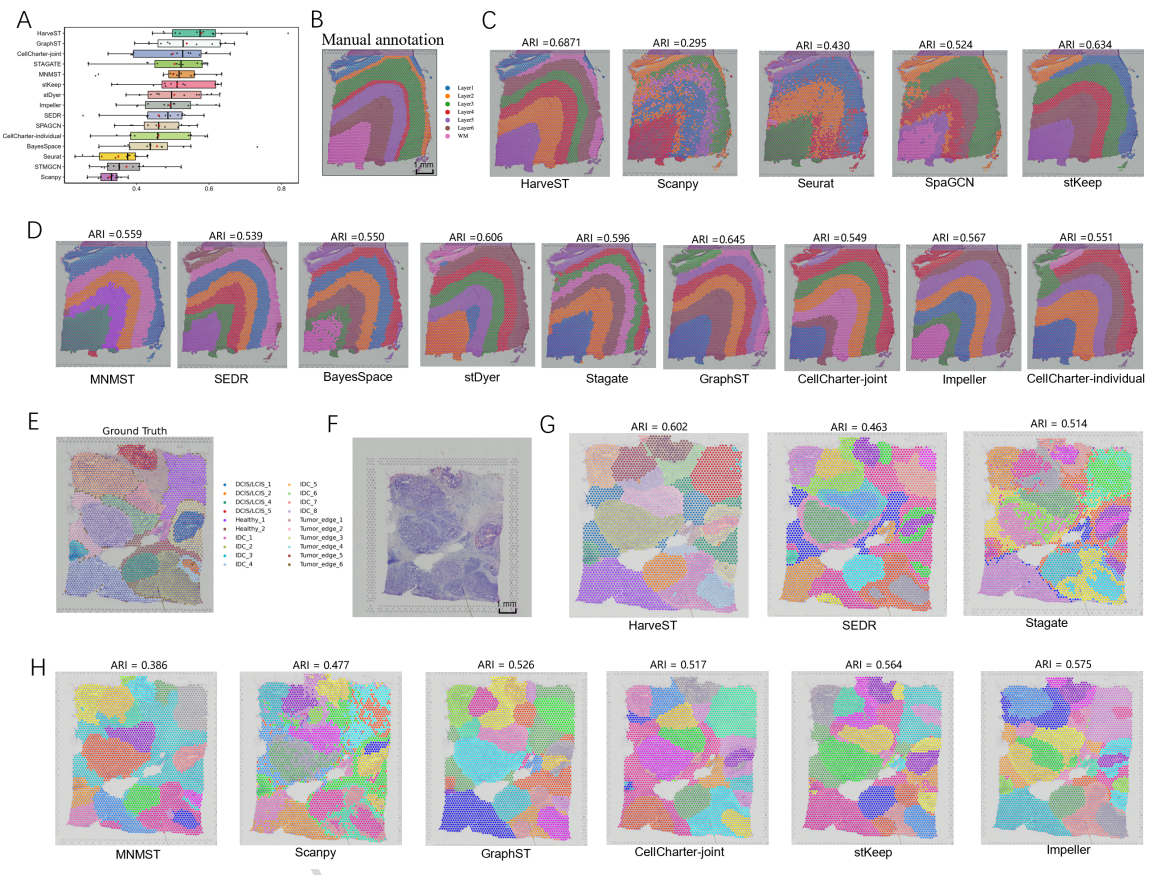
[67] Stickels, R. *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqv2 [data set] (2021). URL https://singlecell.broadinstitute.org/single_cell/study/SCP815. Study ID: SCP815.

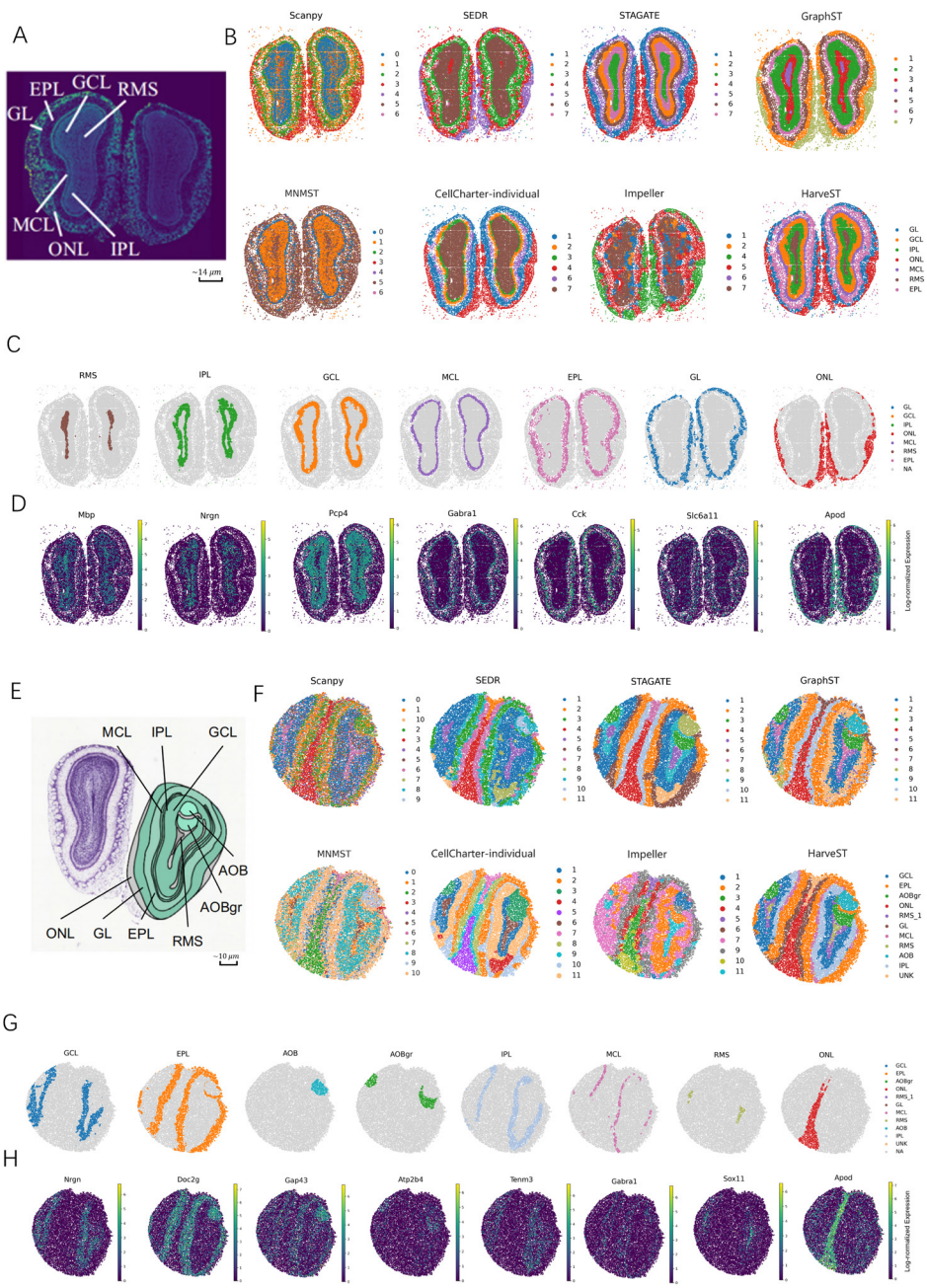
[68] Moncada, R. *et al.* Integrating microarray-based spatial transcriptomics and single-cell rna-seq reveals tissue architecture in pancreatic ductal adenocarcinomas [data set] (2020). URL https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111672. Accession: GSE111672.

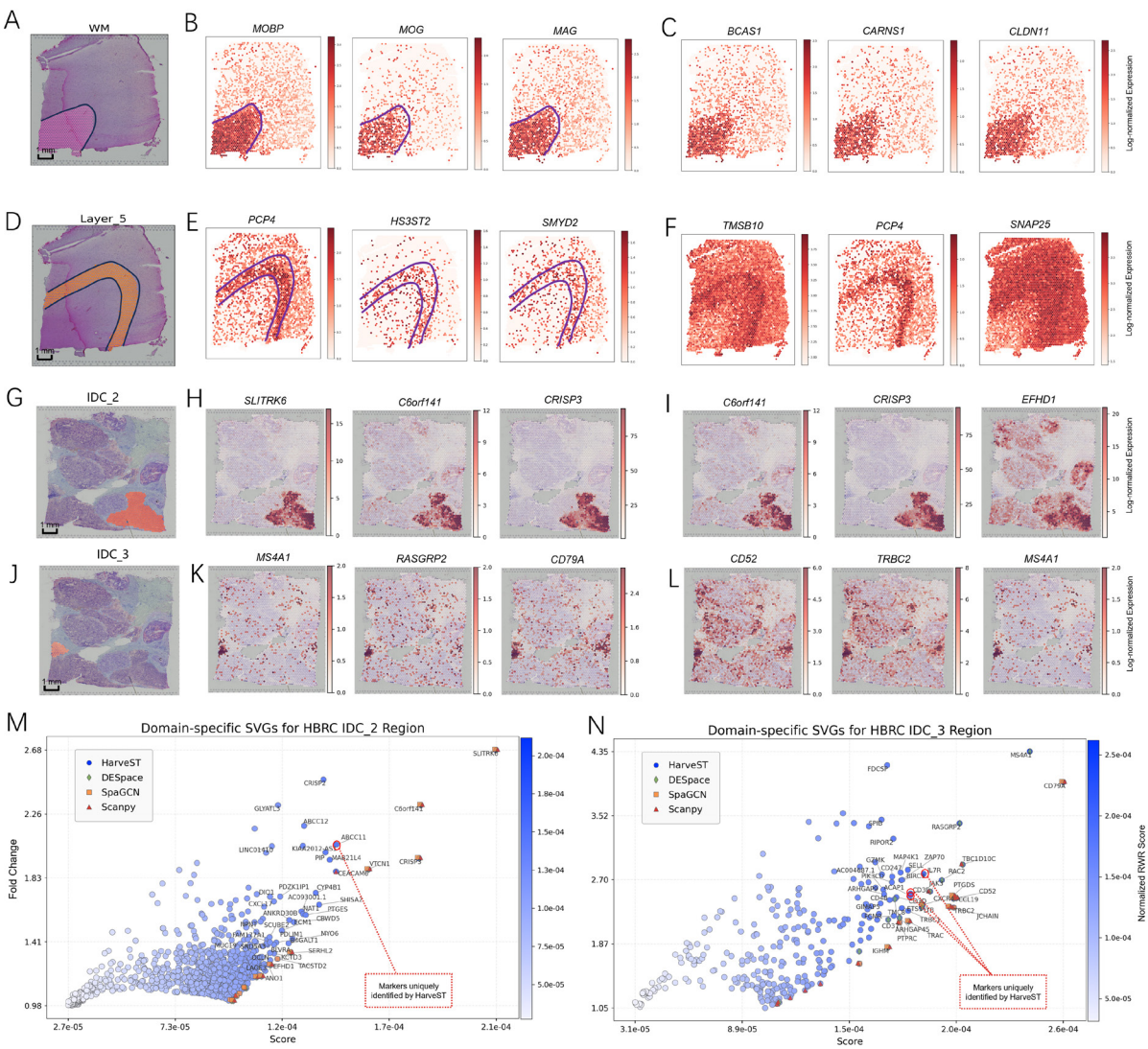
[69] Feng, J., Yu, T. & Zhang, Y. Harvest: Heterogeneous graph learning framework for spatial transcriptomics [data set/software] (2025). URL https://doi.org/10.5281/zenodo.18532348.

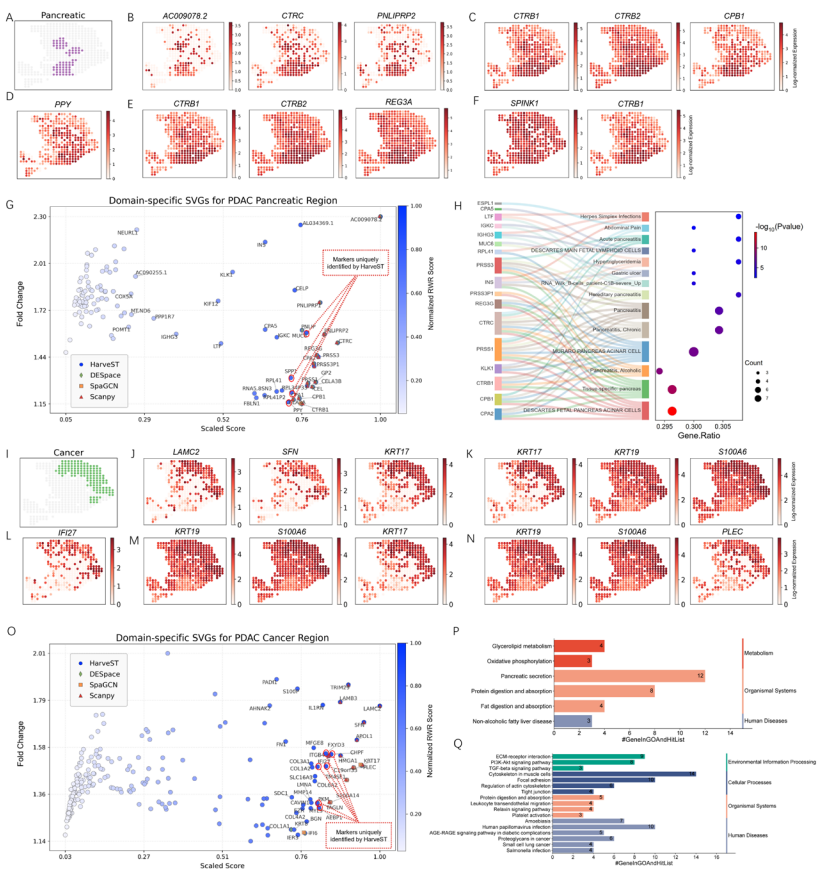
ARTICLE IN PRESS


A
Spatial Location
Spot-Spot Graph
Expression Matrix
Quantile Discretize
Spot-Gene Graph
Heterogeneous Graph
Mutual Information
Gene-Gene Graph
B
Heterogeneous Graph Attention
Heterogeneous Graph Attention Encoder-Decoder
Latent Feature
GMM
Partially Supervised
Boundary Refinement
C
RWR on Normalized Heterogeneous Graph
Random Walk Paths
Spatial Domain 1
Spatial Domain 2
Spatial Domain 3
Ranked genes
D
Spatial domains
domain-specific marker SVG
MOBP
domain-specific marker SVGs
Functional pathways
Functional gene modules


A
HarveST
GraphST
CellCharter-joint
STAGATE
MNMST
stKeep
stDyer
Impeller
SEDR
SPAGCN
CellCharter-individual
BayesSpace
Seurat
STMGCN
Scanpy
0.4
0.6
0.8
B
Manual annotation
Layer1
Layer2
Layer3
Layer4
Layer5
Layer6
WM
1 mm
C
ARI =0.6871
ARI =0.295
ARI =0.430
ARI =0.524
ARI =0.634
HarveST
Scanpy
Seurat
SpaGCN
stKeep
D
ARI =0.559
ARI =0.539
ARI =0.550
ARI =0.606
ARI =0.596
ARI =0.645
ARI =0.549
ARI =0.567
ARI =0.551
MNMST
SEDR
BayesSpace
stDyer
Stagate
GraphST
CellCharter-joint
Impeller
CellCharter-individual
E
Ground Truth
DCIS/LCIS_1
DCIS/LCIS_2
DCIS/LCIS_4
DCIS/LCIS_5
Healthy_1
Healthy_2
IDC_1
IDC_2
IDC_3
IDC_4
IDC_5
IDC_6
IDC_7
IDC_8
Tumor_edge_1
Tumor_edge_2
Tumor_edge_3
Tumor_edge_4
Tumor_edge_5
Tumor_edge_6
F
1 mm
G
ARI = 0.602
ARI = 0.463
ARI = 0.514
HarveST
SEDR
Stagate
H
ARI = 0.386
ARI = 0.477
ARI = 0.526
ARI = 0.517
ARI = 0.564
ARI = 0.575
MNMST
Scanpy
GraphST
CellCharter-joint
stKeep
Impeller


A
EPL
GCL
RMS
GL
MCL
ONL
IPL
~14 μm
B
Scanpy
SEDR
STAGATE
GraphST
MNMST
CellCharter-individual
Impeller
HarveST
GL
GCL
IPL
ONL
MCL
RMS
EPL
C
RMS
IPL
GCL
MCL
EPL
GL
ONL
NA
D
Mbp
Nrgn
Pcp4
Gabra1
Cck
Slc6a11
Apod
Log-normalized Expression
E
MCL
IPL
GCL
AOB
AOBgr
ONL
GL
EPL
RMS
~10 μm
F
Scanpy
SEDR
STAGATE
GraphST
MNMST
CellCharter-individual
Impeller
HarveST
GCL
EPL
AOBgr
ONL
RMS_1
GL
MCL
RMS
AOB
IPL
UNK
G
GCL
EPL
AOB
AOBgr
IPL
MCL
RMS
ONL
H
Nrgn
Doc2g
Gap43
Atp2b4
Tenm3
Gabra1
Sox11
Apod
Log-normalized Expression


A
WM
B
MOBP
MOG
MAG
C
BCAS1
CARNS1
CLDN11
Log-normalized Expression
D
Layer_5
E
PCP4
HS3ST2
SMYD2
F
TMSB10
PCP4
SNAP25
Log-normalized Expression
G
IDC_2
H
SLITRK6
C6orf141
CRISP3
I
C6orf141
CRISP3
EFHD1
Log-normalized Expression
J
IDC_3
K
MS4A1
RASGRP2
CD79A
L
CD52
TRBC2
MS4A1
Log-normalized Expression
M
Domain-specific SVGs for HBRC IDC_2 Region
HarveST
DESpace
SpaGCN
Scanpy
Markers uniquely identified by HarveST
Fold Change
Score
Normalized RWR Score
N
Domain-specific SVGs for HBRC IDC_3 Region
HarveST
DESpace
SpaGCN
Scanpy
Markers uniquely identified by HarveST
Score
Normalized RWR Score


A
Pancreatic
B
AC009078.2
CTRC
PNLIPRP2
C
CTRB1
CTRB2
CPB1
Log-normalized Expression
D
PPY
E
CTRB1
CTRB2
REG3A
F
SPINK1
CTRB1
G
Domain-specific SVGs for PDAC Pancreatic Region
Fold Change
Scaled Score
Normalized FNNR Score
Markers uniquely identified by HarveST
HarveST
DESpace
SpaGCN
Scanpy
H
-log10(Pvalue)
Count
Gene.Ratio
I
Cancer
J
LAMC2
SFN
KRT17
K
KRT17
KRT19
S100A6
L
IFI27
M
KRT19
S100A6
KRT17
N
KRT19
S100A6
PLEC
O
Domain-specific SVGs for PDAC Cancer Region
Fold Change
Scaled Score
Normalized FNNR Score
Markers uniquely identified by HarveST
HarveST
DESpace
SpaGCN
Scanpy
P
Glycerolipid metabolism
Oxidative phosphorylation
Pancreatic secretion
Protein digestion and absorption
Fat digestion and absorption
Non-alcoholic fatty liver disease
Metabolism
Organismal Systems
Human Diseases
#GeneInGOAndHitList
Q
ECM-receptor interaction
PI3K-Akt signaling pathway
TGF-beta signaling pathway
Cytokine-cytokine receptor
Focal adhesion
Regulation of actin cytoskeleton
Tight junction
Protein digestion and absorption
Leukocyte transendothelial migration
Relaxin signaling pathway
Pancreatic secretion
Amoebiasis
Human papillomavirus infection
AGE-RAGE signaling pathway in diabetic complications
Proteoglycans in cancer
Small cell lung cancer
Salmonella infection
Environmental Information Processing
Cellular Processes
Organismal Systems
Human Diseases
#GeneInGOAndHitList