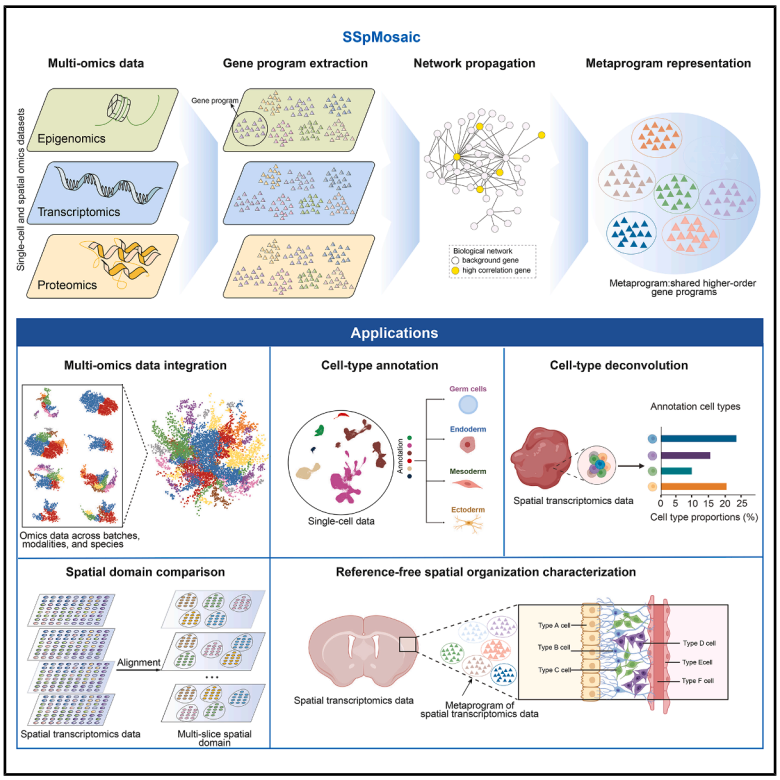
Cell Genomics

Technology

Robust integration and annotation of single-cell and spatial omics data using interpretable gene programs

Graphical abstract



Highlights

- Unifies single-cell and spatial omics using interpretable gene programs
- Enables robust single-cell integration and precise cell-type annotation
- Achieves resolution-agnostic deconvolution from spot to subcellular level
- Identifies conserved spatial ecotypes without single-cell references

Authors

Yuelei Zhang, Wenxuan Ming, Bianjiong Yu, ..., Yanhong Ni, Runzhi Deng, Dijun Chen

Correspondence

njdrz@nju.edu.cn (R.D.), dijunchen@nju.edu.cn (D.C.)

In brief

Zhang et al. introduce SSpMosaic, a computational framework that leverages interpretable gene programs to unify single-cell and spatial omics data. SSpMosaic enables robust integration, accurate annotation, and reference-free spatial mapping, advancing understanding of cellular organization and functional architecture in complex tissues.

Zhang et al., 2026, Cell Genomics 6, 101105
April 8, 2026
https://doi.org/10.1016/j.xgen.2025.101105

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Technology

Robust integration and annotation of single-cell and spatial omics data using interpretable gene programs

Yuelei Zhang,[1,5] Wenxuan Ming,[1,5] Bianjiong Yu,[1] Lele Wang,[2] Kaiyan Lu,[1] Lei Xu,[3] Yanhong Ni,[1] Runzhi Deng,[1,*] and Dijun Chen[1,3,4,6,*]

[1]Nanjing Stomatology Hospital, Medical School and the State Key Laboratory of Pharmaceutical Biotechnology, Nanjing University, Nanjing, Jiangsu 210023, China
[2]West China School of Public Health and West China Fourth Hospital, Sichuan University, Chengdu, Sichuan 610041, China
[3]Department of Gastroenterology, Nanjing Drum Tower Hospital, Nanjing University, Nanjing, Jiangsu 210008, China
[4]Chemistry and Biomedicine Innovation Center, Nanjing University, Nanjing, Jiangsu 210023, China
[5]These authors contributed equally
[6]Lead contact
*Correspondence: njdrz@nju.edu.cn (R.D.), dijunchen@nju.edu.cn (D.C.)
https://doi.org/10.1016/j.xgen.2025.101105

SUMMARY

Cellular identity emerges from the dynamic coordination of context-aware gene programs that encode biological functions across molecular layers. To decode this complexity, we present SSpMosaic, a computational framework that establishes metaprograms (higher-order, cross-dataset aligned gene program representations) as universal anchors for biological state representation. Leveraging these metaprograms, SSpMosaic enables consistent, accurate integration across batches, modalities, and species. Critically, SSpMosaic accurately annotates cell types within query datasets, enabling discovery and annotation of novel cell states through metaprogram-based transfer learning. The framework achieves resolution-agnostic spatial transcriptomics deconvolution, precisely mapping cell-type distributions from spot-level (Visium) to subcellular scales (CosMx/Visium HD). As a paradigm-shifting application, we integrate single-nucleus transcriptomics, chromatin accessibility, and spatial transcriptomics to resolve multi-stage spatial domain dynamics across tissue slices. Finally, SSpMosaic enables reference-free spatial characterization, identifying conserved spatial ecotypes across tissue slices and annotating cellular niches without requiring matched single-cell data.

INTRODUCTION

Single-cell and spatial omics technologies have revolutionized our understanding of cellular heterogeneity, tissue architecture, and disease mechanisms.[1,2] In cancer biology, single-cell RNA sequencing (scRNA-seq) has revealed transcriptional diversity and complex tumor-immune-stromal interactions,[3,4] while spatial transcriptomics has mapped microenvironmental organization underlying processes like immune evasion and metastasis.[5,6] However, integrating these diverse datasets remains challenging, limiting their full potential.

Each omics modality captures a unique but complementary aspect of cellular biology. For example, scRNA-seq provides high-resolution transcriptional profiles but lacks spatial context. Single-cell ATAC sequencing (scATAC-seq) maps chromatin accessibility to infer regulatory potential, yet lacks direct gene expression measurement. Spatial transcriptomics preserves tissue architecture but often at lower resolution or with gene sparsity. Integrating these layers is essential for a comprehensive view of cellular organization but remains challenging due to batch effects, modality-specific noise, varying resolutions, and high dimensionality.

Traditional deep learning models for multi-omics integration often function as ''black boxes,'' prioritizing predictive accuracy over transparency and biological interpretability, thereby limiting their utility for mechanistic discovery. A promising alternative leverages gene programs—sets of co-regulated or co-expressed genes that define specific cellular states or processes.[7–11] This approach reduces the complexity and noise of high-dimensional data, enhances interpretability, and uncovers shared regulatory mechanisms across modalities. However, existing frameworks based on gene programs or metaprograms[7,8,12,13] often lack the flexibility to handle multi-omics complexity or incorporate spatial dependencies, restricting their application to cutting-edge technologies.

To address these gaps, we present SSpMosaic, a unified framework that leverages metaprograms as dynamic biological anchors within a meta-graph architecture for integrated single-cell and spatial omics analysis. Using an unsupervised strategy, SSpMosaic first identifies cell-specific gene programs within

individual datasets, then applies network propagation to derive shared "metaprograms" that serve as robust anchors for diverse downstream tasks—including cross-batch/cross-modality integration, cell-type annotation, and multi-resolution spatial analysis. Crucially, SSpMosaic overcomes a fundamental limitation in spatial omics by resolving multi-stage tissue dynamics through integrated chromatin accessibility, single-nucleus transcriptomics, and spatial architecture mapping across serial tissue slices. Furthermore, SSpMosaic pioneers reference-free spatial characterization, enabling the identification of conserved tissue ecotypes and cellular niches even in the absence of matched single-cell data—directly addressing the critical bottleneck of reference dependency in current spatial analyses. We demonstrate that SSpMosaic outperforms existing methods in accuracy, scalability, and biological interpretability, efficiently handling datasets with millions of cells.

DESIGN

Overview of the SSpMosaic workflow

We developed SSpMosaic, a robust and interpretable computational framework for integrating and annotating single-cell and spatial omics data. This workflow seamlessly integrates diverse datasets to extract biologically meaningful insights into molecular and cellular mechanisms (Figure 1).

SSpMosaic begins by harmonizing feature spaces across batches and modalities into a unified gene representation. It then constructs cluster-specific differential co-expression gene programs, which encapsulate distinct transcriptional and regulatory signatures. To enable cross-dataset integration, a network propagation algorithm assesses similarities between these programs across datasets. Hierarchical clustering of the resulting similarity matrix yields "metaprograms" (MPs)—higher-order representations of shared regulatory patterns. Cells or spatial spots are thus assigned MP labels, creating a unified framework for cross-dataset state and function interpretation (Figure 1A).

The derived metaprograms serve as robust anchors for a variety of downstream analytical tasks. For instance, SSpMosaic employs a graph-based strategy to compute nearest neighbors among MP labels, constructing a harmonization graph that captures relationships of cells or spots across batches and modalities. By capturing shared regulatory or expression patterns, this approach ensures robust integration while preserving the biological variability inherent in the data (Figure 1B). Additionally, metaprograms enhance annotation of cell types or spatial spots by linking gene programs with well-defined cellular identities or biological functions (Figure 1C). In single-cell analysis, SSpMosaic extracts gene programs that define specific cell types in a reference dataset and leverages the aligned metaprograms between reference and query datasets to achieve precise annotation of the query dataset (Figure 1C, top). For spatial transcriptomics, SSpMosaic adopts non-negative least squares (NNLS) optimization to deconvolve individual spatial spots into contributions from metaprograms derived from reference single-cell data (Figure 1C, bottom).

Spatial spots with similar contributions of metaprograms indicate coherent cell states and functional organization, while spatially colocalized regions with coherent cell-type composition are referred to as spatial domains[5] or spatial ecotypes.[14] The resulting deconvolution data can then be utilized for unsupervised clustering, enabling the alignment of spots across tissue slices into coherent regions, referred to as spatial domains. Each spatial domain is characterized by a unique pattern of spatially resolved gene programs, reflecting distinct transcriptional signatures that provide insights into the functional organization and cellular interactions within the tissue microenvironment (Figure 1D). When single-cell references are absent, SSpMosaic directly infers metaprograms from spatial data, enabling reference-free, data-driven spatial characterization and precise annotation (Figure 1E). Thus, metaprograms provide a robust foundation for comparative spatial domain analysis across samples.

RESULTS

SSpMosaic enables robust integration of single-cell data across batches, modalities, and species

To evaluate SSpMosaic's single-cell integration performance, we analyzed datasets across diverse batches, modalities, and species, benchmarking it against six high-performing methods[15] (FastMNN,[16] Harmony,[17] Seurat CCA,[18] Scanorama,[19] scGen,[20] and BBKNN[21]) and focusing on integration accuracy, batch correction, scalability, and biological interpretability. Specifically, performance was quantified using five metrics: adjusted Rand index (ARI), cell-type ASW (average silhouette width), normalized mutual information (NMI) for biological information preservation, graph connectivity (GC), and batch/modality ASW for batch or modality effect removal. These metrics assess the balance between batch-effect removal and preservation of biological signal. An overall integration metric (Avg score)[22] was calculated as the weighted average of these metrics.

In the cross-batch integration scenarios, we applied SSpMosaic to human colorectal cancer (CRC) scRNA-seq data from four independent batches.[23–25] SSpMosaic demonstrated exceptional performance in preserving cellular distinctions while effectively correcting batch effects (Figure 2A). It achieved an average integration score of 0.87, outperforming the six other methods tested by at least 4% (Figure 2B). Notably, SSpMosaic obtained the highest ARI of 0.91 and NMI of 0.93, alongside competitive batch ASW of 0.74 and cell-type ASW of 0.79 (Figure S1A). Importantly, it maintained distinct boundaries between closely related cell types, such as CD8 T cells and natural killer (NK) cells, where other methods showed significant overlap or misclassification (Figure 2A).

Integrating single-cell data across species presents significant challenges due to both biological differences, such as gene orthology and regulatory variations, as well as technical biases from distinct experimental platforms. We benchmarked SSpMosaic using a comprehensive single-cell atlas of human and mouse brains, comprising 380,000 cells across 11 independent batches.[26–29] SSpMosaic excelled in accurately distinguishing cell types across species, particularly in identifying subtle subtypes of neurons, such as excitatory (Ext) and inhibitory (Inh) neurons, which other methods struggled to differentiate (Figure 2C). This superior performance was reflected in its higher integration score of 0.89 (ARI: 0.93, NMI: 0.93, batch ASW: 0.87,

CellPress
OPEN ACCESS

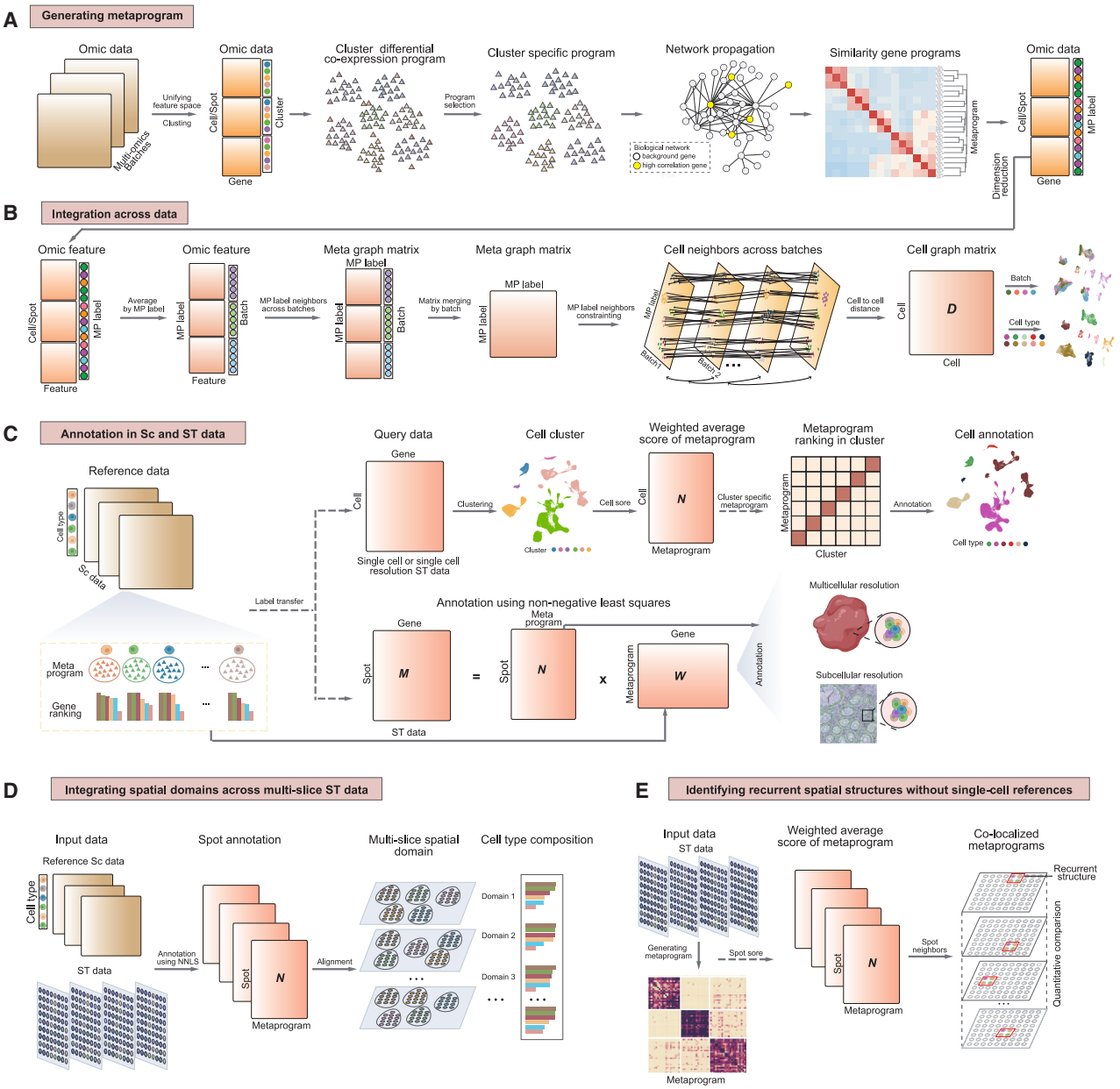


Figure 1. Schematic overview of SSpMosaic

(A) Construction of metaprograms across single-cell and spatial omics datasets. First, differential co-expression gene programs are constructed within each dataset. Cluster-specific programs are retained to enhance biological relevance. Next, a network propagation approach is employed to assess similarities among programs from different clusters across datasets. Finally, hierarchical clustering is applied to the similarity matrix, generating metaprograms (MPs) that represent higher-order functional gene programs, facilitating cross-sample and cross-modality integration. In this manner, cells or spatial spots are labeled with distinct MPs based on the origin of clusters.

(B) Workflow of data integration through the anchors of metaprograms. Graph-based neighbor identification among MP labels constructs a harmonization graph (matrix *D*) across batches/modalities.

(C) Cell/spatial annotation using reference metaprograms in (A). Single-cell query data are scored via weighted averaging of metaprogram genes; spatial spots are deconvolved via NNLS optimization using expression matrix *M* and metaprogram weight matrix *W* to obtain scoring matrix *N*.

(D) Cross-slice spatial transcriptomics analysis. Unsupervised clustering of spot-metaprogram scores (matrix *N*) identifies cell-type-consistent spatial domains and compositions.

(E) Reference-free spatial characterization. Direct metaprogram inference from multi-slice data enables co-localization analysis via metaprogram score patterns.

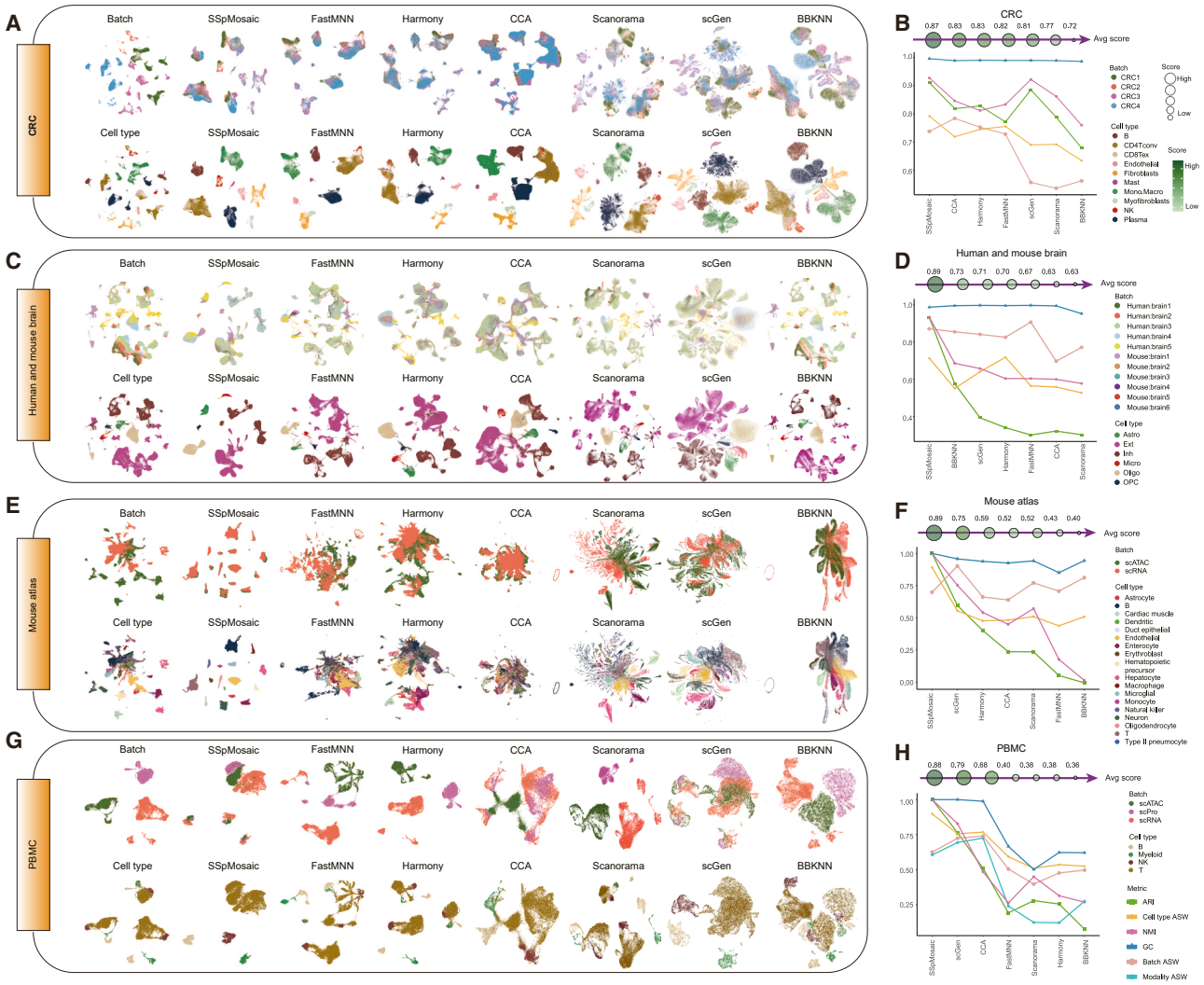


Figure 2. Integration performance across batches, modalities, and species

(A) UMAP visualization of seven methods on CRC cross-batch data (colored by batch/cell type).
(B) Quantitative evaluation of six integration metrics on CRC data.
(C) UMAP visualization of cross-species integration (human/mouse brain, 11 batches).
(D) Cross-species integration metrics across seven methods.
(E) UMAP visualization of mouse atlas data integration. Cells are colored by modality and cell type.
(F) Quantitative metrics for mouse atlas integration.
(G) UMAP visualization of PBMC multi-modal integration.
(H) PBMC integration performance across seven methods.

and cell-type ASW: 0.71; Figures 2D and S1B). The abovementioned results highlight SSpMosaic's ability to seamlessly integrate cross-species datasets while maintaining biological fidelity, making it a powerful tool for accurately capturing and comparing cellular dynamics across species.

Next, we assessed the cross-modality integration performance of SSpMosaic by evaluating its ability to harmonize datasets from different experimental modalities. Specifically, we integrated mouse scRNA-seq[30] and scATAC-seq data[31] from a comprehensive atlas spanning 18 cell types across multiple tissues (including heart, kidney, liver, and brain). SSpMosaic successfully aligned these datasets, effectively removing modality-specific artifacts while preserving cell-type-specific signals, achieving an overall integration score of 0.89 (Figures 2E and 2F). Competing methods, such as FastMNN, Harmony, CCA, and Scanorama, often over-corrected technical differences, leading to poor clustering performance and lower ARI and NMI scores (Figures 2F and S1C). BBKNN notably struggled with cell-type resolution, failing to accurately differentiate cell populations. While scGen showed relative effectiveness in modality alignment, it underperformed in preserving cell-type-specific signals compared to SSpMosaic (Figure 2E).

CellPress
OPEN ACCESS

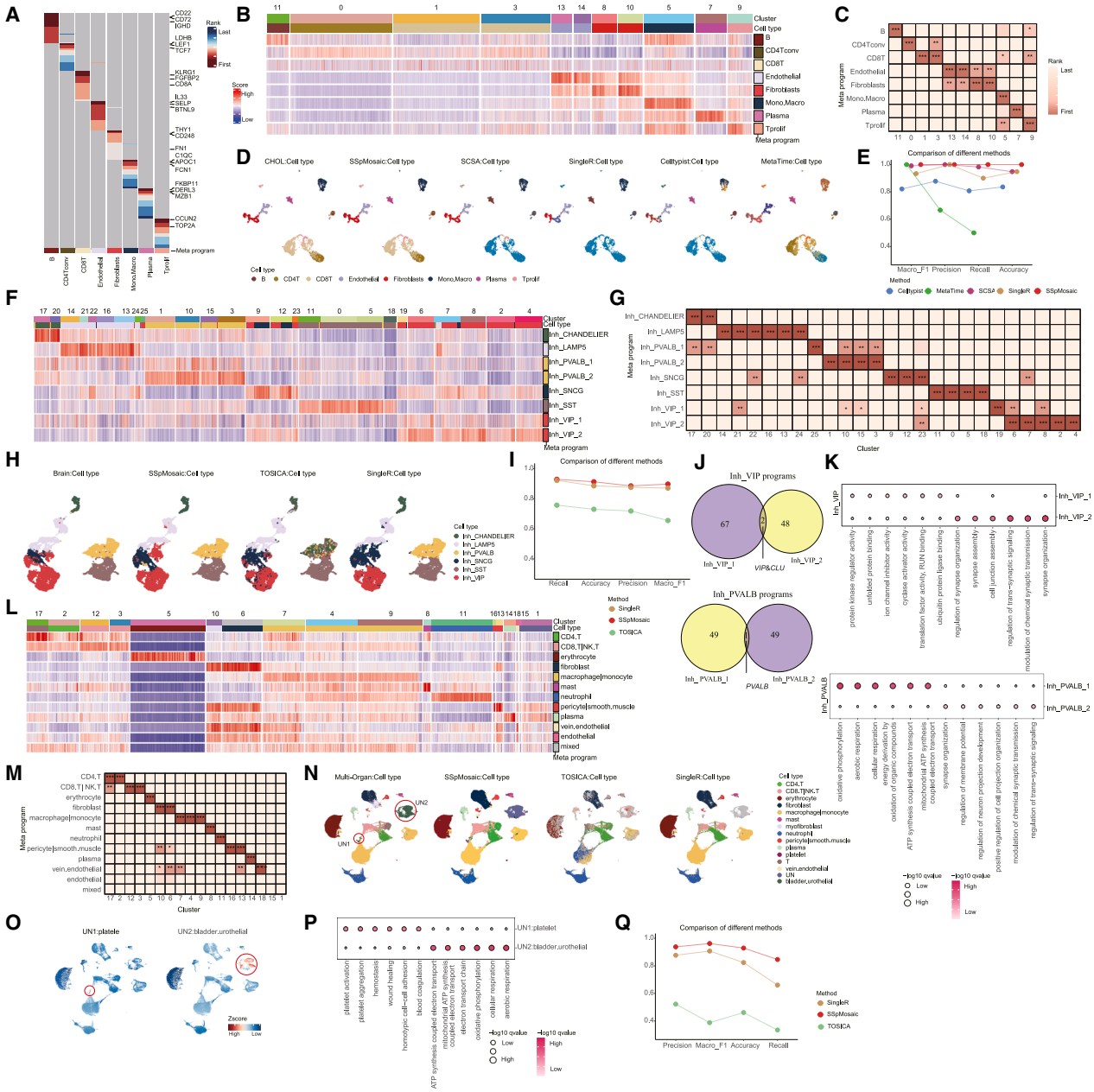


Figure 3. Annotation of scRNA data and performance comparison of SSpMosaic with other methods

(A) Gene ranking of metaprograms from pan-cancer scRNA datasets as reference, with high-weight genes indicated on the right.

(B) Heatmap displaying weighted average scores of all metaprograms across cells in CHOL dataset.

(C) Heatmap showing the correspondence between cell clusters and metaprograms from SSpMosaic in CHOL dataset.

(D) UMAP visualization showing ground truth labels and annotation results from five annotation methods in CHOL dataset.

(E) Performance comparison of five annotation methods in CHOL dataset using four evaluation metrics (Maro_F1, Precision, Recall, and Accuracy).

(F) Heatmap displaying weighted average scores of all metaprograms across cells in inhibitory neuron subtype dataset.

(G) Heatmap showing the correspondence between cell clusters and cell-type metaprograms in inhibitory neuron subtype dataset.

(H) UMAP visualization showing true labels and annotation results from three annotation methods in inhibitory neuron subtype query dataset.

(I) Performance comparison of SSpMosaic against two other automatic annotation tools on inhibitory neuron subtype dataset using four evaluation metrics.

(J) Venn diagram shows gene overlap between two metaprograms corresponding to the same cell subtype.

(K) GO enrichment bubble plot demonstrates distinct functional enrichment of two metaprograms from the same cell subtype, showing top six enriched pathways.

(L) Heatmap displaying weighted average scores of all metaprograms across cells in Tabula Sapiens cross-tissue query dataset.

(M) Heatmap showing the correspondence between cell clusters and cell-type metaprograms in Tabula Sapiens cross-tissue query dataset.

(legend continued on next page)

We further evaluated SSpMosaic using tri-modal PBMC data, which included single-cell RNA, ATAC, and protein modalities across four batches.[32] In the tested data, discrepancies in the expression of cell-type-specific markers among different modalities are particularly evident.[22] For instance, markers such as CD28 and CD40LG, which indicated T cells in RNA data, were absent in protein data. Similarly, CCR4, specific to the subset of T regulatory cells in transcriptomics, exhibited broader expression in protein data (Figures S1E and S1F). These variations can obscure biological signals and hinder cross-modality data integration and accurate cell-type identification. SSpMosaic leverages its metaprogram-based approach to overcome these discrepancies, aligning gene programs across modalities and effectively bridging modality-specific differences. Indeed, SSpMosaic achieved robust modality alignment, effective batch correction, and clear cell-type delineation (Figure 2G), achieving an integration score of 0.88 (Figure 2H). Although its batch ASW (0.63) and modality ASW (0.61) were slightly lower than those of Seurat CCA and scGen (batch ASW: 0.74 and 0.72; modality ASW: 0.73 and 0.70; Figures 2H and S1D), the overall integration score highlights SSpMosaic's ability to balance batch or modality correction with the preservation of critical biological signals.

SSpMosaic provides a metaprogram-based strategy for accurate cell-type annotation

The learned metaprograms from single-cell datasets are biologically interpretable higher-order features closely associated with cell types. As such, metaprograms inferred from reference datasets offer robust proxies for achieving precise and interpretable annotations of query datasets.

To evaluate this capability, we firstly applied SSpMosaic to pan-cancer single-cell transcriptomics datasets (as reference datasets) from the TISCH2 database,[33] which includes manually curated cell-type annotations. Metaprograms corresponding to each cell type were derived from the reference datasets, capturing unique co-expressed gene programs reflective of distinct cellular identities and functions (Figure 1C). Within each metaprogram, genes were ranked based on their concurrence with the gene program's overall expression pattern, prioritizing genes that contributed most strongly to the metaprogram's functional and cellular specificity. Interestingly, known cell-type markers were consistently ranked among the top genes in their associated metaprograms, highlighting the biological relevance and accuracy of the derived features (Figure 3A). This concordance underscores the ability of SSpMosaic to capture key gene programs that are both functionally relevant and cell-type-specific, reinforcing the interpretability and precision of its annotations. We applied this framework to annotate cell types in the independent cholangiocarcinoma (CHOL) dataset[34] to test generalization across datasets. Each CHOL cell was scored via weighted average expression of reference metaprogram genes, revealing cluster-specific metaprograms (Figure 3B). On the basis of the standard deviation difference in metaprogram scores across cell clusters, CHOL cell clusters were generally assigned to specific cell types based on their corresponding metaprogram activity (Figure 3C), thereby effectively linking distinct clusters to their underlying cellular identities (Figure 3D). We found that the cell-type assignments generated by SSpMosaic were highly consistent with manual expert annotations and demonstrated overall superior performance compared to five other widely used reference-free annotation methods, including MetaTime,[35] SingleR,[36] and SCSA[37] (Figures 3D and 3E). The above analyses highlight the robustness and accuracy of SSpMosaic-derived metaprograms in capturing cellular identities at the major cell-type level, showcasing their reliability for precise cell-type annotation across datasets.

Furthermore, we evaluated SSpMosaic's ability to predict cell types at the sub-cell-type level. Metaprograms were derived from a reference dataset,[38] capturing finely resolved gene expression patterns associated with specific sub-cellular states and identities. These metaprograms were then applied to annotate a high-resolution dataset of human neuronal subtypes,[27,29,38] enabling the identification of distinct neuronal subpopulations with enhanced granularity and biological relevance (Figures 3F–3H). We also compared SSpMosaic with two reference-based methods, TOSICA[39] and SingleR, demonstrating that SSpMosaic consistently outperformed these approaches in both accuracy and resolution of sub-cell-type predictions (Figure 3I). In particular, SingleR misclassified Inh_VIP and Inh_LAMP5 as Inh_SNCG, while TOSICA incorrectly annotated the majority of Inh_PVALB as Inh_CHANDELIER (Figure 3H). Of note, SSpMosaic further recognized fine subsets within neuronal subtypes, dividing Inh_PVALB into Inh_PVALB_1 and Inh_PVALB_2 and Inh_VIP into Inh_VIP_1 and Inh_VIP_2 (Figure 3H). Gene programs for these subsets retained common genes for the corresponding sub-type markers (*PVALB* for Inh_PVALB, *VIP* and *CLU* for Inh_VIP) while revealing distinct gene signatures that further differentiated the subsets (Figure 3J). The distinction was further supported by functional enrichment analysis of cell-subset gene programs, which identified unique biological pathways and molecular processes associated with each subset (Figure 3K). For example, in the subtype of Inh_VIP, Inh_VIP_1 was associated with stress response pathways, while Inh_VIP_2 was linked to typical neuronal functions such as synaptic organization and signal transduction (Figure 3K). In a nutshell, the abovementioned analyses demonstrate SSpMosaic's ability to capture subtle differences in gene expression that are crucial for understanding the complexity of cellular heterogeneity at the sub-cell-type level.

Lastly, we investigated the potential of SSpMosaic in annotating large-scale datasets and discovering previously

(N) UMAP visualization showing cross-tissue annotation results on Tabula Sapiens (red circles: absent reference types).
(O) UMAP visualization showing metaprogram scores from SSpMosaic for two UN-annotated cell clusters in Tabula Sapiens cross-tissue query dataset.
(P) GO enrichment bubble plot showing top six enriched pathways for metaprograms corresponding to UN-annotated cell clusters.
(Q) Performance comparison of SSpMosaic against two other automatic annotation tools on Tabula Sapiens cross-tissue dataset using four evaluation metrics.
For (C), (G), and (M), metaprograms are ordered by decreasing standard deviation difference. "****", "**", and "*" indicate metaprograms ranked first, second, and third in each cell cluster, respectively. Only metaprograms passing the standard deviation threshold are shown.

CellPress
OPEN ACCESS

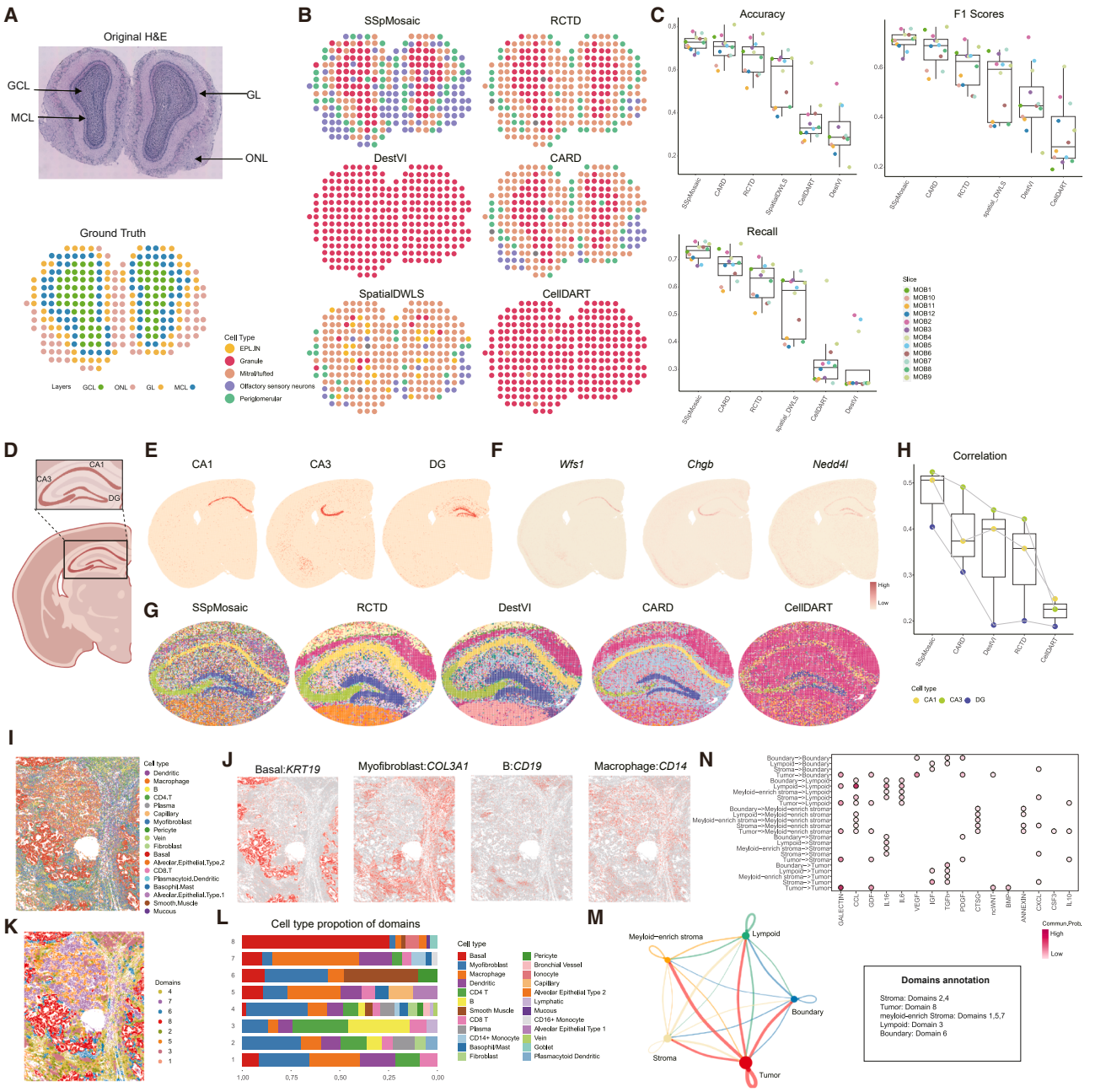


Figure 4. Spatial deconvolution and annotation performance of SSpMosaic across different tissue types

(A) Anatomical structure of MOB layers from inner to outer: GCL, MCL, GL, and ONL.

(B) Main cell-type inference by six deconvolution methods at each spot in MOB10 dataset.

(C) Performance comparison of six deconvolution methods across 12 MOB slices using Accuracy, F1 scores, and Recall metrics.

(D) Schematic illustration of mouse hippocampal organization and its three main cell types.

(E) Metaprogram scores of SSpMosaic for three major hippocampal cell types in Visium HD mouse brain sections.

(F) Expression levels of cell-type-specific markers (CA1: *Wfs1*; CA3: *Chgb*; DG: *Nedd4l*) for three main hippocampal cell types in Visium HD mouse brain data.

(G) Main cell-type inference by six deconvolution methods at each spot in the hippocampal region of Visium HD mouse brain data.

(H) Boxplots showing Pearson correlations between inference scores and corresponding marker expression levels for the three main hippocampal cell types across different methods in Visium HD mouse brain slices.

(I) Cell-type inference by SSpMosaic in CosMx human NSCLC data.

(J) Spatial expression patterns of markers for four major cell types in NSCLC data.

(legend continued on next page)

uncharacterized or undefined cell types. To explore this, we applied SSpMosaic to a cross-tissue dataset that included nearly 200,000 cells.[40] By leveraging reference metaprograms learned from an annotated dataset with diverse cell types, SSpMosaic was able to assign cell-type annotations across this large dataset (Figures 3L and 3M). This approach not only confirmed known cell types but also facilitated the identification of novel or undefined cell clusters (denoted as UN) in the query dataset, based on unique expression patterns of gene programs (Figure 3N). In the query dataset, two clusters (labeled as UN1 and UN2) corresponded to no reference metaprograms (Figure 3M), highlighting the potential of SSpMosaic to reveal previously uncharacterized cell populations, while SingleR and TOSICA failed to identify them (Figure 3N). Further analysis revealed two independent gene programs associated with these two undefined cell clusters (Figure 3O). Enrichment analysis showed that UN1 exhibited high expression of programs associated with platelet-related pathways, and thus, UN1 was denoted as platelets. On the other hand, UN2 displayed gene signatures linked to aerobic respiration, suggesting that UN2 may represent an uncharacterized subtype of bladder urothelial cells (Figures 3P and S2). Quantitative analysis highlighted the exceptional performance of SSpMosaic compared to other methods (Figure 3Q). These findings highlight the ability of SSpMosaic to identify novel cellular populations and provide functional insights into previously unannotated clusters.

SSpMosaic facilitates cell-type deconvolution for spatial transcriptomics across different resolutions

Disentangling the spatial localization of cell types and characterizing complex tissue architecture using spatial transcriptomic data remains a significant challenge, particularly due to the inherent variability in spatial resolution, tissue heterogeneity, and differences in molecular profiles across datasets. SSpMosaic addresses this challenge by using metaprograms instead of reference expression matrices, ensuring adaptable and accurate annotation across different spatial transcriptomics datasets (Figure 1C). To evaluate its performance, we tested SSpMosaic on spatial transcriptomic datasets spanning spot-level (10x Visium and 10x Genomics), single-cell (CosMx and NanoString), and sub-cellular (Visium HD, 10x Genomics) resolution, demonstrating its capability for accurate cell-type annotation across diverse spatial scales (Figure 4A).

We first applied SSpMosaic to mouse olfactory bulb (MOB) spatial transcriptomics datasets (n = 12 slices) at a spot-level resolution of 55 μm,[37] using matched scRNA-seq data from the same tissue[41] for deconvolution. The MOB data are characterized by four distinct morphological layers: granule cell layer (GCL), mitral cell layer (MCL), glomerular layer (GL), and olfactory nerve layer (ONL) (Figure 4A). SSpMosaic demonstrated high accuracy in annotating cell types corresponding to these morphological layers (Figures 4B and S3A), successfully identifying granule cells in the GCL, periglomerular cells in the GL, olfactory sensory neurons in the ONL, and mitral/tufted cells in the MCL.[42] Overall, SSpMosaic demonstrated superior performance compared to state-of-the-art deconvolution methods, including CARD,[43] RCTD,[44] DestVI,[45] SpatialDWLS,[46] and CellDART,[47] achieving higher accuracy, F1 scores, and recall scores across 12 slices (Figure 4C). For instance, while CARD, the second-best method, struggled to clearly define the ONL layer, SSpMosaic accurately captured the olfactory sensory neurons in this region, showcasing its precision and robustness in resolving complex tissue architecture (Figure 4B).

We then evaluated SSpMosaic using a high-resolution dataset—a coronal brain section containing the hippocampus measured using 10x Visium HD at a resolution of 16 μm (Figure 4D). Using a hippocampus scRNA-seq dataset[26] for deconvolution, SSpMosaic accurately delineated key hippocampal regions, including cornu ammonis 1 (CA1), CA3, and dentate gyrus (DG) neurons (Figure 4E), with validation provided by the expression patterns of known marker genes (Figure 4F). In comparison, competing methods such as RCTD and DestVI were unable to clearly capture the main structures of the hippocampus, while CARD detected diffused signals in the CA3 region (Figure 4G). We further assessed the performance of deconvolution methods by calculating Pearson's correlation coefficient between the deconvoluted cell-type proportions and the corresponding marker expression levels within the three major hippocampal regions (CA1, CA3, and DG). This metric provides a quantitative evaluation of how accurately each method captures the spatial distribution of cell types, reflecting its ability to align deconvolution results with established biological markers. SSpMosaic consistently outperformed other methods, achieving the highest correlation coefficients across all regions, demonstrating its superior capability in preserving spatial fidelity and accurately mapping cell-type distributions (Figure 4H).

Furthermore, we evaluated the performance of SSpMosaic at single-cell resolution using Nanostring CosMx data from non-small cell lung cancer (NSCLC). This dataset measured mRNA expression for 960 genes across over 100,000 cells, providing a high-resolution spatial transcriptomic landscape.[48] Applying the same metaprogram-based strategy used for single-cell data annotation, SSpMosaic effectively predicted the cell type of each cell by leveraging an annotated NSCLC scRNA-seq dataset[49] as the reference (Figure 4I). The annotated cell types, including basal cells, fibroblasts, B cells, and macrophages, exhibited high concordance with the expression patterns of known marker genes such as *KRT19*, *COL3A1*, *CD19*, and *CD14*, respectively (Figures 4J and S3B). In addition, SSpMosaic allowed clustering analysis of the cell × metaprogram score matrix, grouping NSCLC sub-cellular regions into eight distinct spatial domains (Figure 4K). Each domain represented a region with a coherent and unique cell-type composition, reflecting the underlying heterogeneity and functional organization of the tumor

(K) Spatial domains identified through clustering based on SSpMosaic-inferred cell types in NSCLC data.
(L) Stacked bar plot showing the proportion of different cell types within each spatial domain in NSCLC data.
(M) Right: spatial domain annotations; left: intercellular communication networks between different spatial regions in NSCLC data. Line thickness represents communication strength.
(N) Bubble plot showing signaling pathway communication probabilities between different spatial regions.

microenvironment (Figure 4L). For example, domain 3 was enriched in lymphoid cells, including $CD4^+$ and $CD8^+$ T cells as well as B cells, indicating an immune-active region. In contrast, domain 8, representing the tumor core, was composed predominantly of tumor cells with minimal immune cell infiltration, highlighting its immune-suppressive nature (Figure 4L). Notably, the boundary regions between the tumor core and surrounding tissues (domain 6) were abundant in smooth muscle cells and pericytes, suggesting a vasculature-rich zone that may play a critical role in tumor angiogenesis and tissue remodeling (Figure 4L). Communication analysis among spatial domains (Figures 4M and 4N) identified heightened GALECTIN pathway activity within the tumor core (domain 8), suggesting its role in promoting immune evasion and tumor progression.[50] At the tumor boundary (domain 6), vascular endothelial growth factor (VEGF) signaling was markedly elevated, indicating active angiogenesis and vascular remodeling,[51] which aligns with the spatial arrangement of vascular structures, including pericytes and smooth muscle cells, observed in this region. These findings highlight key signaling pathways underlying the functional specialization of different spatial domains within the NSCLC tumor microenvironment.

In summary, SSpMosaic provides both deconvolution-based and metaprogram-based approaches, offering a versatile framework for the analysis of spatial transcriptomics data across diverse resolutions, from spot-level to sub-cellular granularity.

SSpMosaic elucidates spatial domain dynamics and spatially resolved gene programs

To evaluate the capability of SSpMosaic in uncovering spatial domain dynamics and spatially resolved gene programs, we applied it to a comprehensive single-cell and spatial multi-omic map of myocardial infarction.[52] This dataset included single-nucleus RNA sequencing (snRNA-seq) and transposase-accessible chromatin sequencing (snATAC-seq), and spatial transcriptomics data were generated from tissue samples of the ischemic zone (IZ), fibrotic zone (FZ), border zone (BZ), and remote zone (RZ) from patients with acute myocardial infarction, as well as non-transplanted donor hearts as controls.

SSpMosaic harmonized the snRNA-seq and snATAC-seq data by aligning them into shared metaprograms, facilitating the seamless integration of transcriptional and chromatin accessibility information (Figure 5A). Comparison of gene programs revealed consistent patterns of nuclear transcription and chromatin accessibility across cell types, exemplified by well-established marker genes with known cell-type specificity (Figure 5B), which confirmed the alignment of transcriptional and epigenomic landscapes achieved by SSpMosaic. The metaprograms derived from snRNA-seq and snATAC-seq data were subsequently applied to deconvolute the spatial transcriptomics data (Figure 5A). Of note, the cell-type annotation by SSpMosaic demonstrated high consistency with findings from the original study (Figure 5C), which underscores the robustness and reliability of SSpMosaic's metaprogram-based integration approach. For instance, the control, RZ, and BZ zones showed similar patterns dominated by cardiomyocyte cells, while IZ and FZ zones exhibited enrichment of myeloid and fibroblast cells, respectively (Figure 5C).[52,53] The IZ was particularly characterized by a significant reduction in the number of cardiomyocyte cells and an increased proportion of myeloid cells, which is indicative of an acute inflammatory response. Conversely, the FZ was marked by a higher presence of fibroblasts and a partial restoration of cardiomyocyte cells[52] (Figure 5D).

UMAP visualization, based on metaprogram contributions, revealed that SSpMosaic grouped spatial spots with similar cell-type composition into eleven distinct spatial domains (Figure 5E), which represent potential functional units within the tissue that are conserved across different slices, thereby facilitating cross-sample comparisons.[54] Notably, distinct sample groups exhibited enrichment for specific spatial domains (Figure S4A). For instance, domains 1, 3, 8, and 7 were enriched with cardiomyocyte cells; domains 2 and 5 with fibroblast cells; domains 4 and 10 with myeloid and lymphoid cells; and domains 9, 6, and 11 with vascular smooth muscle cells (vSMCs), mast cells, and adipocytes, respectively—all of which are rare cell types in the myocardium. Hierarchical clustering of spatial domain compositions further supported this observation by revealing four major groups: myogenic (domains 1, 3, 7, and 8), fibrotic (2 and 5), inflammatory (4 and 10), and rare cell groups (6, 9, and 11) (Figure 5F). This grouping not only aligns closely with findings from the original study[52] but also provides enhanced resolution and granularity in defining domain-specific compositions and gene program activities (Figure 5G), underscoring SSpMosaic's ability to capture nuanced spatial domain dynamics and their functional implications across different myocardial regions. Consistent with the spatial domain annotations, these programs included five groups: the mural group (enriched in domain 9, involved in smooth muscle contraction, with *MYH11* as a hallmark gene)[55]; the myocardial group (highly expressed in muscle-derived domains, enriched in muscle differentiation pathways, with *ACTN2* as a key gene)[56]; the fibro group (significantly expressed in fibrotic domains, enriched in fibrosis-related genes like *AEBP1*)[57]; the inflammatory group (focused on inflammatory domains, enriched in complement and macrophage activation pathways, containing important immune genes like *C1R* and *C1S*)[58]; and the MI-Fibro group (specifically highly expressed in ischemic domains, centered on extracellular-matrix-related pathways)[59] (Figures 5F and S4). These results demonstrate the ability of SSpMosaic to uncover functionally distinct spatial domain activities within complex tissues.

To quantify the spatial distribution patterns of these domains across different sample groups, we employed Moran's I, a measure of spatial autocorrelation. The analysis revealed that control, remote, and border zone slices exhibited Moran's I values near zero, indicating a random distribution of domains (Figures 5H and S4). In contrast, domains 2, 4, 9, and 5 showed higher Moran's I value, suggesting a more structured distribution. Specifically, domains 2, 4, and 5, which are linked to fibrosis and inflammation, demonstrate a concentrated spatial distribution, particularly evident in the ischemic and fibrotic groups. Additionally, domain 9, enriched with vascular wall cells, also displayed a high Moran's I value, indicating a concentrated distribution of the vascular system (Figures 5H and S4A). Further spatial communication analysis revealed the molecular mechanisms underlying the observed spatial distribution differences (Figure 5I). Interestingly, we observed that the communication intensity was reduced in the FZ, while the communication patterns

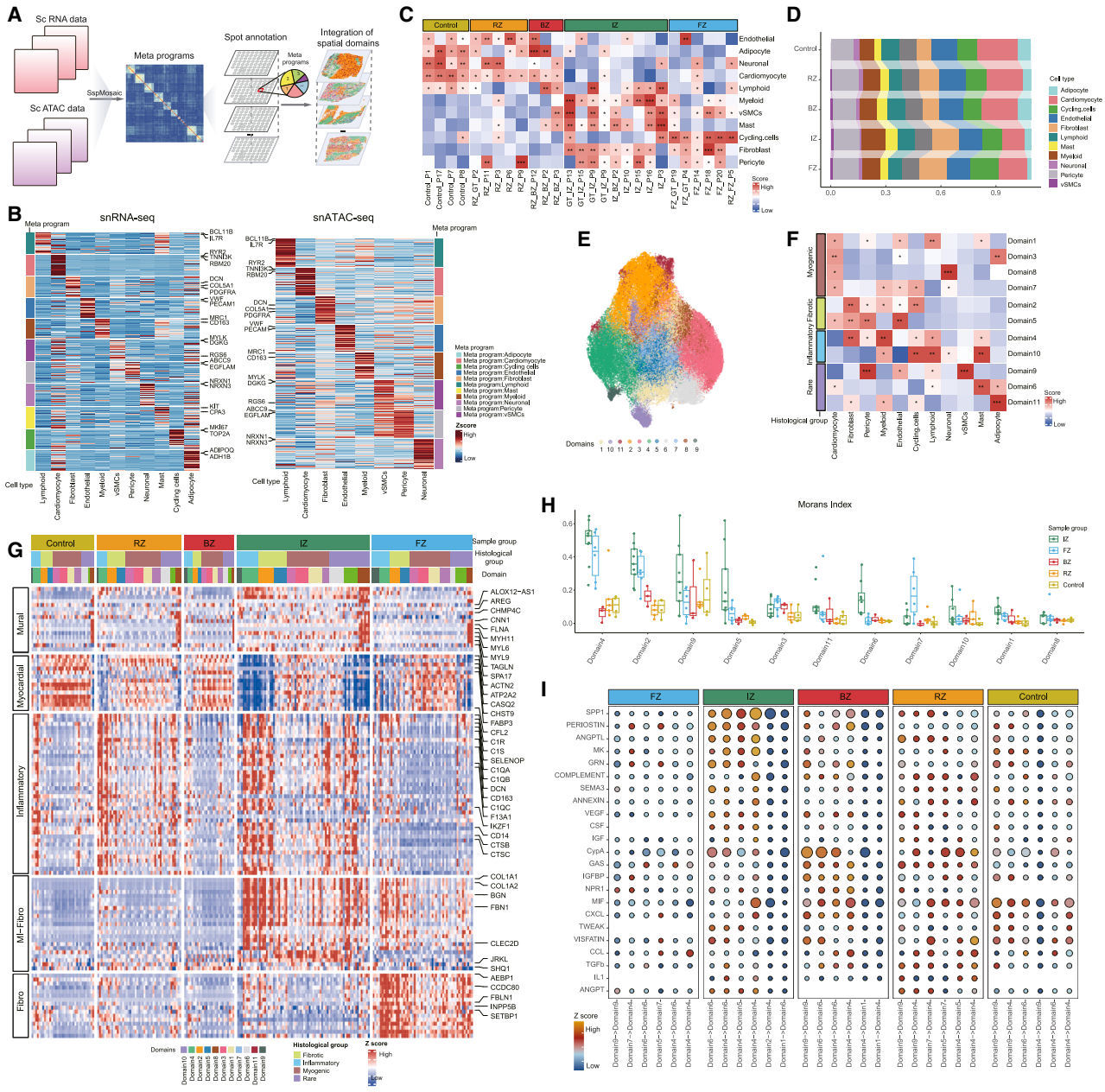


Figure 5. Myocardial infarction spatial domain analysis

(A) Integrated workflow for scRNA-seq, scATAC-seq, and spatial transcriptomics.

(B) The heatmaps display the expression of high-weight genes from SSpMosaic metaprograms across the cell-type compositions in scRNA and scATAC data.

(C) Heatmap showcases the cell-type scores computed by SSpMosaic's deconvolution across each slice. The scores represent the z-scaled average values for all spots within a sample (*: $0 < Z$ score <1; **: $1 \leq Z$ score <2; ***: Z score ≥ 2).

(D) Stacked bar plot illustrating the proportion of different cell types across sample groups.

(E) UMAP visualization showing spots clustered into 11 distinct niches based on cell type.

(F) Heatmap of the average z-scaled cell type scores across spatial domains, histological group clustering by domains (*: $0 < Z$ score <1; **: $1 \leq Z$ score <2; ***: Z score ≥ 2).

(G) Heatmap showing the expression patterns of high-weight genes within domain-specific gene programs from different slices. The top-weighted gene markers are indicated on the right.

(H) Boxplot visualizing the Moran's I index for spatial domains across different sample groups.

(I) Bubble chart depicting communication probabilities between different spatial domains. Only the six communication types that show the most significant differences between sample groups are displayed (ANOVA p value <0.01).

in the IZ, BZ, control, and RZ were similar (Figure 5I). Notably, the SPP1 pathway exhibited the highest activity in the IZ,[52] particularly significant in communications from domain 4 to domains 4 and 5. Given that domain 4 was enriched in myeloid cells and domains 4 and 5 were both enriched in fibroblast cells, this finding indicates that during the tissue remodeling process following myocardial infarction, *SPP1* plays a crucial role in facilitating the interaction between myeloid and fibroblast cells.[23]

In summary, SSpMosaic provides a methodological framework dissecting the functional organization and molecular activity within complex tissues, enabling a unified understanding of tissue architecture and cellular interactions.

Extend SSpMosaic for reference-free spatial organization characterization

To broaden its applicability, we extended SSpMosaic to enable the integration of spatial transcriptomics and the characterization of spatial patterns without relying on a reference single-cell dataset. This extension leverages SSpMosaic's ability to infer gene programs directly from spatial transcriptomics data. These gene programs can then be mapped to cell types or cell states by evaluating their proximity to known marker genes or annotated through gene set enrichment analysis (GSEA).

As a proof of concept, we applied SSpMosaic to a glioblastoma (GBM) dataset comprising 26 spatial transcriptomics slices generated using the 10x Visium platform.[60] Gene programs were independently inferred for each slice (Figure 6A), followed by the identification of 17 metaprograms, each of which typically encompassed programs derived from multiple GBM samples (Figures 6B and S5A). These metaprograms were further annotated by overlapping their gene sets with those defined in the original study[60] or through enriched biological functions identified via GSEA (Figures 6C and 6D). SSpMosaic successfully recovered all the metaprograms defined by non-negative matrix factorization (NMF) in the original study[60] (Figure 6D). The identification was supported by recent independent studies, including malignant metaprograms such as oligodendrocyte progenitor-like malignant cells (OPCs), astrocyte-like malignant cells (ACs), mesenchymal cells (MESs), and neural progenitor cells (NPCs),[13,61] as well as non-malignant metaprograms such as hypoxia-related MES (MES.Hyp)[62] (Figure 6E). In addition, three additional metaprograms were identified, including gene programs associated with unfolded protein response (UPR),[63] tumor-associated macrophages (TAMs), and ion channel programs (ICs)[63] (Figures 6C and 6E).

In multicellular systems, such as the glioblastoma microenvironment, cellular function and tissue homeostasis rely on spatial interactions between neighboring cell types. Spatially colocalized metaprograms may reflect interactions between cell types within specific spatial contexts, coordinating tissue functions. To explore this, we performed pairwise colocalization analysis of weight average scores of metaprograms (Figure 6F). Significant colocalization was observed among the metaprograms of TAM, UPR, and MES.Hyp, highlighting hypoxia as a critical microenvironmental feature in GBM development. This finding underscores hypoxia's role in driving tumor cell proliferation, invasion, angiogenesis, and resistance to treatment.[64] We found that UPR high-score spots showed random distribution (Figure S5B), and TAM high-score spots were frequently encapsulated by MES hypoxic regions, forming pseudopalisade-like structures.[65]To further quantify the degree of neighboring cell-type mixture within spatial domains, we developed an "encapsulation index" (Figure 6G). High encapsulation index values indicate a greater degree of intermixing and spatial interaction, suggesting potential cell-cell communication or shared functional roles within the tissue microenvironment. For the aforementioned colocalized metaprograms of TAM, UPR, and MES.Hyp, using the MES.Hyp metaprogram as the proxy, the encapsulation index for TAM was significantly higher than that for UPR (Figure S5C). This indicates that TAM spots were frequently encapsulated by MES hypoxic regions, indicative of the formation of pseudopalisade-like structures.[65] However, the encapsulation index for TAM by MES.Hyp varied across slices, leading to the classification of the slices into two groups: encapsulated TAM-hypoxia (E-TAM-hyp) and intermixed TAM-hypoxia (I-TAM-hyp) (Figure 6H), as exemplified in Figure 6I. While both E-TAM-hyp and I-TAM-hyp slices exhibited TAM, UPR, and MES.Hyp colocalization, E-TAM-hyp slices showed distinct features in hypoxic-TAM regions. Specifically, E-TAM-hyp slices displayed elevated VISFATIN, ANGPTL, SPP1, and MIF signaling intensities compared to I-TAM-hyp slices (Figure 6I). These elevated signals suggest enhanced M2 macrophage polarization and recruitment in E-TAM-hyp slices.[66–69] Additionally, the increased VISFATIN and MIF, both upstream regulators of interleukin-1β (IL-1β) synthesis, further support their role in pseudopalisade formation.[49,65,70]

To further validate the robustness of SSpMosaic in reference-free settings with limited input, especially in practically relevant small-sample scenarios, we analyzed a stricturing Crohn disease spatial transcriptomics dataset.[71] From the original collection of 20 tissue slices generated on the 10x Genomics Visium platform across 10 patients, we selected five representative slices (three adjacent-to-stricture normal and two fibrotic) to mimic small-sample conditions (Figure S6A). Using the same pipeline as in the glioblastoma analysis, we inferred gene programs from each slice and integrated them into 15 metaprograms (Figure S6B), which were annotated to 13 cell types (Figures S7–S11) through GSEA and high-weight gene validation (Figures S6C and S6D). These metaprograms not only recapitulated all seven major cell types reported in the original study (epithelial cells, fibroblasts, myeloid cells, B cells, T cells, pericytes, and glia) but also uncovered additional biologically plausible populations that were absent from the paired single-cell reference, including neurons (marked by *SNAP25*[72] and *GAP43*[73]) and two distinct myocyte subsets with contractile versus metabolic signatures (Figures S6B–S6D). Spatial distributions of metaprogram scores aligned with histological compartments—for example, stromal cells in the lamina propria and myocytes in the muscle layers(Figure S6E). For quantitative evaluation, we applied pointwise mutual information (STAR Methods) analysis, which confirmed that each metaprogram showed the strongest association with markers of its assigned cell type across all slices (Figure S6F). Notably, fibrotic slices exhibited significantly higher T cell-B cell colocalization than adjacent tissues[74] (Figures S6G and S6H), consistent with known expansion of tertiary lymphoid structures in fibrotic Crohn

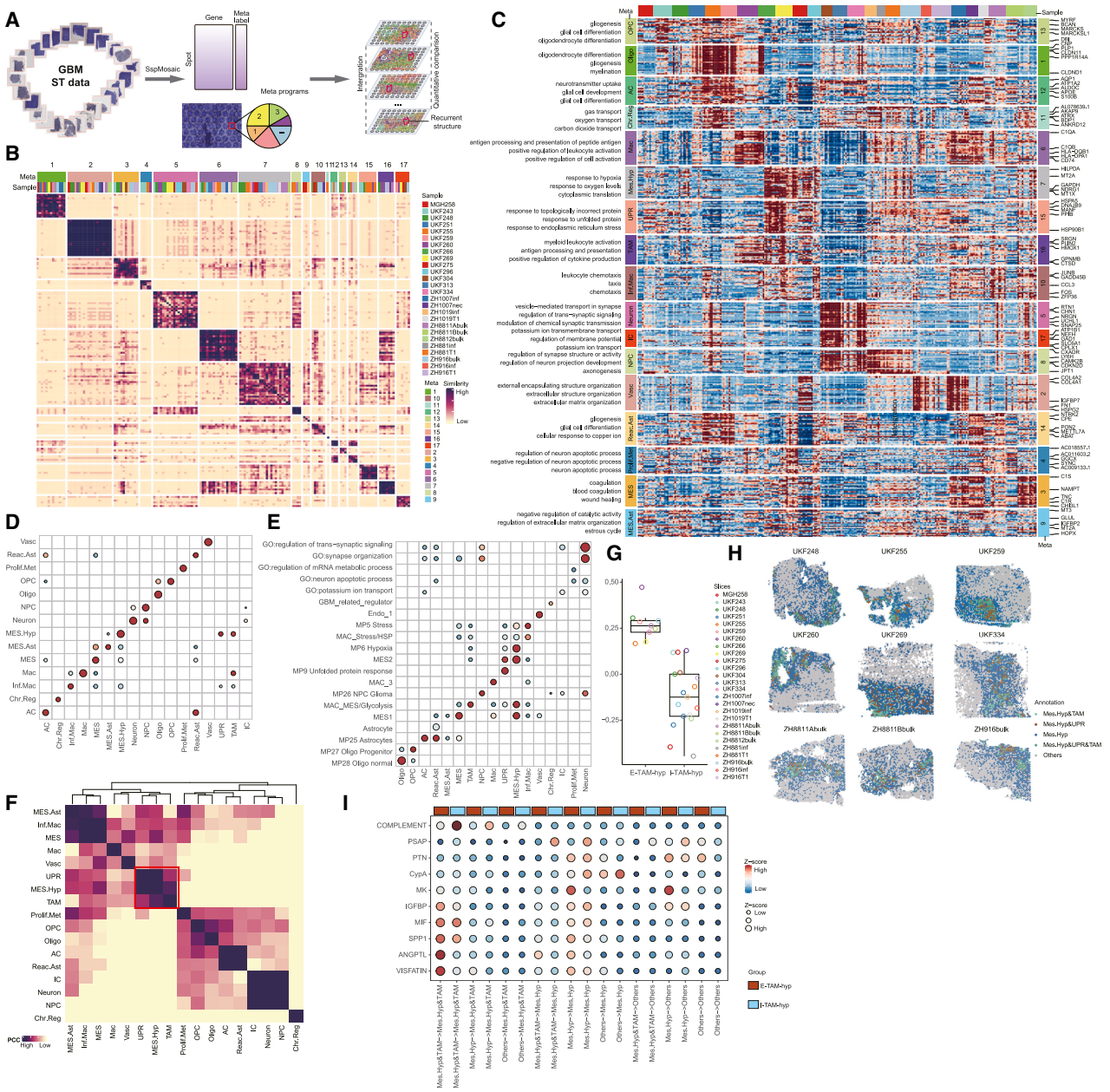


Figure 6. Integration of multi-slice GBM spatial transcriptomics and identification of recurrent spatial structures

(A) Schematic workflow for integrating multi-slice GBM spatial transcriptomics data and identifying recurrent spatial structures.

(B) Heatmap showing similarity of metaprograms identified by the SSpMosaic algorithm across 26 GBM tissue slices.

(C) Heatmap depicting expression levels of high-weight genes for each metaprogram. Right: top five weighted genes per metaprogram. Left: significantly enriched pathways for each metaprogram.

(D) Dot plot illustrating enrichment of SSpMosaic metaprograms (columns) in 14 metaprograms (rows) defined by spatial transcriptomics in the original study.

(E) Dot plot showing enrichment of SSpMosaic metaprograms (columns) in gene sets associated with programs and cell types defined by previous single-cell studies (rows).

(F) Heatmap of Pearson correlation coefficients between metaprogram scores.

(G) Box and scatterplots depict the TAM spot encapsulation index within MES.Hyp regions, categorizing slices into E-TAM-hyp and I-TAM-hyp groups. The nine highest encapsulation index slices (E-TAM-hyp group) are shown in (H).

(I) Dot plot showing inter-region communication probabilities of ligand-receptor pairs between E-TAM-hyp and I-TAM-hyp groups. "&" indicates co-localization between metaprograms.

disease.[75] Together with the GBM analysis, these results establish that SSpMosaic can robustly characterize spatial organization without reliance on single-cell references, delivering interpretable, spatially coherent, and biologically meaningful insights across both large-scale and limited-sample contexts.

DISCUSSION

Single-cell and spatial multi-omics data integration aims to combine different layers of molecular information (e.g., transcriptomics, epigenomics, proteomics, and metabolomics) from the same or similar cells to gain a comprehensive understanding of cellular states and regulatory mechanisms within their native tissue context. However, each omics modality captures distinct biological features, often exhibiting differences in data sparsity, noise levels, batch effects, and spatial resolution. These challenges hinder direct integration and cross-modality alignment, necessitating the development of sophisticated computational frameworks that preserve biological signal while harmonizing diverse datasets.

Genes serve as the fundamental molecular blueprint of cellular function, with their expression patterns determining cell identity, state, and interactions within complex tissues. Cellular characteristics, such as identity and function, are rarely dictated by individual genes alone; instead, they arise from the coordinated activity of multiple functionally related genes. These gene sets, often referred to as gene programs, reflect the intricate molecular mechanisms that govern cellular processes. Gene programs from different omics layers (e.g., transcriptomics, epigenomics, and proteomics) can be integrated by embedding them into a shared latent space, facilitating cross-modality alignment and consistent interpretation of molecular data across diverse datasets and experimental conditions. Building on this concept, we develop SSpMosaic, a computational framework that leverages interpretable metaprograms—higher-order representations of gene programs learned across datasets—to achieve robust data integration and precise cell-type annotation. Crucially, SSpMosaic maintains stability across key analytical parameters: varying clustering resolutions (major or minor clusters) and differential expression thresholds (high or low logFC) consistently yielded specific programs that captured intra-cluster heterogeneity (Figures S12–S18). We evaluate its performance under diverse scenarios, demonstrating its ability to preserve critical biological signals while effectively mitigating technical noise. This framework facilitates various downstream tasks by harmonizing multi-omics data at both single-cell and spatial levels, enabling comprehensive analysis of cellular heterogeneity and tissue organization.

The success of SSpMosaic in the diverse applications is largely due to its ability to learn and apply biologically interpretable gene programs that capture functional gene programs associated with cellular identities. This method avoids the limitations of relying solely on individual marker genes, which may not always be present or detectable in all datasets. Notably, when benchmarked against specialized program-learning methods (Spectra,[7] expiMap,[8] and cNMF[76]), SSpMosaic demonstrated superior biological fidelity in two critical dimensions. For cell-type specificity, SSpMosaic can distinguish closely related immune subtypes, such as $CD4^+$ Tconv from $CD8^+$ T cells, while preserving shared cytotoxic programs with NK cells. In terms of heterogeneity capture, it comprehensively maps continuum states within homogeneous populations. For example, it identified four complementary plasma cell programs covering the entire plasma cell population. By contrast, Spectra conflated T cell subtypes, expiMap showed limited substate resolution due to reliance on predefined gene sets, and cNMF exhibited reduced sensitivity to transitional biology (Figure S19). Notably, SSpMosaic effectively resolves continuum biology beyond discrete clusters as demonstrated in human hematopoietic development.[77] When applied to hematopoietic differentiation time courses (D0–D6), metaprogram activity scores precisely aligned with pseudotime trajectories: in the mesenchymal lineage, sequential activation progressed from hPSC → MPC → D4_Mes → D6_Mes; in the endothelial-hematopoietic lineage, activation followed hPSC → MPC → D4_EC → HPC → D6_EC (Figure S20). Most impressively, it identified rare transitional states (398-cell HPC population) through stage-specific programs containing definitive markers like *SPN* (CD43), while capturing transient regulators (*RUNX1* in D4_EC) that define temporal maturation windows (Figure S20). By focusing on co-expressed gene sets, SSpMosaic provides a more comprehensive and robust approach to cell-type annotation, which can be critical for accurately classifying cells in heterogeneous tissues, such as tumors or complex organs. Moreover, the ability to transfer metaprograms across single-cell and spatial omics datasets enables cell-type deconvolution, spatial domain detection, and identification of spatially resolved gene programs to reveal tissue organization and cellular interactions, increasing the scalability and versatility of this approach.

Overall, SSpMosaic introduces several key innovations that distinguish it from existing methods. First, its network propagation algorithm enables robust evaluation of program similarities both within and across datasets, resulting in more accurate metaprogram construction. For cross-batch integration, the default neighbor parameter (metaprogram neighbor = 1) restricts search to identical metaprogram labels, prioritizing biological specificity by preventing implausible connections between fundamentally distinct cell types (e.g., T cells to neurons). This design was empirically validated in colorectal cancer data, where metaprogram neighbor = 1 achieved higher integration metrics (avg score = 0.87, ARI = 0.91, and cell-type ASW = 0.79) compared to higher values (metaprogram neighbor = 2/3; Figure S21). While increased metaprogram neighbor could capture relationships between related types (e.g., NK cell-CD8Tex similarity, which reflects that both are cytotoxic cell types), it compromised cell-type purity (Figure S21). Second, the implementation of a unified framework for both single-cell and spatial omics data analysis eliminates the need for complex tool integration workflows, streamlining data processing, enhancing reproducibility, improving computational efficiency, and identifying potential functional units of cellular interaction and tissue architecture, as highlighted in the multi-omic analysis of myocardial infarction. Third, the adaptive nature of gene programs allows SSpMosaic to handle spatial data at various resolutions without sacrificing accuracy, enabling seamless integration of spatial omics data across different scales. Last but not least, by leveraging metaprograms to assess spatial feature similarities, the framework enables cross-sample comparisons and

annotation of tissue architectures from spatial transcriptomics data, even in the absence of single-cell reference data. In the 26 glioblastoma spatial transcriptomics datasets, SSpMosaic identified spatially colocalized gene programs associated with macrophage polarization and tumor progression.

In conclusion, SSpMosaic represents a powerful and versatile framework for the integration and interpretation of single-cell and spatial omics data. By leveraging interpretable gene programs and metaprograms, it provides a more comprehensive understanding of cellular identities, functions, and interactions within complex tissue environments. As single-cell and spatial omics technologies continue to advance, the potential applications of SSpMosaic will likely expand, contributing to our understanding of cellular heterogeneity, tissue dynamics, and the molecular underpinnings of disease progression in health and disease.

LIMITATIONS

While SSpMosaic offers several advantages, there are still challenges that warrant further investigation. One such challenge is the assumption that gene programs are conserved across different conditions or disease states, which may not always hold true due to the dynamic nature of gene expression in response to environmental, developmental, or pathological factors. Additionally, the generalizability of metaprogram transfer across diverse datasets requires further validation across a broader range of biological systems and experimental conditions. Finally, modeling higher-order gene interactions and regulatory networks across multi-omics data remains a complex task, particularly in the context of rare cell types or dynamic biological processes. Addressing these challenges will be crucial for enhancing the scalability, precision, and applicability of SSpMosaic in future studies.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the lead contact, Dijun Chen (dijunchen@nju.edu.cn).

Materials availability

This study did not develop any new or unique reagents.

Data and code availability

The publicly available data utilized in this study can be accessed through various sources. For multi-batch and multi-modal data, CRC data are available at TISCH, while mouse brain scRNA data can be found in ArrayExpress: E-MTAB-11115. Human brain scRNA datasets are accessible via the Allen Brain Atlas: https://portal.brain-map.org/atlases-and-data/rnaseq, GEO: GSE229169, and the Cellxgene Collection: https://cellxgene.cziscience.com/collections/ceb895f4-ff9f-403a-b7c3-187a9657ac2c. Mouse atlas scATAC data can be referenced in the Mouse ATAC Atlas: https://atlas.gs.washington.edu/mouse-atac/, and mouse atlas scRNA data are available in Tabula Muris: https://tabula-muris.sf.czbiohub.org/data. PBMC data are accessible through Github: https://github.com/PeterZZQ/scMoMaT/tree/main/data/real/ASAP-PBMC. In terms of single-cell annotation data, CHOL data are also available at TISCH, the human neural cell subtype dataset is in the Cellxgene Collection: https://cellxgene.cziscience.com/collections/cae8bad0-39e9-4771-85a7-822b0e06de9f, and the cross-human tissue atlas dataset can be Cellxgene Collection: https://cellxgene.cziscience.com/collections/e5f58829-1a66-40b5-a624-9046778e74f5. Regarding spatial annotation data, the mouse olfactory bulb dataset is available on CNGB FTP: https://ftp.cngb.org/pub/SciRAID/stomics/STDS0000017/stomics/, with the corresponding scRNA reference found at GEO: GSE121891. The NSCLC dataset is provided by Nanostring-nsclc: https://nanostring.com/products/cosmx-spatial-molecular-imager/ffpe-dataset/nsclc-ffpe-dataset/, with its scRNA reference data accessible through Synapse: https://www.synapse.org/Synapse:syn21041850/files/, while the mouse brain dataset can be accessed via 10X Genomics: https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-libraries-of-mouse-brain-he. Additionally, myocardial infarction multi-omic data are available at Zenodo: https://zenodo.org/records/6578047, and the glioblastoma dataset can be found at Github: https://github.com/tiroshlab/Spatial_Glioma. The stricturing Crohn disease data are also available at Zenodo: https://zenodo.org/records/14509802, and the human hematopoietic development dataset can be found at GEO: GSE145859. The original code for SSpMosaic is publicly available on GitHub: https://github.com/compbioNJU/SSpMosaic and on Zenodo: https://doi.org/10.5281/zenodo.17549304.

ACKNOWLEDGMENTS

We would like to acknowledge the Center for Information Technology and the High Performance Computing Center of Nanjing University for providing high-performance computing resources. This research was not funded by any formal grant or external funding agency but was made possible through internal institutional support and collaboration.

AUTHOR CONTRIBUTIONS

D.C. conceived the project. D.C., Y.N., and R.D. contributed to the resources. D.C. supervised the project. D.C. and R.D. acquired the funding. Y.Z. and W.M. developed the algorithms and implemented the software. Y.Z. and W.M. performed the formal analysis with contribution from B.Y., L.W., K.L., and L.X. D.C., Y.Z., and W.M. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Overview of SSpMosaic
  - Unifying feature space
  - Constructing differential co-expression gene programs
  - Cell cluster-specific program selection
  - Calculating program distances using network propagation
  - Data integration
  - Cell type annotation
  - Annotation of spatial transcriptomics data
  - Encapsulation index
  - The overall tendency M_l for clustering is measured as
  - Pointwise mutual information calculation
  - Colocalization score calculation
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2025.101105.

Received: April 16, 2025
Revised: July 31, 2025
Accepted: November 21, 2025
Published: December 19, 2025

REFERENCES

1. Rao, A., Barkley, D., França, G.S., and Yanai, I. (2021). Exploring tissue architecture using spatial transcriptomics. Nature *596*, 211–220. https://doi.org/10.1038/s41586-021-03634-9.
2. Baysoy, A., Bai, Z., Satija, R., and Fan, R. (2023). The technological landscape and applications of single-cell multi-omics. Nat. Rev. Mol. Cell Biol. *24*, 695–713. https://doi.org/10.1038/s41580-023-00615-w.
3. Van de Sande, B., Lee, J.S., Mutasa-Gottgens, E., Naughton, B., Bacon, W., Manning, J., Wang, Y., Pollard, J., Mendez, M., Hill, J., et al. (2023). Applications of single-cell RNA sequencing in drug discovery and development. Nat. Rev. Drug Discov. *22*, 496–520. https://doi.org/10.1038/s41573-023-00688-4.
4. Russo, M., Chen, M., Mariella, E., Peng, H., Rehman, S.K., Sancho, E., Sogari, A., Toh, T.S., Balaban, N.Q., Batlle, E., et al. (2024). Cancer drug-tolerant persister cells: from biological questions to clinical opportunities. Nat. Rev. Cancer *24*, 694–717. https://doi.org/10.1038/s41568-024-00737-z.
5. Zhang, Y., Yu, B., Ming, W., Zhou, X., Wang, J., and Chen, D. (2024). SpaTopic: A statistical learning framework for exploring tumor spatial architecture from spatially resolved transcriptomic data. Sci. Adv. *10*, eadp4942. https://doi.org/10.1126/sciadv.adp4942.
6. Moses, L., and Pachter, L. (2022). Museum of spatial transcriptomics. Nat. Methods *19*, 534–546. https://doi.org/10.1038/s41592-022-01409-2.
7. Kunes, R.Z., Walle, T., Land, M., Nawy, T., and Pe'er, D. (2024). Supervised discovery of interpretable gene programs from single-cell data. Nat. Biotechnol. *42*, 1084–1095. https://doi.org/10.1038/s41587-023-01940-3.
8. Lotfollahi, M., Rybakov, S., Hrovatin, K., Hediyeh-Zadeh, S., Talavera-López, C., Misharin, A.V., and Theis, F.J. (2023). Biologically informed deep learning to query gene programs in single-cell atlases. Nat. Cell Biol. *25*, 337–350. https://doi.org/10.1038/s41556-022-01072-x.
9. Wouters, J., Kalender-Atak, Z., Minnoye, L., Spanier, K.I., De Waegeneer, M., Bravo González-Blas, C., Mauduit, D., Davie, K., Hulselmans, G., Najem, A., et al. (2020). Robust gene expression programs underlie recurrent cell states and phenotype switching in melanoma. Nat. Cell Biol. *22*, 986–998. https://doi.org/10.1038/s41556-020-0547-3.
10. Morabito, S., Reese, F., Rahimzadeh, N., Miyoshi, E., and Swarup, V. (2023). hdWGCNA identifies co-expression networks in high-dimensional transcriptomics data. Cell Rep. Methods *3*, 100498. https://doi.org/10.1016/j.crmeth.2023.100498.
11. Li, J., Wang, J., Zhang, P., Wang, R., Mei, Y., Sun, Z., Fei, L., Jiang, M., Ma, L., E, W., et al. (2022). Deep learning of cross-species single-cell landscapes identifies conserved regulatory programs underlying cell types. Nat. Genet. *54*, 1711–1720. https://doi.org/10.1038/s41588-022-01197-7.
12. Li, H., Courtois, E.T., Sengupta, D., Tan, Y., Chen, K.H., Goh, J.J.L., Kong, S.L., Chua, C., Hon, L.K., Tan, W.S., et al. (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat. Genet. *49*, 708–718. https://doi.org/10.1038/ng.3818.
13. Gavish, A., Tyler, M., Greenwald, A.C., Hoefflin, R., Simkin, D., Tschernichovsky, R., Galili Darnell, N., Somech, E., Barbolin, C., Antman, T., et al. (2023). Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. Nature *618*, 598–606. https://doi.org/10.1038/s41586-023-06130-4.
14. Gulati, G.S., D'Silva, J.P., Liu, Y., Wang, L., and Newman, A.M. (2024). Profiling cell identity and tissue architecture with single-cell and spatial transcriptomics. Nat. Rev. Mol. Cell Biol. *26*, 11–31. https://doi.org/10.1038/s41580-024-00768-2.
15. Luecken, M.D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M.F., Strobl, D.C., Zappia, L., Dugas, M., Colomé-Tatché, M., and Theis, F.J. (2022). Benchmarking atlas-level data integration in single-cell genomics. Nat. Methods *19*, 41–50. https://doi.org/10.1038/s41592-021-01336-8.
16. Haghverdi, L., Lun, A.T.L., Morgan, M.D., and Marioni, J.C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat. Biotechnol. *36*, 421–427. https://doi.org/10.1038/nbt.4091.
17. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.R., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296. https://doi.org/10.1038/s41592-019-0619-0.
18. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.
19. Hie, B., Bryson, B., and Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat. Biotechnol. *37*, 685–691. https://doi.org/10.1038/s41587-019-0113-3.
20. Lotfollahi, M., Wolf, F.A., and Theis, F.J. (2019). scGen predicts single-cell perturbation responses. Nat. Methods *16*, 715–721. https://doi.org/10.1038/s41592-019-0494-8.
21. Polanski, K., Young, M.D., Miao, Z., Meyer, K.B., Teichmann, S.A., and Park, J.E. (2020). BBKNN: fast batch alignment of single cell transcriptomes. Bioinformatics *36*, 964–965. https://doi.org/10.1093/bioinformatics/btz625.
22. He, Z., Hu, S., Chen, Y., An, S., Zhou, J., Liu, R., Shi, J., Wang, J., Dong, G., Shi, J., et al. (2024). Mosaic integration and knowledge transfer of single-cell multimodal data with MIDAS. Nat. Biotechnol. *42*, 1594–1605. https://doi.org/10.1038/s41587-023-02040-y.
23. Zhang, L., Li, Z., Skrzypczynska, K.M., Fang, Q., Zhang, W., O'Brien, S.A., He, Y., Wang, L., Zhang, Q., Kim, A., et al. (2020). Single-Cell Analyses Inform Mechanisms of Myeloid-Targeted Therapies in Colon Cancer. Cell *181*, 442–459.e29. https://doi.org/10.1016/j.cell.2020.03.048.
24. Wu, S., Xue, S., Tang, Y., Zhao, W., Zheng, M., Cheng, Z., Hu, X., Sun, J., and Ren, J. (2024). Mitogen-activated protein kinase kinase kinase 1 facilitates the temozolomide resistance and migration of GBM via the MEK/ERK signalling. J. Cell Mol. Med. *28*, e70173. https://doi.org/10.1111/jcmm.70173.
25. Wu, T.D., Madireddi, S., de Almeida, P.E., Banchereau, R., Chen, Y.J.J., Chitre, A.S., Chiang, E.Y., Iftikhar, H., O'Gorman, W.E., Au-Yeung, A., et al. (2020). Peripheral T cell expansion predicts tumour infiltration and clinical response. Nature *579*, 274–278. https://doi.org/10.1038/s41586-020-2056-8.
26. Kleshchevnikov, V., Shmatko, A., Dann, E., Aivazidis, A., King, H.W., Li, T., Elmentaite, R., Lomakin, A., Kedlian, V., Gayoso, A., et al. (2022). Cell2location maps fine-grained cell types in spatial transcriptomics. Nat. Biotechnol. *40*, 661–671. https://doi.org/10.1038/s41587-021-01139-4.
27. Zemke, N.R., Armand, E.J., Wang, W., Lee, S., Zhou, J., Li, Y.E., Liu, H., Tian, W., Nery, J.R., Castanon, R.G., et al. (2023). Conserved and divergent gene regulatory programs of the mammalian neocortex. Nature *624*, 390–402. https://doi.org/10.1038/s41586-023-06819-6.
28. Zhu, K., Bendl, J., Rahman, S., Vicari, J.M., Coleman, C., Clarence, T., Latouche, O., Tsankova, N.M., Li, A., Brennand, K.J., et al. (2023). Multi-omic profiling of the developing human cerebral cortex at the single-cell level. Sci. Adv. *9*, eadg3754. https://doi.org/10.1126/sciadv.adg3754.
29. Yao, Z., van Velthoven, C.T.J., Kunst, M., Zhang, M., McMillen, D., Lee, C., Jung, W., Goldy, J., Abdelhak, A., Aitken, M., et al. (2023). A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. Nature *624*, 317–332. https://doi.org/10.1038/s41586-023-06812-z.
30. Tabula Muris Consortium; Overall coordination; Logistical coordination; Organ collection and processing; Library preparation and sequencing; Writing group; Supplemental text writing group; Principal investigators sequencing; Computational data analysis; Cell type annotation (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature *562*, 367–372. https://doi.org/10.1038/s41586-018-0590-4.

31. Cusanovich, D.A., Hill, A.J., Aghamirzaie, D., Daza, R.M., Pliner, H.A., Berletch, J.B., Filippova, G.N., Huang, X., Christiansen, L., DeWitt, W.S., et al. (2018). A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell *174*, 1309–1324.e18. https://doi.org/10.1016/j.cell.2018.06.052.

32. Mimitou, E.P., Lareau, C.A., Chen, K.Y., Zorzetto-Fernandes, A.L., Hao, Y., Takeshima, Y., Luo, W., Huang, T.S., Yeung, B.Z., Papalexi, E., et al. (2021). Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. Nat. Biotechnol. *39*, 1246–1258. https://doi.org/10.1038/s41587-021-00927-2.

33. Han, Y., Wang, Y., Dong, X., Sun, D., Liu, Z., Yue, J., Wang, H., Li, T., and Wang, C. (2023). TISCH2: expanded datasets and new tools for single-cell transcriptome analyses of the tumor microenvironment. Nucleic Acids Res. *51*, D1425–D1431. https://doi.org/10.1093/nar/gkac959.

34. Ma, L., Hernandez, M.O., Zhao, Y., Mehta, M., Tran, B., Kelly, M., Rae, Z., Hernandez, J.M., Davis, J.L., Martin, S.P., et al. (2019). Tumor Cell Biodiversity Drives Microenvironmental Reprogramming in Liver Cancer. Cancer Cell *36*, 418–430.e6. https://doi.org/10.1016/j.ccell.2019.08.007.

35. Zhang, Y., Xiang, G., Jiang, A.Y., Lynch, A., Zeng, Z., Wang, C., Zhang, W., Fan, J., Kang, J., Gu, S.S., et al. (2023). MetaTiME integrates single-cell gene expression to characterize the meta-components of the tumor immune microenvironment. Nat. Commun. *14*, 2634. https://doi.org/10.1038/s41467-023-38333-8.

36. Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. *20*, 163–172. https://doi.org/10.1038/s41590-018-0276-y.

37. Stahl, P.L., Salmen, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science *353*, 78–82. https://doi.org/10.1126/science.aaf2403.

38. Nascimento, M.A., Biagiotti, S., Herranz-Pérez, V., Santiago, S., Bueno, R., Ye, C.J., Abel, T.J., Zhang, Z., Rubio-Moll, J.S., Kriegstein, A.R., et al. (2024). Protracted neuronal recruitment in the temporal lobes of young children. Nature *626*, 1056–1065. https://doi.org/10.1038/s41586-023-06981-x.

39. Chen, J., Xu, H., Tao, W., Chen, Z., Zhao, Y., and Han, J.D.J. (2023). Transformer for one stop interpretable cell type annotation. Nat. Commun. *14*, 223. https://doi.org/10.1038/s41467-023-35923-4.

40. Tabula Sapiens Consortium*; Jones, R.C., Karkanias, J., Krasnow, M.A., Pisco, A.O., Quake, S.R., Salzman, J., Yosef, N., Bulthaup, B., Brown, P., et al. (2022). The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. Science *376*, eabl4896. https://doi.org/10.1126/science.abl4896.

41. Tepe, B., Hill, M.C., Pekarek, B.T., Hunt, P.J., Martin, T.J., Martin, J.F., and Arenkiel, B.R. (2018). Single-Cell RNA-Seq of Mouse Olfactory Bulb Reveals Cellular Heterogeneity and Activity-Dependent Molecular Census of Adult-Born Neurons. Cell Rep. *25*, 2689–2703.e3. https://doi.org/10.1016/j.celrep.2018.11.034.

42. Nagayama, S., Homma, R., and Imamura, F. (2014). Neuronal organization of olfactory bulb circuits. Front. Neural Circuits *8*, 98. https://doi.org/10.3389/fncir.2014.00098.

43. Ma, Y., and Zhou, X. (2022). Spatially informed cell-type deconvolution for spatial transcriptomics. Nat. Biotechnol. *40*, 1349–1359. https://doi.org/10.1038/s41587-022-01273-7.

44. Cable, D.M., Murray, E., Zou, L.S., Goeva, A., Macosko, E.Z., Chen, F., and Irizarry, R.A. (2022). Robust decomposition of cell type mixtures in spatial transcriptomics. Nat. Biotechnol. *40*, 517–526. https://doi.org/10.1038/s41587-021-00830-w.

45. Lopez, R., Li, B., Keren-Shaul, H., Boyeau, P., Kedmi, M., Pilzer, D., Jelinski, A., Yofe, I., David, E., Wagner, A., et al. (2022). DestVI identifies continuums of cell types in spatial transcriptomics data. Nat. Biotechnol. *40*, 1360–1369. https://doi.org/10.1038/s41587-022-01272-8.

46. Dong, R., and Yuan, G.C. (2021). SpatialDWLS: accurate deconvolution of spatial transcriptomic data. Genome Biol. *22*, 145. https://doi.org/10.1186/s13059-021-02362-7.

47. Bae, S., Na, K.J., Koh, J., Lee, D.S., Choi, H., and Kim, Y.T. (2022). CellDART: cell type inference by domain adaptation of single-cell and spatial transcriptomic data. Nucleic Acids Res. *50*, e57. https://doi.org/10.1093/nar/gkac084.

48. He, S., Bhatt, R., Brown, C., Brown, E.A., Buhr, D.L., Chantranuvatana, K., Danaher, P., Dunaway, D., Garrison, R.G., Geiss, G., et al. (2022). High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. Nat. Biotechnol. *40*, 1794–1806. https://doi.org/10.1038/s41587-022-01483-z.

49. Travaglini, K.J., Nabhan, A.N., Penland, L., Sinha, R., Gillich, A., Sit, R.V., Chang, S., Conley, S.D., Mori, Y., Seita, J., et al. (2020). A molecular cell atlas of the human lung from single-cell RNA sequencing. Nature *587*, 619–625. https://doi.org/10.1038/s41586-020-2922-4.

50. Marino, K.V., Cagnoni, A.J., Croci, D.O., and Rabinovich, G.A. (2023). Targeting galectin-driven regulatory circuits in cancer and fibrosis. Nat. Rev. Drug Discov. *22*, 295–316. https://doi.org/10.1038/s41573-023-00636-2.

51. Goel, H.L., and Mercurio, A.M. (2013). VEGF targets the tumour cell. Nat. Rev. Cancer *13*, 871–882. https://doi.org/10.1038/nrc3627.

52. Kuppe, C., Ramirez Flores, R.O., Li, Z., Hayat, S., Levinson, R.T., Liao, X., Hannani, M.T., Tanevski, J., Wünnemann, F., Nagai, J.S., et al. (2022). Spatial multi-omic map of human myocardial infarction. Nature *608*, 766–777. https://doi.org/10.1038/s41586-022-05060-x.

53. Amrute, J.M., Luo, X., Penna, V., Yang, S., Yamawaki, T., Hayat, S., Bredemeyer, A., Jung, I.-H., Kadyrov, F.F., Heo, G.S., et al. (2024). Targeting immune–fibroblast cell communication in heart failure. Nature *635*, 423–433. https://doi.org/10.1038/s41586-024-08008-5.

54. Ma, Y., and Zhou, X. (2024). Accurate and efficient integrative reference-informed spatial domain detection for spatial transcriptomics. Nat. Methods *21*, 1231–1244. https://doi.org/10.1038/s41592-024-02284-9.

55. Muhl, L., Mocci, G., Pietila, R., Liu, J., He, L., Genove, G., Leptidis, S., Gustafsson, S., Buyandelger, B., Raschperger, E., et al. (2022). A single-cell transcriptomic inventory of murine smooth muscle cells. Dev. Cell *57*, 2426–2443.e2426. https://doi.org/10.1016/j.devcel.2022.09.015.

56. Galdos, F.X., Lee, C., and Wu, S.M. (2024). Cardiac ACTN2 enhancer regulates cardiometabolism and maturation. Nat. Cardiovasc. Res. *3*, 616–618. https://doi.org/10.1038/s44161-024-00483-3.

57. Kattih, B., Boeckling, F., Shumliakivska, M., Tombor, L., Rasper, T., Schmitz, K., Hoffmann, J., Nicin, L., Abplanalp, W.T., Carstens, D.C., et al. (2023). Single-nuclear transcriptome profiling identifies persistent fibroblast activation in hypertrophic and failing human hearts of patients with longstanding disease. Cardiovasc. Res. *119*, 2550–2562. https://doi.org/10.1093/cvr/cvad140.

58. Morgan, B.P., and Harris, C.L. (2015). Complement, a target for therapy in inflammatory and degenerative diseases. Nat. Rev. Drug Discov. *14*, 857–877. https://doi.org/10.1038/nrd4657.

59. Frangogiannis, N.G., and Kovacic, J.C. (2020). Extracellular Matrix in Ischemic Heart Disease, Part 4/4: JACC Focus Seminar. J. Am. Coll. Cardiol. *75*, 2219–2235. https://doi.org/10.1016/j.jacc.2020.03.020.

60. Greenwald, A.C., Darnell, N.G., Hoefflin, R., Simkin, D., Mount, C.W., Gonzalez Castro, L.N., Harnik, Y., Dumont, S., Hirsch, D., Nomura, M., et al. (2024). Integrative spatial analysis reveals a multi-layered organization of glioblastoma. Cell *187*, 2485–2501.e26. https://doi.org/10.1016/j.cell.2024.03.029.

61. Neftel, C., Laffy, J., Filbin, M.G., Hara, T., Shore, M.E., Rahme, G.J., Richman, A.R., Silverbush, D., Shaw, M.L., Hebert, C.M., et al. (2019). An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. Cell *178*, 835–849.e21. https://doi.org/10.1016/j.cell.2019.06.024.

62. Nomura, M., Spitzer, A., Johnson, K., Garofano, L., Nehar-Belaid, D., Oh, Y.T., Anderson, K.J., Najac, R.D., Bussema, L., Varn, F., et al. (2023). EPCO-37. DISSECTING GBM EVOLUTION FOLLOWING STANDARD-OF-CARE BY LARGE-SCALE LONGITUDINAL SINGLE NUCLEUS

RNA-SEQUENCING. Neuro Oncol. *25*, v132. https://doi.org/10.1093/neuonc/noad179.0499.

63. Severin, C., Benitez-Hernandez, J., Dumitru, C.D., Beltran-Raygoza, N., Flandin-Blety, P., Escoubet, L., and Rolfe, M. (2020). DDRE-26. TARGETING THE UNFOLDED PROTEIN RESPONSE POTENTIATES MARIZOMIB ANTITUMOR ACTIVITY IN GLIOBLASTOMA. Neuro Oncol. *22*, ii67. https://doi.org/10.1093/neuonc/noaa215.271.
64. Wang, W., Li, T., Cheng, Y., Li, F., Qi, S., Mao, M., Wu, J., Liu, Q., Zhang, X., Li, X., et al. (2024). Identification of hypoxic macrophages in glioblastoma with therapeutic potential for vasculature normalization. Cancer Cell *42*, 815–832.e12. https://doi.org/10.1016/j.ccell.2024.03.013.
65. Sattiraju, A., Kang, S., Giotti, B., Chen, Z., Marallano, V.J., Brusco, C., Ramakrishnan, A., Shen, L., Tsankov, A.M., Hambardzumyan, D., et al. (2023). Hypoxic niches attract and sequester tumor-associated macrophages and cytotoxic T cells and reprogram them for immunosuppression. Immunity *56*, 1825–1843.e6. https://doi.org/10.1016/j.immuni.2023.06.017.
66. Song, C.Y., Wu, C.Y., Lin, C.Y., Tsai, C.H., Chen, H.T., Fong, Y.C., Chen, L.C., and Tang, C.H. (2024). The stimulation of exosome generation by visfatin polarizes M2 macrophages and enhances the motility of chondrosarcoma. Environ. Toxicol. *39*, 3790–3798. https://doi.org/10.1002/tox.24236.
67. Wang, C., Li, Y., Wang, L., Han, Y., Gao, X., Li, T., Liu, M., Dai, L., and Du, R. (2024). SPP1 represents a therapeutic target that promotes the progression of oesophageal squamous cell carcinoma by driving M2 macrophage infiltration. Br. J. Cancer *130*, 1770–1782. https://doi.org/10.1038/s41416-024-02683-x.
68. Hai, L., Hoffmann, D.C., Wagener, R.J., Azorin, D.D., Hausmann, D., Xie, R., Huppertz, M.C., Hiblot, J., Sievers, P., Heuer, S., et al. (2024). A clinically applicable connectivity signature for glioblastoma includes the tumor network driver CHI3L1. Nat. Commun. *15*, 968. https://doi.org/10.1038/s41467-024-45067-8.
69. Cui, G., Liu, W., Sun, X., Bai, Y., Ding, M., Zhao, N., Guo, J., Qu, D., Wang, S., Qin, L., and Yang, Y. (2025). RNA-seq shows Angiopoietin-like 4 promotes hepatocellular carcinoma progression by inducing M2 polarization of tumor-associated macrophages. Int. J. Biol. Macromol. *287*, 138523. https://doi.org/10.1016/j.ijbiomac.2024.138523.
70. Galvao, I., Dias, A.C., Tavares, L.D., Rodrigues, I.P., Queiroz-Junior, C.M., Costa, V.V., Reis, A.C., Ribeiro Oliveira, R.D., Louzada-Junior, P., Souza, D.G., et al. (2016). Macrophage migration inhibitory factor drives neutrophil accumulation by facilitating IL-1beta production in a murine model of acute gout. J. Leukoc. Biol. *99*, 1035–1043. https://doi.org/10.1189/jlb.3MA0915-418R.
71. Kong, L., Subramanian, S., Segerstolpe, Å., Tran, V., Shih, A.R., Carter, G.T., Kunitake, H., Twardus, S.W., Li, J., Gandhi, S., et al. (2025). Single-cell and spatial transcriptomics of stricturing Crohn's disease highlights a fibrosis-associated network. Nat. Genet. *57*, 1742–1753. https://doi.org/10.1038/s41588-025-02225-y.
72. Brinkmalm, A., Brinkmalm, G., Honer, W.G., Frölich, L., Hausner, L., Minthon, L., Hansson, O., Wallin, A., Zetterberg, H., Blennow, K., and Öhrfelt, A. (2014). SNAP-25 is a promising novel cerebrospinal fluid biomarker for synapse degeneration in Alzheimer's disease. Mol. Neurodegener. *9*, 53. https://doi.org/10.1186/1750-1326-9-53.
73. Karns, L.R., Ng, S.C., Freeman, J.A., and Fishman, M.C. (1987). Cloning of complementary DNA for GAP-43, a neuronal growth-related protein. Science *236*, 597–600. https://doi.org/10.1126/science.2437653.
74. Kolz, A., de la Rosa, C., Syma, I.J., McGrath, S., Kavaka, V., Schmitz, R., Thomann, A.S., Kerschensteiner, M., Beltran, E., Kawakami, N., and Peters, A. (2025). T-B cell cooperation in ectopic lymphoid follicles propagates CNS autoimmunity. Sci. Immunol. *10*, eadn2784. https://doi.org/10.1126/sciimmunol.adn2784.
75. Zhao, L., Jin, S., Wang, S., Zhang, Z., Wang, X., Chen, Z., Wang, X., Huang, S., Zhang, D., and Wu, H. (2024). Tertiary lymphoid structures in diseases: immune mechanisms and therapeutic advances. Signal Transduct. Target. Ther. *9*, 225. https://doi.org/10.1038/s41392-024-01947-5.
76. Kotliar, D., Veres, A., Nagy, M.A., Tabrizi, S., Hodis, E., Melton, D.A., and Sabeti, P.C. (2019). Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. eLife *8*, e43803. https://doi.org/10.7554/eLife.43803.
77. Shen, J., Xu, Y., Zhang, S., Lyu, S., Huo, Y., Zhu, Y., Tang, K., Mou, J., Li, X., Hoyle, D.L., et al. (2021). Single-cell transcriptome of early hematopoiesis guides arterial endothelial-enhanced functional T cell generation from human PSCs. Sci. Adv. *7*, eabi9787. https://doi.org/10.1126/sciadv.abi9787.
78. Cao, Y., Wang, X., and Peng, G. (2020). SCSA: A Cell Type Annotation Tool for Single-Cell RNA-seq Data. Front. Genet. *11*, 490. https://doi.org/10.3389/fgene.2020.00490.
79. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.
80. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15. https://doi.org/10.1186/s13059-017-1382-0.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
CRC data	Han et al.[33]	http://tisch.comp-genomics.org/home/
Mouse brain scRNA data	Kleshchevnikov et al.[26]	https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-11115
Human brain scRNA datasets	Zemke et al.,[27] Zhu et al.,[28] Yao et al.[29]	https://portal.brain-map.org/atlases-and-data/rnaseq; GEO: GSE229169; https://cellxgene.cziscience.com/collections/ceb895f4-ff9f-403a-b7c3-187a9657ac2c
Mouse atlas scATAC data	Cusanovich et.al[31]	https://atlas.gs.washington.edu/mouse-atac/
Mouse atlas scRNA data	Tabula Muris et al.[30]	https://tabula-muris.sf.czbiohub.org/data
PBMC	Mimitou et al.[32]	https://github.com/PeterZZQ/scMoMaT/tree/main/data/real/ASAP-PBMC
CHOL data	Ma et al.[34]	http://tisch.comp-genomics.org/home/
The human neural cell subtype dataset	Zemke et al.,[27] Yao et al.,[29] Nascimento et al.[38]	https://cellxgene.cziscience.com/collections/cae8bad0-39e9-4771-85a7-822b0e06de9f
The cross-human tissue atlas dataset	Tabula Sapiens et al.[40]	https://cellxgene.cziscience.com/collections/e5f58829-1a66-40b5-a624-9046778e74f5
The mouse olfactory bulb spatial dataset	Stahl et al.[37]	https://ftp.cngb.org/pub/SciRAID/stomics/STDS0000017/stomics/
The mouse olfactory bulb scRNA dataset	Tepe et al.[41]	GEO: GSE121891
The NSCLC spatial dataset	He et al.[48]	https://nanostring.com/products/cosmx-spatial-molecular-imager/ffpe-dataset/nsclc-ffpe-dataset/
The NSCLC scRNA dataset	Travaglini et al.[49]	https://www.synapse.org/Synapse:syn21041850/files/
Mouse coronal brain spatial dataset	10X Genomics	https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-libraries-of-mouse-brain-he
Myocardial infarction multi-omic data	Kuppe et al.[52]	https://zenodo.org/records/6578047
Stricturing Crohn's disease dataset	Kong et al.[71]	https://zenodo.org/records/14509802
Human hematopoietic development dataset	Shen et al.[77]	GEO: GSE145859
The glioblastoma dataset	Greenwald.et al.[60]	https://github.com/tiroshlab/Spatial_Glioma
Software and algorithms		
BBKNN (version: 1.6.0)	Polanski et al.[21]	https://github.com/Teichlab/bbknn
Harmony (verson: 1.2.1)	Korsunsky et al.[17]	https://github.com/immunogenomics/harmony
scGen (version: 2.1.1)	Lotfollahi et al.[20]	https://github.com/theislab/scgen
Scanorama (version: 1.7.4)	Hie et al.[19]	https://github.com/brianhie/scanorama
SeuratWrappers (version: 0.3.0)	Haghverdi et al.[16]	https://github.com/satijalab/seurat-wrappers
Omicverse (version:1.5.4)	Cao et al.[78]	https://github.com/Starlitnightly/omicverse
MetaTime (version:1.3.0)	Zhang et al.[35]	https://github.com/yi-zhang/MetaTiME
SingleR (version:1.4.1)	Aran et al.[36]	https://github.com/dviraran/SingleR

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
TOSICA (version:1.0.0)	Chen et al.[39]	https://github.com/JackieHanLab/TOSICA
Seurat (version:4.2.0 & 5.1.0)	Hao et al.[79]	https://github.com/satijalab/seurat
Scanpy (version:1.9.5)	Wolf et al.[80]	https://github.com/scverse/scanpy
CARD (version:1.0.0)	Ma et al.[43]	https://github.com/YMa-lab/CARD
Spacexr (version:2.2.1)	Cable et al.[44]	https://github.com/dmcable/spacexr
Giotto (version:1.1.2)	Dong et al.[46]	https://github.com/drieslab/Giotto
scvi-tools (version:1.0.4)	Lopez et al.[45]	https://github.com/scverse/scvi-tools
SSpMosaic (version: 0.0.0.9)	This paper	GitHub: https://github.com/compbioNJU/SSpMosaic; Zenodo: https://doi.org/10.5281/zenodo.17549304
cNMF (version: 1.7.0)	Kotliar et al.[76]	https://github.com/dylkot/cNMF
scArches (version: 0.6.1)	Lotfollahi et al.[8]	https://docs.scarches.org/en/latest/
Spectra (version: 0.1.0)	Kunes et al.[7]	https://github.com/dpeerlab/spectra
CellDART (version:0.1.2)	Bae et al.[47]	https://github.com/mexchy1000/CellDART

METHOD DETAILS

Overview of SSpMosaic

SSpMosaic is designed to integrate multiple types of omics data, including single-cell transcriptomics (scRNA-seq or snRNA-seq), single-cell chromatin accessibility (scATAC-seq), single-cell proteomics, or spatial transcriptomics, which may originate from one or more experimental batches. In the preprocessing phase, SSpMosaic performs a series of steps to prepare data from diverse modalities, ensuring comparability across datasets. To facilitate cross-modality integration, SSpMosaic maps molecular features to a unified gene space by harmonizing gene expression profiles, chromatin accessibility signals, and protein abundance measurements. Subsequently, SSpMosaic applies unsupervised clustering to each preprocessed dataset to identify cell or spatial clusters and determines differentially expressed genes (DEGs) for each cluster. These DEGs are then used to construct differentially co-expressed gene programs, referred to as gene programs, which encapsulate functionally relevant transcriptional signatures. Following gene program identification, SSpMosaic quantifies program activity for each cell or spatial spot, implementing filtering strategies to define and extract cluster-specific programs while reducing technical noise. Furthermore, to improve interpretability and integration across datasets, SSpMosaic employs a network propagation algorithm to assess similarities between programs across different clusters, generating a structured distance matrix that supports hierarchical clustering. This process results in the formation of metaprograms, which are higher-order representations of gene programs with weighted gene contributions. Leveraging these metaprograms, SSpMosaic provides a unified framework for various downstream analytical tasks, including single-cell type annotation, cross-batch and cross-modality data integration, annotation of cell types in spatial transcriptomics, and comparative spatial transcriptomic analysis across different tissue sections. By integrating multi-omics data in a biologically meaningful manner, SSpMosaic enhances the resolution and interpretability of cellular and spatial heterogeneity, facilitating deeper insights into complex biological systems.

Unifying feature space

In multi-omics data, discrepancies in feature spaces arise due to the distinct molecular entities captured by different modalities. For example, transcriptomic features correspond to gene expression levels, whereas chromatin accessibility profiling measures open chromatin regions, which are often represented as genomic coordinates rather than discrete genes. Similarly, proteomic data quantify protein abundance, introducing another layer of complexity in cross-modality integration.

To ensure comparability across different omics datasets, we standardize all feature spaces at the gene level. This is achieved by mapping chromatin accessibility peaks and protein markers to their associated genes, enabling a unified representation of molecular features across modalities. By harmonizing feature definitions, SSpMosaic facilitates seamless data integration, preserves biological relevance, and enhances the interpretability of shared molecular programs across single-cell and spatial omics data.

For single-cell proteomics data, antibody-derived tags (ADTs) reflect the expression of proteins associated with specific gene products. Due to the technical characteristics of different sequencing platforms, the correspondence between ADTs and genes is not always straightforward. For ADTs that have a one-to-one correspondence with genes, we define the abundance of a gene as the total count of its corresponding ADTs. In cases where multiple ADTs correspond to a single gene, we calculate the gene's abundance by averaging the counts of these ADTs.

For single-cell chromatin accessibility data, features are represented as peaks, which denote specific regions on chromosomes. To compute a gene activity score for a specific gene g, we employ a weighted sum approach. Peaks outside the ±100 kb range of the

transcription start site (TSS) of gene *g* receive a weight of 0. Peaks that overlap with the gene body or the region upstream of the TSS by 5 kb are assigned a weight of 1. Otherwise, weights are distributed according to the following exponential function:

$$e^{-\frac{Distance}{5000}}$$

where *Distance* refers to the distance between the peak and the TSS of gene *g*.

For cross-species data integration, differences in gene nomenclature and evolutionary divergence present significant challenges. To address this, SSpMosaic employs orthologous gene mapping, ensuring that functionally equivalent genes across species are aligned within a common feature space. We restrict the analysis to strict one-to-one ortholog mapping, leveraging high-confidence annotations from the Ensembl database. This approach minimizes ambiguity in gene correspondences, facilitating more accurate cross-species comparisons and enhancing the transferability of gene programs and metaprograms across evolutionary contexts.

Constructing differential co-expression gene programs

We utilize gene co-expression programs as representatives of functional units, based on the hypothesis that genes with high positive correlation among their transcripts likely reflect specific biological functions due to regulatory relationships within the cell. Consequently, gene co-expression programs can be regarded as functional programs. Additionally, since proteins serve as the ultimate products of gene expression and perform various biological functions, while the accessibility of genes is the prerequisite for their transcription. Thus, a strong positive correlation between protein abundance and gene accessibility also applies to the aforementioned assumption. Therefore, the gene co-accessibility and co-translation programs identified will also be considered functional programs.

In the SSpMosaic workflow, we begin by calculating differential genes for each cluster. For the cell or spot set $C_{i,j}$ in batch *i* and cluster *j*, the corresponding set of differential genes is denoted as $G_{i,j}$. The expression matrix for these differential genes is referred to as $X_{i,j}$, where the rows represent genes and the columns represent cells. We then compute the co-expression network for $X_{i,j}$ to identify differential co-expression programs for each cluster using hdWGCNA.[10]

The details of co-expression program calculation are as follows: first construct metacells are constructed to address the sparsity of single-cell data, all metacell sets denoted as $T_{i,j}$.Gene expressions are averaged across all constituent cells in $T_{i,j}$, resulting in the metacell gene expression matrix, where the rows correspond to$G_{i,j}$ and the columns correspond to $T_{i,j}$. We then build the co-expression network using $X_{i,j}^{meta}$. For genes *a* and *b*, their expression vectors in $X_{i,j}^{meta}$ are denoted as x_a and x_b, respectively. We calculate the expression correlation *cor*(*a*,*b*) using the Pearson correlation coefficient.

Under our foundational hypothesis, we seek genes with high positive correlations, thus requiring a transformation of the correlation to ensure that negative correlations yield lower absolute values. Then a soft-thresholding strategy is applied to emphasize strong correlations, ultimately yielding the adjacency matrix Cor:

$$cor'(a,b) = \frac{1+cor(a,b)}{2}$$

$$Cor_{a,b} = sign(cor(a,b)) \cdot (cor'(a,b))^{\beta}$$

cor′(*a*,*b*) represents the transformed correlation value between *a* and *b*, $Cor_{a,b}$ represents the corresponding values of gene *a* and *b* in the adjacency matrix Cor, β is a soft-thresholding hyperparameter.

Next, topological overlap is assessed to make sure the high-correlation gene pair also share a significant number of common high-correlation genes in the correlation network, the topological overlap between *a* and *b* is calculated as follow:

$$TOM_{a,b} = \frac{\left| Cor_{a,b} + \sum_{g \neq a,b} Cor_{a,g} Cor_{g,b} \right|}{min(k_a, k_b) - |Cor_{a,b}|}$$

Here, k_a andk_b represent the weighted degrees of genes *a* and *b*:

$$k_a = \sum_{g \in G_{i,j}} |Cor_{g,a}|$$

We compute the pairwise topological overlap for genes in $G_{i,j}$ to derive the topological overlap matrix TOM. Subsequently, we calculate the TOM distance matrix as 1-TOM and apply the Dynamic Tree Cut algorithm for hierarchical clustering to obtain gene programs. We repeat these operations for each cluster in each batch, ultimately identifying their corresponding programs.

Cell cluster-specific program selection

Upon identifying gene programs, we generate a series of gene sets for each cluster in each dataset, which serve as potential candidates for cluster-specific gene programs. We then utilize the concept of information gain for the selection of gene sets, ultimately resulting in cluster-specific programs, while ensuring that as many clusters as possible have their own matching programs. We denote the k-th gene set corresponding to cluster *j* in batch *i* as $GS_{i,j,k}$. For all cells in batch *i*, we define the collection of cells as

C_i, with $C_{i,j}$ representing all cells in cluster j of batch i. Next, we evaluate the specificity of these gene programs within their respective clusters to filter for cluster-specific programs. We utilize the "Addmodulescore" function to calculate the score of $GS_{i,j,k}$ across C_i, resulting in a score vector denoted as $V_{i,j,k}$. We first compute the standard deviation of this score vector, $SD(V_{i,j,k})$. Subsequently, we remove the cells in $C_{i,j}$ from $V_{i,j,k}$, with the remaining elements represented as $V'_{i,j,k}$. We then calculate the standard deviation of $V'_{i,j,k}$, denoted as $SD\left(V'_{i,j,k}\right)$. The difference in standard deviations is defined as:

$$\Delta_{i,j,k} = SD\left(V_{i,j,k}\right) - SD\left(V'_{i,j,k}\right)$$

When $GS_{i,j,k}$ is specifically expressed in $C_{i,j}$ an increase in $\Delta_{i,j,k}$ is expected. We set a default threshold of 0.01, retaining the program $GS_{i,j,k}$ if $\Delta_{i,j,k}>0.01$. The default threshold of 0.01 for $\Delta_{i,j,k}$ was determined through a data-driven approach and extensive testing to ensure it effectively balances the specificity and sensitivity in identifying cluster-specific gene programs (Figure S22).

With lower threshold, although each cluster can retain its corresponding program, the expression specificity of some programs may not be strong enough. With higher threshold, while the specificity of programs is ensured, some clusters may not have their corresponding programs. Setting the threshold at 0.01 not only ensures that the retained programs are specifically overexpressed in the corresponding clusters but also enables as many clusters as possible to have corresponding programs after selection.

This criterion for $\Delta_{i,j,k}$ is applied to all programs to filter those with high expression specificity, while ensuring that as many cluster-associated programs as possible are retained.

In addition to $\Delta_{i,j,k}$, we also provide two alternative metrics for program selection. One metric $\Delta'_{i,j,k}$ also utilizes the concept of information gain, but replaces the standard deviation with the mean value:

$$\Delta'_{i,j,k} = mean\left(V_{i,j,k}\right) - mean\left(V'_{i,j,k}\right)$$

For the other metric, we denote the score vector of the cells in $C_{i,j}$ as $V''_{i,j,k}$. We then calculate the proportion of elements in $V''_{i,j,k}$ that exceed zero, denoted as $\Delta''_{i,j,k}$:

$$\Delta''_{i,j,k} = \frac{\left|\left\{x > 0 | x \in V''_{i,j,k}\right\}\right|}{\left|V''_{i,j,k}\right|}$$

These two metrics are not enabled by default. Subsequently, we merge programs within the same cluster, ultimately keeping just one program $GS_{i,j}$ for each cluster.

Calculating program distances using network propagation

To quantify the similarity between gene sets, we employ a network propagation-based algorithm to analyze gene interactions and signal transmission within a network, allowing for a comprehensive evaluation of relationships between gene sets. We first provide a background network (e.g., a protein-protein interaction network using the STRING PPI network v12.0 which contains 10,036 genes and 1,828,234 interactions) or other relevant gene sets. If two gene sets represent similar biological functions, the majority of genes in these sets should be closely located within the background network, despite potential additions or omissions.

We construct the following mathematical model to describe this process. First, we define the adjacency matrix W of the background network, which is a symmetric matrix. To reduce false positives in the background network and satisfy the conservation of heat during the network propagation process, we convert W into a Boolean matrix. This also reduces memory usage from $O(N^2)$ to O(E) (N = number of genes, E = number of edges), as seen in the STRING network where sparsity cuts effective computational complexity by ~95%—with only ~1.8% of gene pairs interacting—thus enhancing both accuracy and efficiency.

Next, we initialize the heat across the network. For each gene set, the initial heat vector is created, where the genes in the gene set are assigned a value of 1/(size of the gene set), and all other elements are initialized to 0. To ensure that the total heat remains constant at 1 during propagation, we standardize W to obtain the normalized matrix W':

$$W'_{i,j} = W_{i,j} \Big/ \sum\nolimits_k W_{k,j}$$

We then implement the Random Walk with Restart (RWR) algorithm for network propagation, expressed as:

$$H^t = \alpha W' H^{t-1} + (1 - \alpha) Y_n$$

Here, H^t represents the heat vector at iteration t, Y_n corresponds to the initial heat vector for gene program G_n with $Y_n = H^0$, and α is a tunable hyperparameter within the range [0,1] This parameter dictates the fraction of total heat each node passes to its neighboring nodes during an iteration, retaining 1-α for itself. By default, α is set to 0.5.

For biological background networks with numerous nodes and low connectivity, the computational burden for achieving convergence is substantial. To address the slow convergence issue, we directly solve for the converged heat vector using a closed-form solution, based on the condition:

$$H^t = H^{t-1} = H$$

The final heat distribution vector H can be expressed as:

$$H = (I - \alpha W')^{-1}(1 - \alpha)Y_n$$

This indicates that the final heat distribution H depends solely on the input gene set, reflecting the overall influence of the gene program. By applying this process to each gene program, we derive a collection of final heat distribution vectors. In practice, the matrix $(I - \alpha W')^{-1}(1 - \alpha)$ is precomputed at the start of the workflow, which allows each gene program's final heat distribution vector to be computed in a single step via matrix-vector multiplication, avoiding redundant calculations.

To measure the similarity between gene programs, we utilize the cosine distance between the final heat distribution vectors. For any two gene programs n and m, with corresponding final heat distribution vectors H_n andH_m, the cosine distance is defined as:

$$Distance(n, m) = 1 - \frac{H_n \cdot H_m}{\|H_n\|\|H_m\|}$$

Based on these cosine distances, we construct a distance matrix between programs. We then use the silhouette coefficient to automatically select the optimal number of clusters. Subsequently, we apply hierarchical clustering to further refine the selection of metaprograms that represent functionally similar gene programs (Figure S23).

Data integration

In data integration task of SSpMosaic, we first derive metaprograms as described above and each cell in every batch is assigned a unique MP label, which we subsequently use in place of the original cluster or cell type labels during integration.

To identify neighbors of metaprograms, we define the expected number of neighbors as N. Denote the PCA reduction matrix for batch i as P_i and the PCA matrix for cells labeled with metaprogram m in batch i as $P_{i,m}$. We compute the mean of each column in $P_{i,m}$ to derive the representative vector $V_{i,m}$ for metaprogram m in batch i. The Euclidean distance between these representative vectors quantifies the distances among different MP labels within the same batch.

This process is repeated for all MP labels in batch i, resulting in a metaprogram distance matrix. We then identify the N nearest neighbors for each metaprogram in this matrix, yielding the neighbor set for metaprogram m in batch i, denoted as $U_{i,m}$. By default, n is set to 1, resulting in $U_{i,m}$ containing only m itself.

Next, we repeat the neighbor-finding process within each batch and unite the neighbor sets across batches for the same MP label to form the final neighbor set U_m.

For any cell c from batch i with MP label M and PCA vector V_c, we seek its k -nearest neighbors across batches. For a cell d from batch h, with PCA vector V_d, we impose that MP label of cell d belongs to U_m and compute the distance using Euclidean distance between PCA vectors.

We utilize the Approximate Nearest Neighbor (ANN) algorithm ANNOY to find the k closest cells to c in batch h, recording the distances between them. Moreover, when searching for neighbors of c in batch h, if both belong to the same modality (or species), we look for a reduced number of neighbors k_1; otherwise, we search for a larger number k_2 (where $k_1 \leq k_2$) to balance the effects of modality or species.

Repeating this for all cells across batches results in a k-nearest neighbor distance matrix D, where $D_{d,c}$ with a non-zero distance value indicates that cell d is among the k-nearest neighbors of cell c. Note that$D$is not necessarily symmetric as nearest neighbor relationships can be unidirectional.

Finally, we convert the directed graph matrix D into a symmetric matrix to facilitate integration and enable downstream analyses. We initialize a zero matrix $GRAPH$ of the same shape as D and define the neighbor set of cells c as S_c (excluding c itself). For any d in S_c, we transform the distance into a similarity measure:

$$similarity(c, d) = exp\left(-\frac{Distance(c, d) - m_c}{\beta_c}\right)$$

Where $m_c$represents the minimum distance from c to its other neighbors. The parameter β_c is chosen such that:

$$\sum_{d \in S_c} similarity(c, d) = \lambda \, log_2|S_c|$$

Where λ is set to 1 by default. Reducing λ will hasten the decay of similarity for distant neighbors, thereby emphasizing closer neighbors' importance.

Ultimately, we calculate the corresponding values of cell c and d in the $GRAPH$ matrix:

$$GRAPH_{c,d} = GRAPH_{d,c} = similarity(c, d) + similarity(d, c) - similarity(c, d) * similarity(d, c)$$

This process is repeated for all cells, resulting in the *GRAPH* matrix, which serves as the integrated output for subsequent analyses, including UMAP visualization (Figure 1B). It is important to note that the GRAPH matrix is directly used as the input for UMAP, ensuring that the visualization reflects the integrated cell-cell similarity information derived from the data integration process.

Cell type annotation

After constructing gene programs and calculating the distance matrix, we proceed with the automated annotation of cell types. The core idea of this process is to identify metaprogram s that characterize cell types by screening programs within a reference single-cell dataset, and then applying this annotation to the query dataset. The specific steps are as follows:

In the query dataset, we first perform unsupervised clustering to determine the cell set $C_{i,j}$ for each cluster. Next, we screen for metaprograms corresponding to each cluster and proceed with cell type annotation. For each cluster in the query dataset, we calculate the scoring value of the metaprogram. Let $C_{i,j}$ represent all cells in cluster j of batch i and GS_m represent the gene set of the metaprogram m, with its corresponding weight vector denoted as W_m.For each gene in GS_m, the corresponding weight value in W_m is the frequency of its occurrence in metaprogram m. In particular, to emphasize the cross-sample commonality, the weight values of genes that occur once are set to 0.5. Expression values of the genes in GS_m across $C_{i,j}$ are denoted as $E_{C_{i,j},m}$. We compute the weighted average expression to obtain the score vector $V_{i,j,m}$ for the metaprogram m in $C_{i,j}$:

$$V_{i,j,m} = \frac{1}{\sum_{g \in GS_m} W_{m,g}} \sum_{g \in GS_m} E_{C_{i,j},g} W_{m,g}$$

Here, g denotes a certain gene in GS_m and $E_{C_{i,j},g}$ is the expression value of gene g for cell s within $C_{i,j}$. We repeat these operations for each cluster in each batch and for each metaprogram.

To filter for cluster-specific metaprograms, we reference the previous program screening steps, basing on the concept of information gain, calculating the standard deviation difference to identify specifically expressed metaprograms. We calculate the standard deviation of the score vector of C_i, $V_{i,m}$ as $SD(V_{i,m})$. We then remove the cells in $C_{i,j}$ from this vector, resulting in a resulting vector $V'_{i,j,m}$ and calculate its standard deviation $SD\left(V'_{i,j,m}\right)$. We compute the standard deviation difference $\Delta_{i,j,m}$:

$$\Delta_{i,j,m} = SD(V_{i,m}) - SD\left(V'_{i,j,m}\right)$$

If $\Delta_{i,j,m}$>0.01(the default threshold), we retain the metaprogram m for cluster j of batch i, otherwise, it is discarded.

Among the retained metaprograms, we select the one with the highest score as the cell type annotation for that cluster. Specifically, for cluster j of batch i, we choose the metaprogram m^* that maximizes $\Delta_{i,j,m}$.

If there is no retained metaprogram, we annotate all cells in cluster j of batch i as UN.

Setting a lower threshold here may lead to more miscellaneous genes being retained in the gene programs, generating more mixed signals during the network propagation stage. This would ultimately weaken the programs' ability to represent specific cell types and reduce the specificity of program scores, potentially leading to more UN annotations. Conversely, setting a higher threshold would require program scores to exhibit stronger specificity to be selected for annotation, thus resulting in more programs being annotated as UN.

We identified 0.01 as the optimal threshold through extensive testing (Figure S22), which effectively removes miscellaneous genes to avoid mixed signals while preventing excessive UN annotations caused by overly stringent criteria.

Ultimately, each cluster in the query dataset is assigned the most suitable metaprogram annotation.

Annotation of spatial transcriptomics data

For a given spatial transcriptomics data, let X denote the gene expression matrix, G the set of all genes in the slice, and C the set of all spots in the slice. The rows of the matrix represent the genes G, and the columns represent the spots C. Additionally, let META be the set of all metaprogram s, and let m be a certain metaprogram, GS_m represent the gene set corresponding to metaprogram m.

First, we construct a weight matrix for the gene set of metaprogram s, denoted as W_{META}. In this matrix, rows represent genes and columns represent metaprograms, with the value in each position indicating the weight of a gene in its corresponding metaprogram. The gene set G' used in this matrix includes only those genes present in both the metaprogram s and the slice:

$$G' = G \cap \left(\bigcup_{m \in META} GS_m\right)$$

The gene weights are defined in the same manner as described in the "Cell Type Annotation". The gene weight is defined by its frequency of occurrence in the metaprogram. To emphasize commonality, the weight of genes that appear only once is set to 0.5.

Next, we align the gene expression matrix X and the gene weight matrix W_{META}. We filter the rows of X to retain only those corresponding to genes in G'. The filtered matrix is denoted as X', thereby ensuring that the genes in both matrices are unified.

Next, we apply non-negative least squares (NNLS) to compute the scores of all metaprograms across each spot. For a given spot c in the ST data, let X'_c be the gene expression vector from X' corresponding to spot c, and let V_c be the vector of scores for all metaprograms in spot c. The optimal score vector V_{c^*} is obtained by minimizing the sum of squared deviations:

$$V_{c^*} = \arg\min_{V_c \geq 0} \left(W_{META} V_c - X'_c\right)^T \left(W_{META} V_c - X'_c\right)$$

This process is repeated for all spots in the ST slice, resulting in the metaprogram score matrix R_{META}, where rows correspond to spots and columns correspond to metaprograms.

We then proceed to annotate the spots based on the metaprogram score matrix. Each spot can be labeled according to the metaprogram with the highest score, or if necessary, we can first normalize the scores and then assign labels. Two different normalization methods are provided: Z score normalization and min-max normalization. For metaprogram m, the Z score normalization is defined as follows:

$$R'_{META,m} = \frac{R_{META,m} - mean\left(R_{META,m}\right)}{SD\left(R_{META,m}\right)}$$

Repeating this process for all metaprograms results in the Z score normalized matrix R'_{META}. The label for each spot is then assigned to the metaprogram with the highest normalized score.

Alternatively, using min-max normalization, for metaprogram m, the normalized scores are computed as:

$$R''_{META,m} = \frac{R_{META,m} - min\left(R_{META,m}\right)}{max\left(R_{META,m}\right) - min\left(R_{META,m}\right)}$$

After repeating this process for all metaprogram s, we obtain the min-max normalized matrix R''_{META}. Each spot is then labeled with the metaprogram that has the highest normalized score in this matrix.

Encapsulation index

In spatial transcriptomics, let C be the set of all spots in a slice, and LABEL denote their labels. The distance between the spots is measured using the Euclidean distance between spatial coordinates.

To assess the encapsulation of a label l at spot c, we identify its six nearest neighbors S_c (excluding c) based on 10X Visium data. The corresponding labels are $LABEL_{S_c}$. The proportion $M_{c,others}$of neighbors labeled as ‘‘others’’ quantifies encapsulation:

$$M_{c,others} = \frac{\left|\{x = others | \; x \in LABEL_{S_c}\}\right|}{\left|LABEL_{S_c}\right|}$$

To calculate the average encapsulation $M_{l,others}$ across all spots labeled with l:

$$M_{l,others} = \frac{1}{|C_l|} \sum\nolimits_{c \in C_l} M_{c,others}$$

C_l denotes the set of all spots labeled with l. This can also be expressed as:

$$M_{l,others} = \frac{\left|\left\{x = others | \; x \in \bigcup_{c \in C_l} LABEL_{S_c}\right\}\right|}{6||C_l||}$$

A lower $M_{l,others}$ indicates greater encapsulation. To evaluate clustering, for each spot c with label l, we compute the proportion $M_{c,l}$ of neighbors also labeled l:

$$M_{c,l} = \frac{\left|\{x = l | \; x \in LABEL_{S_c}\}\right|}{\left|LABEL_{S_c}\right|}$$

The overall tendency M_l for clustering is measured as

$$M_l = \frac{1}{|C_l|} \sum\nolimits_{c \in C_l} M_{c,l}$$

For 10X Visium data, this can be reformulated as:

$$M_l = \frac{\left|\left\{x = l | \; x \in \bigcup_{c \in C_l} LABEL_{S_c}\right\}\right|}{6||C_l||}$$

For the whole slice, a higher value of M_l indicates a higher tendency of spots labeled with l to cluster together.

A higher M_l signifies stronger clustering. Finally, we define the Encapsulation Index (EI) for label *l* as the difference between the clustering tendency and the encapsulation degree concerning "others":

$$EI_l = M_l - M_{l,others}$$

Ultimately, a higher EI_lindicates that the spots labeled *l* are more clustered and encapsulated.

Pointwise mutual information calculation

To quantitatively evaluate the coherence between marker gene expressions and metaprogram scores,we calculate the pointwise mutual information between marker gene expressions and metaprogram scores as follows:

$$PMI(i,j) = log_2 \frac{P(i,j)}{P(i) * P(j)}$$

where i denotes a marker gene, j denotes a metaprogram,$P(i,j)$ denotes the joint probability of the marker gene i and the metaprogram j presenting in the same spot, $P(i)$ denotes the probability of the marker gene i present in spots, $P(j)$ denotes the probability of the metaprogram j present in spots. To minimize confounding by low-expression noise, we only consider a gene/metaprogram to be present in a spot if its expression/score is above the 80th percentile.

Colocalization score calculation

Let $S \in R^{N*K}$ denote the metaprogram scoring matrix, where N is the number of spots, K is the number of metaprograms, and $S_{i,j}$ represents the score of metaprogram j in spot i.We first normalized the matrix column-wise (per metaprogram) using *Z* score transformation to standardize scores across different metaprograms, resulting in the *Z* score normalized matrix *Z*:

$$Z_{i,j} = \frac{S_{i,j} - \mu_j}{\sigma_j}$$

where μ_j is the mean score of metaprogram j across all spots, and σ_j is the corresponding standard deviation.

Then for each spot (row), we applied the softmax function to obtain the meta program proportional values that sum to 1 within each spot, resulting in the softmax transformed matrix *T*:

$$T_{i,j} = \frac{exp(Z_{i,j})}{\sum_{k=1}^{K} exp(Z_{i,k})}$$

The colocalization score between metaprogram j and k is then measured as the product of the proportional values:

$$L_{i,j\&k} = T_{i,j} * T_{i,k}$$

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analyses were performed using R (v.4.3.2). The comparison of T cell marker expressions between modalities was conducted using a two-sided Wilcoxon rank-sum test. The ANOVA was applied to compare cell-cell communication probabilities between sample groups. The comparison of the encapsulation index between UPR and TAM was performed using a two-sided paired *t* test. The comparison of T cell–B cell colocalization scores between fibrotic slices and adjacent slices was conducted using a two-sided *t* test. A p-value <0.01 was considered statistically significant. Correlations between SSpMosaic metaprograms were performed with Pearson's correlation. More detailed quantitative and statistical analyses are described in the relevant sections of the STAR Methods and in figure legends.