CAESAR: a cross-technology and cross-resolution framework for spatial omics annotation

Jin Liu

liujinlab@cuhk.edu.cn

The Chinese University of Hong Kong, Shenzhen https://orcid.org/0000-0002-5707-2078

Wei Liu

Sichuan University

Xiaoran Chai

Duke-NUS Medical School

Xiao Zhang

The Chinese University of Hong Kong, Shenzhen

Zhixiang Lin

The Chinese University of Hong Kong

Article

Keywords:

Posted Date: October 22nd, 2024

DOI: https://doi.org/10.21203/rs.3.rs-5086440/v1

Additional Declarations: There is **NO** Competing Interest.

CAESAR: a cross-technology and cross-resolution framework for spatial omics annotation

Xiao Zhang[1†], Wei Liu[2†], Xiaoran Chai[3], Zhixiang Lin[4] and Jin Liu[1*]

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

[2] School of Mathematics, Sichuan University, Chengdu, China.

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

[4]Department of Statistics, The Chinese University of Hong Kong, Hong Kong, China

[†] Equal contributions

[*]Corresponding author. Email: liujinlab@cuhk.edu.cn

Abstract

The biotechnology of spatial omics has advanced rapidly over the past few years, with enhancements in both throughput and resolution. However, existing annotation pipelines in spatial omics predominantly rely on clustering methods and lack the flexibility to integrate extensive annotated information from single-cell RNA sequencing (scRNA-seq) due to discrepancies in spatial resolutions, species, or modalities. Here we introduce the CAESAR suite, an open-source software package that provides image-based spatial co-embedding of locations and genomic features. It uniquely transfers labels from scRNA-seq reference data, enabling the annotation of spatial omics datasets across different technologies, resolutions, species, and modalities, based on the conserved relationship between signature genes and cells/locations at an appropriate level of granularity. Notably, CAESAR enriches for location-level pathways, allowing for the detection of gradual biological pathway activation within spatially defined domain types. We demonstrate the advantages of CAESAR through a comprehensive analysis of five spatial omics datasets encompassing diverse technologies, resolutions, and modalities. Across these applications, CAESAR achieved substantial improvements in annotation accuracy (45.45%-4333.33%) by transferring cell-type labels from either multiple reference data, or across different species and modalities. As a result, CAESAR effectively recovers intricate structures in mouse olfactory bulb and embryo, and unveils tumor microenvironment heterogeneity, with exceptional efficiency and flexibility.

Introduction

Spatial omics is accomplished via a set of breakthrough technologies that enable the spatial profiling of molecular parameters, including gene and protein expression and chromatin structure. One of the techniques used, spatially resolved transcriptomics (SRT), requires a range of advanced technologies that enhance the throughput of expression profiling , from targeted to transcriptome-wide gene measurements, and improve the spatial resolution, from low resolution to subcellular resolution [1–3]. In parallel with the evolution of SRT technologies, other spatial omics technologies, such as spatial-ATAC-seq [4] and spatial-CITE-seq [5], have also seen rapid advancements. By mapping cell/domain types in a scalable manner, emerging spatial omics technologies offer unprecedented opportunities to characterize transcriptomic and cellular landscapes within a spatial context. Many spatial omics methods have been developed that incorporate routine analytical steps, such as the detection of spatially variable genes, dimensionality reduction, clustering, differential gene expression analysis, and gene set enrichment analysis [6–10]. However, most of these methods are "cluster-centric", predominantly relying on accurately defined clustering to identify meaningful gene features. This reliance becomes problematic when samples contain cells that are undergoing active state transitions, a phenomenon commonly observed in tumor or developmental datasets [11–14]. Moreover, a large number of single-cell RNA sequencing (scRNA-seq) datasets have been thoroughly characterized, providing abundant transcriptomic information with annotations for both human and mouse samples.

To annotate scRNA-seq datasets using these predefined references, the use of cluster-centric methods for cell annotation has been proposed. These methods typically either transfer cell-type labels from reference data to target data [15–17] or model marker-gene expression patterns in the target data [18–20]. The former strategy requires an additional batch-removal step, while the latter demands access to high-quality marker genes. Due to discrepancies between spatial omics and scRNA-seq data, annotating spatial omics data from diverse technologies, spatial resolutions, species, or modalities that leverage reference information from predefined scRNA-seq datasets is challenging. To fully harness the potential of these emerging technologies and drive breakthrough discoveries in molecular biology, co-embedding has emerged as a promising approach to overcome the limitations of clustering-centric pipelines [21–23]. Existing co-embedding methods based on multiple correspondence analysis (MCA) [21, 23] or multi-relation graph models [22] are employed for various tasks, including signature gene detection, pathway enrichment analysis, and multimodality co-embedding. However, as cluster-agnostic methods, they often fail to fully incorporate spatial information or histology images during co-embedding, leading them to potentially overlook valuable information. Moreover, these methods are limited in their ability to use labels from rich reference datasets to annotate spatial omics datasets across different technologies, resolutions, species, and modalities.

To overcome these limitations, we have designed the CAESAR suite, a unified and versatile software package that offers a general spatial co-embedding framework based on a feature-weighted scheme that leverages both spatial information and histology images. By assuming a conserved relationship between genomic features and cells/locations within each cell/domain type at an appropriate level of granularity, the CAESAR suite flexibly annotates spatial omics datasets across technologies, resolutions, species, and modalities by transferring cell-type

labels from predefined scRNA-seq reference data in a cluster-agnostic manner, and detects cell/domain-type-specific signature genes. Moreover, the CAESAR suite includes functions for hypothesis testing to identify pathways enriched in each cell/location or cell/domain type. We illustrate the benefits of using the CAESAR suite through extensive simulations and analyses of a diverse range of example datasets collected using various spatial omics technologies, species, and resolutions: 10x Xenium datasets of four human breast cancer (BC) sections, 10x Visium datasets of four human hepatocellular carcinoma (HCC) sections, Pixel-seq and ST datasets of the mouse olfactory bulb (MOB), and a spatial ATAC-seq dataset of a mouse embryo.

Results

Overview of CAESAR

The CAESAR suite is a novel open-source software package that co-embeds spatial locations and gene features into a unified low-dimensional space, utilizing both histology images and spatial coordinates. Within this space, the relative distance between locations and gene features can be used to characterize transcriptomic specificity, enabling a range of downstream analytical tasks (Fig. 1 and Methods). When cell types/domains are predefined, as in labeled reference datasets, the CAESAR suite detects cell- or domain-type-specific signature genes by evaluating the relative distances between the cells/locations and gene features. In scenarios where reference data originate from multiple batches or sections, heterogeneous batch effects can significantly distort expression patterns, complicating data integration. However, within a single batch or section, the relationships between cells/locations and genomic features remain conserved with respect to the cell or domain types, with batch effects merely introducing systematic noise. Leveraging these conserved relationships, the CAESAR suite exhibits remarkable flexibility in detecting signature genes, annotating cells or locations through knowledge transfer via labeled reference data and seamlessly integrating multiple reference and target datasets. As a proof of concept, we demonstrate that the CAESAR suite is capable of performing spatial annotations, with confidence level assessed via a permutation test, using knowledge transferred from scRNA-seq or SRT reference data to spatial omics data derived from diverse technologies, species, resolutions, and modalities. By analyzing the distances between spatial locations and sets of genes, such as pathway genes, the CAESAR suite detects gradual changes in pathway activation across different spatial domains. This is achieved through permutation and Wilcoxon tests, providing spot-level and cell- or domain-type-level enrichment significance, respectively.

Validation using CosMx data

We conducted comprehensive simulations using a CosMx dataset for lung cancer [24] and rigorously evaluated the performance of the CAESAR suite by comparing it with Cell-ID in terms of dimensionality reduction, signature gene detection, and annotation accuracy. The evaluation metrics included average silhouette width (ASW) [25], signature score (SigScore; see Methods) and classification accuracy (ACC) [26]. Ideally, when given a set of genes specific to a particular cell type, the optimal method should co-embed these genes in close proximity to the corresponding cells. To quantify this specificity, we introduced the SigScore, which attains

a value of 1 when all cell-type-specific genes are top-ranked by their average distance to cells of the corresponding type.

In Scenario 1, we used all fields of view (FOVs) from section Lung5 rep1 as the reference dataset and FOVs from three other sections (Lung5 rep2, Lung13, and Lung12) as target datasets, representing varying levels of heterogeneity between reference and target datasets (Supplementary Fig. S1). In this scenario, Lung5 rep1 and Lung5 rep2, derived from two consecutive sections from the same donor, exhibited a high degree of similarity. In Scenario 2, we binned 5 cells per location in the target datasets while maintaining the same reference data, to evaluate the performance of the CAESAR suite when confronted with a low-resolution target dataset. Conversely, in Scenario 3, the target data remained consistent with Scenario 1, and we binned 5 cells per location in the reference datasets to assess the CAESAR suite's performance using a low-resolution reference dataset. The details of these simulations are provided in the "Methods" section.

We first evaluated the CAESAR suite's ability to generate informative embeddings compared to the MCA employed by Cell-ID and GSDensity in co-embeddings, using ASW as the metric (Fig. 1 c, top panel). The CAESAR suite yielded higher ASW values in the estimated image-based spatial embedding of locations, indicating that its (co)-embeddings better preserved the biological differentiation. We then assessed the performance of both the CAESAR suite and Cell-ID in signature gene detection (Fig. 1 c, middle panel). The CAESAR suite exhibited a higher SigScore, indicating its superior capability in detecting signature genes. While the ACC of all the methods declined with increasing heterogeneity, the CAESAR suite consistently outperformed Cell-ID (Fig. 1 c, bottom panel, and Supplementary Fig. S2). Notably, iCAESAR, which integrates information from multiple reference datasets, demonstrated the most stable performance with minimal variation in its ACC, highlighting the advantage of utilizing multiple references. Furthermore, use of the CAESAR suite resulted in a substantially smaller proportion of unassigned cells than Cell-ID, with iCAESAR providing an even further reduced proportion (Supplementary Fig. S3), indicating the enhanced cell-type detection performance of iCAESAR.

Subsequently, we evaluated the performance of the CAESAR suite for pathway detection in comparison to other methods. In our simulations, pathway gene sets were generated using differentially expressed genes specific to cell types, and the area under the curve (AUC) was used to evaluate the performance in pathway recovery across various pathway scores. As illustrated in Fig. 1d, the CAESAR suite demonstrated superior performance in pathway detection, consistently surpassing Cell-ID and GSDensity in terms of AUC values.

CAESAR suite facilitates spatial annotations using multiple scRNA-seq reference datasets

We applied the CAESAR suite and other methods to analyze five published spatial omics datasets from different sequencing platforms: 10x Xenium, 10x Visium, ST, Pixel-seq, and Spatial ATAC-seq. For spatial annotation, we leveraged scRNA-seq reference data and transferred the labels to spatial omics datasets derived from diverse technologies, species, resolutions, and modalities. Upon annotating the target spatial omics data, the CAESAR suite was used to detect cell- or domain-type-specific signature genes and perform hypothesis testing for the detection of pathways enriched within each cell or location and/or cell or domain type.

To harness the reference scRNA-seq data from 26 human BC patients [27], we first analyzed human BC data generated using 10x Xenium [28] comprising four sections from two BC patients, with two serial replicates for each patient (Supplementary Fig. S4). We observed striking batch effects among the 26 scRNA-seq reference data in UMAP (Fig. 2a) with substantial heterogeneity exhibited in the annotated cell-type proportions, especially for cancer epithelial cells (Fig. 2b). Using the CAESAR suite that integrates all 26 reference datasets (named iCAESAR), we sequentially (a) detected cell-type-specific signature genes in each of the 26 reference datasets, (b) aggregated a signature gene list from the 26 reference datasets by weighting their occurrence across the references, (c) estimated spatial co-embeddings with histology images for the target BC sections, and (d) performed spatial annotations based on the average distance between each location and signature genes identified in step (b), with the entire annotation process performed as shown in Fig. 2c. The resulting co-embedding of cells/locations and the top-ranked signature genes revealed conserved relationships across both the reference and target datasets (Fig. 2c and Supplementary Fig. S5). In the reference data, we detected *CD3E* (in 20 samples, including Samples 1 and 25) and *CD3D* (in 18 samples, including Sample 1) as signature genes for T cells, among others (Supplementary Data 1). *CD3E* functions as a subunit of the T-cell receptor complex, playing a crucial role in CAR-T cell therapy [29], while *CD3D* has been implicated to participate in lymphocyte infiltration and immune checkpoint regulation, and serves as a prognostic biomarker for BC [30]. These signature genes were aggregated into a gene list used for annotating the target BC dataset by iCAESAR (Supplementary Fig. S6). By removing unwanted variations, we visualized the expression patterns of the top five signature genes for each annotated cell type across all four sections and observed the distinct signature profiles for each cell type (Fig. 2d; see Methods). Notably, many of these genes were reported to be differentially expressed across various cell types, i.e., *MS4A1* and *BANK1* in B cells [31, 32]; *CD3E*, *IL7R*, *CD3D*, and *CD247* in T cells [29, 33–35]; and *LYPD3*, *FASN*, and *FOXA1* in cancer epithelial cells [36–38], while the roles of *MLPH* and *SERHL2*, specifically detected in cancer epithelial cells, remain underexplored in BC.

To evaluate the performance of CAESAR in spatial annotation, we applied CAESAR and Cell-ID to each of the 26 references, and iCAESAR to all 26 references to annotate the BC dataset, and generated spatial heatmaps illustrating the cell-type assignments, as shown in Fig. 2e and Supplementary Fig. S7-S8. The majority of the CAESAR annotation results demonstrated high confidence levels (Supplementary Fig. S9). iCAESAR precisely detected cancer epithelial cells and other immune-relevant cell types, while Cell-ID labeled most cells as cancer epithelial in BC sections 1 and 2, with a higher proportion of normal epithelial cells in sections 3 and 4. Notably, the iCAESAR results exhibited a significantly lower proportion of unassigned cells than those of Cell-ID, with an average of 1.22% unassigned cells compared to Cell-ID's 95.04% across the four sections, indicating its enhanced capability in cell type detection (Supplementary Fig. S10). Using all 26 reference datasets, iCAESAR demonstrated further improved stability compared to the use of each reference individually with CAESAR, although both showed substantial improvements in annotation accuracy over Cell-ID, with mean (standard deviation) ACC values of 0.819 (0.055), 0.665 (0.186), and 0.015 (0.066), respectively (Fig. 2f, upper panel). While CAESAR/iCAESAR demonstrated superior performance over Cell-ID in its ability to generate co-embeddings to distinguish among distinct cell types, with

mean ASW scores of 0.115 and 0.042, respectively.

Next, we examined the pathways enriched within the BC dataset. First, we detected significantly enriched pathways within the categories of GO biological process (GOBP), KEGG, Reactome, chemical and genetic perturbations (CGP), and cancer modules (CM) using a graph-based test (see Methods), with 393, 19, 38, 327, and 69 pathways detected, respectively, under an adjusted p-value of less than 0.05. Subsequently, we applied CAESAR to detect differentially enriched pathways among annotated cell types, summarizing the top five most significantly enriched pathways for each cell type using a dot plot (Fig. 2g). Among these, cancer-related module 139 and Doane breast cancer classes up were enriched in cancer epithelial cells, while vasculature development was enriched in perivascular-like cells (PVLs), endothelial and cancer-associated fibroblasts (CAFs). To further examine the enrichment of pathways in each location, we applied CAESAR to perform spot-level enrichment analysis. CAESAR exhibited superior performance to Cell-ID in pathway activity scoring, with mean SigScore values of 0.898 and 0.624, respectively (Fig. 2f, bottom panel). We summarized the cell-type-specific pathway activation data across each section using a spatial heatmap (Fig. 2h and Supplementary Fig. S11-14), which highlighted that vasculature development was highly enriched at the boundary of cancer epithelial cells, while Doane breast cancer classes up was predominantly enriched in cancer epithelial cells. Further enrichment analysis revealed that the cell types from each section were highly enriched in several common pathways, suggesting that the annotations provided by the CAESAR suite were well-aligned across sections (Supplementary Fig. S15-16).

CAESAR suite enables spatial annotations of human HCC data transferred from scRNA-seq in mouse HCC

Next, we applied the CAESAR suite and Cell-ID to analyze four sections of human HCC data obtained from 10X Visium [39]. The dataset comprised two tumor sections (HCC1 and HCC2) and two tumor-adjacent tissue sections (HCC3 and HCC4) from an HCC patient (Supplementary Fig. S17a). To demonstrate the robustness of the CAESAR suite using reference data across species, we performed annotations of the four target HCC sections using either human [40] or mouse [41] scRNA-seq data as references (Fig. 3a; see Methods). Taking manual annotations as the ground truth (Fig. 3b and Supplementary Fig. S17a), the spatial heatmaps generated by CAESAR, using either human (Fig. 3c and Supplementary Fig. S17b) or mouse reference data (Fig. 3d and Supplementary Fig. S17c), exhibited marked improvements over those generated by Cell-ID, which showed a substantial proportion of unassigned cells. Notably, the annotations CAESAR made using mouse reference data closely aligned with those obtained using the human reference data, achieving mean accuracies of 0.702 and 0.669, respectively (Fig. 3e). Compared to those made by Cell-ID, CAESAR achieved a substantial gain in accuracy, 495.5% and 677.7%, respectively. A detailed examination of the annotations based on the human and mouse references revealed that HPC-like cells, an annotation absent from the mouse data, were detected as HCC cells using the mouse reference (Supplementary Fig. S18). HPC-like cells are known to exhibit similarities to HCC cells and contribute to HCC formation through their activation [42, 43]. An analysis of annotation confidence further demonstrated consistent species-agnostic results (Supplementary Fig. S19).

Using CAESAR with a single mouse HCC reference dataset, we first (a) detected cell-type-

specific signature genes within the reference data, (b) estimated spatial co-embedding in the target HCC sections, and (c) performed spatial annotations based on homologous genes of human and mouse. The resulting visualization of co-embeddings for the cells/locations and the top signature genes revealed conserved relationships across both the reference and target datasets (Fig. 3f). For example, the genes *Rnf128* and *Acox2*, which are unique to HCC cells in mice, were detected as signature genes in human HCC sections. Similarly, *Mmp23* and *Tpm2*, associated with fibroblasts, were also detected in the human HCC sections. After removing unwanted variations, we visualized the expression levels of the top six signature genes for each cell type across all four sections (Fig. 3g; see Methods). Several of these genes have been reported to be enriched in specific HCC cell types, such as *IGLC2* in B/Plasma cells [44] and *RNF128* and *ABCB11* in HCC cells [45, 46]. Notably, *RNF128* promotes HCC progression through the activation of the EGFR/MEK/ERK signaling pathway [45] while *ABCB11* is associated with a patient's susceptibility to HCC development [46].

We further applied CAESAR to an enrichment analysis, identifying 1,303, 61, 253, 1,312, 194, and 2,213 significantly enriched pathways in categories in the GOBP, KEGG, Reactome, CGP, CM, and immune signatures database (ImmuneSigDB), respectively, all with an adjusted p-value of less than 0.05. Subsequent analysis revealed significant differences in these pathways at the cell/domain-type level (Fig. 3h). Pathways predominantly enriched in HCC cells contained liver cancer subtypes and survival and proliferation mechanisms, such as the Reactome pathway involving *SREBF* and *SREBP*. The high expression of *SREBP-1* in tumors has been linked to improved 3-year overall and disease-free survival rates in HCC patients, and thus *SREBP-1* potentially promotes tumor progression by enhancing cell growth and metastasis [47–49]. Pathways enriched in stroma/immune cells are involved in the regulation of immune responses, cell signaling, protein interactions, and vasculature development. The spatial heatmaps of these differentially enriched pathways (Fig. 3i and Supplementary Fig. S20) indicated that the vasculature development pathway was prominently activated at the boundaries of HCC cells but not within HCC regions. This suggests that the role of vasculature development at the tumor periphery and within the tumor microenvironment may be consistent across various cancer types [50, 51].

CAESAR suite accurately recovers MOB layers in SRT datasets with low or high resolution

To demonstrate the ability to annotate cell/domain types in SRT data with varying resolutions, we applied the CAESAR suite to an analyses of MOB datasets from the ST or Pixel-seq platform. ST represents an earlier SRT technology with a 100-μm diameter resolution, while Pixel-seq is a more recent technology offering near-single-cell resolution.

We first applied CAESAR and Cell-ID to annotate low-resolution ST MOB dataset using scRNA-seq reference data with coarse-grained labels for five layers: granule cell (GC), mitral and tufted cell (M/TC), Olfactory sensory neurons (OSNs), periglomerular cell (PGC), and external plexiform layer interneuron (EPL-IN) [52]. Compared to the manual annotations (Fig. 4a, left panel), CAESAR demonstrated superior performance in accurately reconstructing the MOB layer structure (Fig. 4a, middle panel), with a heatmap of confusion matrix indicates a strong alignment between the manual annotations and CAESAR predictions (Fig. 4b), whereas

Cell-ID struggled to capture the MOB architecture, resulting in a considerable number of unassigned cells (Fig. 4a, right panel). Further analysis led to a visualization of the conserved relationships between locations and genes across both datasets (Supplementary Fig. S21). For the low-resolution dataset from ST, we applied CAESAR to estimate the cell-mixing proportions for each location (Supplementary Fig. S22-S23), and distinct cell-type distributions across different domains were revealed. For example, Domain GC was predominantly composed of GC, immature neurons, and transitional neurons, while Domain OSNs was primarily occupied by OSNs.

Next, we applied CAESAR and Cell-ID to annotate the high-resolution Pixel-seq MOB dataset using scRNA-seq reference data with fine-grained labels [53]. Compared to the original annotations of the target dataset, CAESAR achieved 45.45% higher annotation accuracy than Cell-ID (Fig. 4c), and the spatial heatmaps reflect the fine structural consistencies with the delineations in the spatial heatmap of the logarithm of unique molecular identifiers (UMIs) (Fig. 4c, left panel). An enhanced visualization of expression with heatmaps for each cell type (Supplementary Figure S24) revealed the distinct spatial patterns of the cell types, particularly OSNs, mesenchymal (Mes), and PGC, consistent with the cell-type probability heatmaps.

In the annotation, we first visualized the co-embeddings of cells and the top two signature genes from the fine-grained MOB reference with the high-resolution Pixel-seq dataset in the UMAP plots (Fig. 4d). In the near-single-cell resolution target data, we observed a considerable overlap of signature genes for each cell type, such as *Dcn* and *Asgr1* for Mes, *Lrrtm1* and *Otop2* for M/TC, and *Penk* and *Icam5* for GC, indicating the preserved relationships between cells/locations and genes. Of note, *Icam5*-knockout mice have been shown to experience experimental autoimmune encephalomyelitis in the chronic phase, highlighting *Icam5*'s neuroprotective role in progressive neurodegeneration [54]. We visualized the expression patterns of the top five signature genes for each annotated cell type, and observed distinct cell-type expression patterns (Supplementary Fig. S25). The spatial distributions of the expression of the top signature genes (Fig. 4e, upper panel) closely aligned with the annotated cell types (Fig. 4e, middle panel).

Finally, we applied CAESAR to detect differentially enriched pathways in the GO database between annotated cell types, with enrichment scores visualized in spatial heatmaps (Fig. 4e, bottom panel, and Supplementary Fig. S26). We found that the neuron neurotransmitter transport was enriched in neural M/TCs, indicating its crucial role in supporting M/TCs' neurotransmission and olfactory signal modulation. The top differentially enriched pathways for each cell type, which are presented in Fig. 4f, indicated Mes cells were significantly enriched in activated transmembrane transporter activity, particularly ion transporter activity, mirroring the mechanism in the nervous system by which neurons use ion transmembrane transport to generate action potentials for information transmission [55, 56]. CAESAR outperformed Cell-ID and GSDensity in scoring the cell-type-specific pathway activity, achieving a median AUC of 0.762, compared to 0.707 for Cell-ID and 0.504 for GSDensity, as illustrated in Fig. 4g.

Annotations of spatial ATAC-seq data using scRNA-seq reference with CAESAR suite

Using the CAESAR suite, we conducted a more challenging cross-modality task invoving an analysis of spatial ATAC-seq data from a mouse embryo, characterized by high sparsity and high noise (Supplementary Fig. S27). The E11 mouse embryo data utilized contained a median of 36,303 unique fragments per 50-μm spot, with a total of 2,162 spots [4]. This spatial ATAC-seq dataset was annotated using scRNA-seq mouse embryo reference data from the Mouse Organogenesis Cell Atlas (MOCA) [57], with annotations derived via Louvain clustering.

The two-dimensional UMAP projections of co-embeddings were made for the cells/spots to illustrate the overlap among the top signature genes across both reference and target datasets (Fig. 5a). We then visualized CAESAR's annotations of the spatial coordinates and compared them with those from Cell-ID (Fig. 5b). CAESAR (ACC = 0.253) significantly outperformed Cell-ID (ACC = 0.090), accurately recovering excitatory neurons, stroma cells, and a primitive erythroid lineage, while using Cell-ID resulted in a substantial proportion of unassigned locations.

Next, we performed spot-level pathway enrichment analysis using the CAESAR suite to detect differential pathways among cell types. The top five differentially enriched pathways were visualized in a dot plot (Fig. 5c). Notably, the chloride transmembrane transport pathway within the GOBP database was highly enriched in both excitatory neurons and postmitotic premature neurons. This pathway is essential for neuronal functionality and excitability, particularly within excitatory neurons [58]. The epithelial-to-mesenchymal transition (EMT) involved in endocardial cushion formation was prominently enriched in stroma cells. This finding aligns with the role of EMT in cardiac development, during which transformed cells function as stromal components critical for the formation of cardiac valves and septa [59]. Additionally, the gas transport pathway was significantly enriched in the primitive erythroid lineage. This lineage represents a pivotal stage in erythropoiesis during embryogenesis, when progenitor cells mature into erythroid precursors, eventually developing into mature red blood cells essential for gas transport during metabolic processes [60]. We further depicted the enrichment scores of the domain-specific pathways, including those associated with excitatory neurons, stromal cells, and the primitive erythroid lineage, in heatmaps (Fig. 5d and Supplementary Fig. S28). Our analysis revealed the progressive activation pattern of pathways within their respective domains. For instance, the enrichment score for EMT involved in endocardial cushion formation exhibited a continuous decline from stromal cells to adjacent domains, particularly those with excitatory neurons (Fig. 5d, middle panel). This observation highlights the exceptional capability of CAESAR to derive spot-level enrichments and offer profound insights into the intricate dynamics of biological pathways.

Discussion

We aimed, via this study, to introduce and demonstrate the CAESAR suite, a novel spatial co-embedding framework that offers a fully integrated and cluster-agnostic suite of tools. This framework is designed to detect cell- or domain-type-specific signature genes, perform spatial annotations of cell or domain types, and facilitate hypothesis testing to uncover pathways

enriched in each cell or location as well as within specific cell or domain type.

In contrast to traditional co-embedding methods based on MCA, such as Cell-ID and GSDensity, the CAESAR suite provides more sophisticated, image- and spatial-aware co-embedding of genomic features and cells/locations by effectively considering both histology image information and spatial coordinates. Moreover, the co-embedding framework in the CAESAR suite is compatible with any dimensionality reduction technique that employs a feature-weighted scheme. When cell/domain-type labels are known, such as in labeled reference data, the CAESAR suite excels in detecting cell/domain-type signature genes by assessing the relative distance between cells/locations and gene features.

Assuming a conserved relationship between genomic features and cells/locations at an appropriate level of granularity, the CAESAR suite, to the best of our knowledge, is the first to enable spatial annotations by transferring cell-type labels from predefined scRNA-seq references to target spatial omics datasets across a wide range of technologies, resolutions, species, and modalities. When multiple references are available, the CAESAR suite also accounts for the uncertainty in detecting cell/domain-type signature genes across multiple batches, thereby mitigating the impact of batch effects, which often problematic in cluster-centric analysis. Our examination of five spatial omics datasets, encompassing diverse technology, resolution, and modalities, i.e., 10x Xenium, 10x Visium, ST, Pixel-seq, and Spatial ATAC-seq, demonstrated CAESAR suite's robust spatial annotation capabilities. Using reference data from 26 batches, we demonstrated the CAESAR suite's capacity to effectively annotate a Xenium dataset of four human BC sections, in which it achieved substantial improvements in accuracy (4333.33%) and SigScore (42.9%) compared to Cell-ID. Similarly, when used to annotate a Visium dataset of four human HCC sections using scRNA-seq reference data from either human or mouse, the CAESAR suite achieved comparable annotation accuracies, with remarkable accuracy gains of 495.5% and 677.7%, respectively, compared to Cell-ID.

The CAESAR suite also offers functions for pathway enrichment analysis at both the location and cell/domain-type levels, enabling the delineation of pathway activation across different domain types. For example, in the Xenium dataset for BC, the CAESAR suite detected the activation of the vasculature development pathway, which was highly active at the boundary of cancer epithelial cells, minimally active in cancer cells, and dormant in non-cancer cells.This finding highlights the critical role of vascular networks in tumor growth and metastasis, where the newly formed vasculature surrounding cancer cells serves not only to sustain tumor survival and expansion but also as a conduit for metastatic tumor cells [61].

As a proof-of-concept, the CAESAR suite provides opportunities for new exciting research routes. Firstly, when sections of spatial omics are from multiple conditions, functions that can be used to perform hypothesis testing between conditions at both gene and pathway levels are needed. Secondly, when datasets with multi-modality measurement on the same section (paired) are available, functions for co-embedding paired datasets are needed.

As a proof-of-concept, the CAESAR suite provides opportunities for new exciting research routes. Firstly, when sections of spatial omics are from multiple conditions, functions that can be used to perform hypothesis testing between conditions at both gene and pathway levels are needed. Secondly, when datasets with multi-modality measurement on the same section (paired) are available, functions for co-embedding paired datasets are needed.

Methods

CAESAR suite overview

The CAESAR suite is an open-source software package comprising diverse functional modules that facilitate the co-embedding of locations and gene features, signature gene detection, spatial annotations through the integration of multiple reference datasets and for multiple SRT target datasets, and pathway enrichment analysis at both the spot-level and cell/domain type-level, as illustrated in Fig. 1a.

Different from existing co-embedding methods, CAESAR model uses a combination of a latent factor model and a feature-weighted scheme to project locations and features onto the same Euclidean space. Specifically, we denote $\mathbf{X} = (x_{sg}) \in \mathbb{R}^{S\times G}$ as the log-normalized gene expression matrix, $\mathcal{L} = (l_s) \in \mathbb{R}^{S\times 2}$ as the spatial coordinate matrix, $\mathbf{E} = (\mathbf{e}_s) \in \mathbb{R}^{S\times d}$ as the feature matrix from histology images extracted by Visual transformer (see Supplementary Notes) and $\mathbf{H} = (\mathbf{h}_s) \in \mathbb{R}^{S\times d}$ as low-dimensional embeddings of locations, where S is the number of spots, G is the number of genes and d is the dimension of image features. We relate gene expression (x_{sg}) to low-dimensional embeddings $(\mathbf{h}_s)$ using a linear factor model:

$$x_{sg} = \mu_g + \mathbf{b}_g^{\mathrm{T}}\mathbf{h}_s + u_{sg}, \tag{1}$$

and relate the low-dimensional embeddings $(\mathbf{h}_s)$ to the spatial coordinates l_s and histology image features $\mathbf{e}_s$ via an intrinsic conditional autoregressive model:

$$\mathbf{h}_s = \sum_{s'\in N_{l_s}} w(\mathbf{e}_s, \mathbf{e}_{s'})\mathbf{h}_{s'} + \boldsymbol{\varepsilon}_s, \tag{2}$$

where $\mathbf{u}_s = (u_{s1}, \cdots, u_{sG})^{\mathrm{T}} \sim N(\mathbf{0}, \Lambda)$ with $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_G)$, N_{l_s} is the neighboring spot set of spot s defined by coordinates, $w(\mathbf{e}_s, \mathbf{e}_{s'}) = \frac{\tilde{w}_{ss'}}{w_{s+}}$ with $\tilde{w}_{ss'} = \exp\{-d^2(\mathbf{e}_s, \mathbf{e}_{s'})/\sigma\}$ and $w_{s+} = \sum_{s'\in N_{l_s}} \tilde{w}_{ss'}$, and $\boldsymbol{\varepsilon}_s \sim N(\mathbf{0}, w_{s+}^{-1}\Phi)$. We designed a variational EM algorithm to infer $\mathbf{h}_s$ using its posterior estimate (see Supplementary Notes). Next, we utilize a gene expression-weighted scheme to derive the embeddings of genes, as formulated below:

$$\mathbf{v}_g = \frac{\sum_{s=1}^{S} w_{sg}x_{sg}\mathbf{h}_s}{\sum_{s=1}^{S} w_{sg}x_{sg}},$$

where $w_{sg} = (1 + \sum_{s'\in N_{l_s}} I(x_{sg} \neq 0))/(1 + n_{l_s})$, $n_{l_s} = |N_{l_s}|$ is the number of neighboring spots of spot s, and $I(x_{sg} \neq 0)$ is an indicator function that equals 1 if $x_{sg} \neq 0$ and 0 otherwise. By accounting for the gene expression ratio in the cell's local microenvironment, the resulting gene embedding focuses more on gene expression-intensive areas. It is important to note that $\mathbf{v}_g$ represents a weighted average of the embeddings of locations, $\mathbf{h}_s$, and thus resides in the same Euclidean space spanned by $\{\mathbf{h}_s, s = 1, \cdots, S\}$. Consequently, computing the distance between $\mathbf{v}_g$ and any $\mathbf{h}_s$ is semantically meaningful. Let $\mathbb{S}_k$ denote the index set of spots corresponding to cell/domain type k. For clarity, suppose that gene g is exclusively expressed in cell/domain type k, implying $x_{sg} = 0$ for $s \notin \mathbb{S}_k$. In this scenario, the embedding of gene g simplifies to:

$$\mathbf{v}_g = \frac{\sum_{s\in\mathbb{S}_k} w_{sg}x_{sg}\mathbf{h}_s}{\sum_{s\in\mathbb{S}_k} w_{sg}x_{sg}}.$$

As a result, the embedding of gene g will closely align with the embeddings of spots belonging to cell/domain type k. Furthermore, for scRNA-seq data, we devise a non-centered linear factor model to jointly embed cells and genes into a shared space (see Supplementary Notes for details).

Signature gene detection

The Euclidean distance $d(\mathbf{v}_g, \mathbf{h}_s)$ captures the degree of specificity between gene g and location s, as the embedding of gene g resides at the weighted centroid of the embeddings of cells that express this gene. Consequently, the specificity of gene g to a particular cell/domain type k is quantified by the mean distance between gene g and the cells belonging to cell/domain type k. This is formally expressed as $\frac{1}{|\mathbb{S}_k|}\sum_{s\in\mathbb{S}_k} d(\mathbf{v}_g, \mathbf{h}_s)$, where $\mathbb{S}_k$ denotes the set of cells constituting the cell/domain type k. After excluding genes with an expression ratio below η_r (set at 0.1 by default) to mitigate the inclusion of infrequently expressed genes and diminish the influence of random noise, the signature genes for cell/domain type k, denoted as $\Gamma_k(\gamma)$, are identified as the top γ genes that exhibit the highest level of specificity. This is accomplished by ranking genes based on their average distance from cells belonging to type k, as given by the formula:

$$\Gamma_k(\gamma) = \left\{ g \mid \mathrm{rank}_g \left(\frac{1}{|\mathbb{S}_k|} \sum_{s\in\mathbb{S}_k} d(\mathbf{v}_g, \mathbf{h}_s) \right) \leq \gamma \right\}.$$

Here, rank_g represents the ranking function that assigns a position to each gene g based on its calculated average distance, with lower distance indicating higher specificity.

Spatial annotation

We first extract the signature gene sets for each cell/domain type from each of R reference datasets, denoted as $\mathbb{L}_r = \{\Gamma_{rk}(\gamma) : k = 1, \cdots, K\}, r = 1, \cdots, R$, where K is the total number of cell/domain types, R signifies the number of available reference datasets and Γ_{rk} is set to empty set when cell/domain type k is absent from the r-th reference dataset. The parameter γ_r represents the number of signature genes chosen for each set Γ_{rk}, which is determined as the maximum value that maintains the overlap of signature genes across $\{\Gamma_{rk}, k = 1, \cdots, K\}$ below a specified threshold t. This threshold is established to regulate the extent of shared signature genes among different cell/domain types. Formally, γ_r is given by:

$$\gamma_r = \max_{\gamma} \{\gamma \mid \forall 1 \leq k_1 < k_2 \leq K, |\Gamma_{rk_1}(\gamma) \cap \Gamma_{rk_2}(\gamma)| \leq t\}.$$

Here, the expression $|\Gamma_{rk_1}(\gamma) \cap \Gamma_{rk_2}(\gamma)|$ calculates the number of genes common to both $\Gamma_{rk_1}(\gamma)$ and $\Gamma_{rk_2}(\gamma)$, ensuring that the intersection does not exceed the threshold t for any pair of cell/domain types k_1 and k_2. The default setting for the threshold t is 1, but it can be adjusted upwards when fine-grained labels are available. By aggregating the signature gene sets derived from the various reference datasets, we can obtain a comprehensive signature gene set for each cell/domain type k, denoted as $\Gamma_k = \bigcup_{r=1}^{R} \Gamma_{rk}$. Additionally, we assign weights to each gene g within Γ_k, denoted as w_{gk}, which are calculated as the proportion of references in which the gene appears as a signature gene for that cell type. Specifically, $w_{gk} = \frac{\tilde{w}_{gk}}{\sum_{g\in\Gamma_k} \tilde{w}_{gk}}$,

where $\tilde{w}_{gk} = \sum_{r=1}^{R} I(g \in \Gamma_{rk})$. This frequency-based weighting approach effectively emphasizes the robust associations between genes and cell/domain types, while mitigating the potential influence of low-quality signature genes that may arise due to data variability or random effects on subsequent annotations.

Subsequently, we compute the Euclidean distances between each location s and gene g in the target data, represented as $d(\mathbf{v}_g, \mathbf{h}_s)$. Here, $\mathbf{v}_g$ and $\mathbf{h}_s$ represent the co-embeddings of gene g and location s, respectively, which are obtained through the spatial co-embedding module within the CAESAR suite. To assess the specificity of a given location s to a particular cell/domain type k, we calculate the weighted average distance between location s and the genes in the signature gene set Γ_k from the reference data. This is expressed as:

$$d(\mathbf{h}_s, \Gamma_k) = \sum_{g \in \Gamma_k} w_{gk} d(\mathbf{h}_s, \mathbf{v}_g).$$

The probability of assigning the label y_s of location s to a specific cell/domain type k is then approximated using a standard normal cumulative distribution function $\Phi(\cdot)$, adjusted for the mean μ_s and standard deviation σ_s:

$$\text{Prob}(y_s = k) = \Phi\left(\frac{d(\mathbf{h}_s, \Gamma_k) - \mu_s}{\sigma_s}\right),$$

where μ_s and σ_s are the mean and the standard deviation of $\{d(\mathbf{h}_s, \Gamma_k), k = 1, \cdots, K\}$. Then, CAESAR suite annotates location s as cell/domain type k with highest probability. For the low-resolution target dataset, the cell mixing proportion of cell/domain type k in location s is obtained by normalizing the above probability, denoted as $\pi_{sk} = \frac{\text{Prob}(y_s=k)}{\sum_{k=1}^{K} \text{Prob}(y_s=k)}$.

The CAESAR suite is unique in its ability to offer confidence levels for annotation results, a vital feature for evaluating the trustworthiness and precision of cell annotations. This enables researchers to base their conclusions on a solid foundation, as they are informed of the quality of the data. Specifically, we commence by generating K control gene sets via random sampling, ensuring that each control set Γ_k^ℓ mirrors the size and gene weights of its corresponding signature gene set Γ_k. Subsequently, we identify the minimal average distance of these control sets to a given spot s, denoted as $\min_{\ell \in \{1,\cdots,K\}} d(\mathbf{h}_s, \Gamma_k^\ell)$. This procedure is repeated L times, and we calculate the confidence level for spot s being annotated as type k as follows:

$$\text{Confidence}(y_s = k) = \frac{1}{L} \sum_{\ell=1}^{L} I\left(\min_{\ell \in \{1,\cdots,K\}} d(\mathbf{h}_s, \Gamma_k) < \min_{\ell \in \{1,\cdots,K\}} d(\mathbf{h}_s, \Gamma_k^\ell)\right).$$

Spots with a confidence level falling below a predefined threshold η_c are designated as "unassigned", with a default threshold of $\eta_c = 0.95$ employed in this study. Researchers have the flexibility to adjust this threshold based on their specific project requirements.

Pathway enrichment analysis at different levels

The CAESAR suite offers comprehensive pathway enrichment analysis at different levels without necessitating clustering. Given that the cell-gene distance serves as a proxy for their association, genes that are specific to a particular subpopulation of cells tend to cluster closely together in the co-embedding space. To assess the extent of enrichment of a pathway in the dataset,

we employ a robust graph-based test [62] to test the degree of clustering of the gene set (denoted as Γ) of this pathway, which is agnostic to the underlying graph structure and adept at handling high-dimensional data. Specifically, we first construct a 5-th minimum-spanning tree graph using all gene embeddings, in which each node i represents the gene g_i with its embedding as node feature. We define the edge weight of (i, j) as $w(i, j) = 1/\max\{d_i, d_j\}$, where d_i is the node degree of node i and $w(i, j) = 0$ if node i and node j are not connected. Let $R_p = \sum_{i,j\in\Gamma} w(i, j)$ be the total weights of edges connecting genes within the pathway, $R_{p^c} = \sum_{i,j\notin\Gamma} w(i, j)$ be the total weights of edges connecting genes outside the pathway. We further define $R_{\text{diff}} = R_p - R_{p^c}$ and $R_w = (1 - q)R_p + qR_{p^c}$, where $q = (n_p - 1)/(G - 2)$ and $n_p = |\Gamma|$. The robust edge-count test statistic is constructed as

$$T_n = \max\left\{\frac{R_w - \mathrm{E}(R_w)}{\sqrt{\mathrm{Var}(R_w)}}, \left|\frac{R_{\text{diff}} - \mathrm{E}(R_{\text{diff}})}{\sqrt{\mathrm{Var}(R_{\text{diff}})}}\right|\right\},$$

by comparing the observed values of R_w and R_{diff} to their expected values and variances, under the null hypothesis of no enrichment. The asymptotic distribution of T_n is used to obtain p-values, which are then adjusted for multiple comparisons using the Cauchy combination [63] when testing multiple sections simultaneously, and the Benjamini-Hochberg procedure for FDR control when testing multiple pathways simultaneously. This approach enables us to identify whether a pathway is highly and specifically expressed within some specific cell subpopulations. However, the specific identity of these subpopulations remains elusive at this juncture.

To uncover the subpopulation where the pathway is abundant, we undertake a spot-level enrichment analysis that not only reveals the enriched subpopulation but also tracks the gradual activation of the pathway across entire spots. Specifically, the CAESAR suite assesses the level of pathway activity at each location, quantifying the specificity of this pathway among various locations. It accomplishes this by generating L size-matched control gene sets through random sampling and subsequently calculating their average distance $d(\mathbf{h}_s, \Gamma^\ell)$ for a given spot s and each set Γ^ℓ, where ℓ ranges from 1 to L. Throughout this study, a default value of $L = 1000$ was employed. Ultimately, the pathway activity level is determined as the proportion of control gene sets whose average distance is greater than the distance between the co-embeddings of a given spot s and its true gene set Γ, given by $\frac{1}{L}\sum_{\ell=1}^{L} I\left(d(\mathbf{h}_s, \Gamma) < d(\mathbf{h}_s, \Gamma^\ell)\right)$. A higher pathway activity score signifies a more pronounced enrichment of the pathway, and the variations in activity levels across different locations indicate the existence of a gradual activation pattern for the tested pathway.

Furthermore, when detailed information about cell subpopulations is accessible (for instance, the cell/domain types annotated by CAESAR), the CAESAR suite can conduct enrichment analysis for a particular pathway at a cell/domain type-specific level. To achieve this, we employ a Wilcoxon test to ascertain whether the pathway activity level within a specific cell type surpasses that of other types. In scenarios where multiple sections are evaluated concurrently, the p-values are aggregated using the Cauchy combination method. This approach can also be extended to identify pathways that are unique to specific biological conditions by comparing their activity levels across varying conditions. As a result, the CAESAR suite is instrumental in identifying pivotal pathways potentially implicated in distinct cellular behaviors or disease states. This not only pinpoints potential therapeutic targets but also offers profound insights into cellular function and the underlying mechanisms of disease.

Unwanted-variation-removal for gene expression

To effectively visualize the expression patterns of signature genes by integrating multiple target sections, such as 10x Xenium BC sections and 10x Visium HCC sections, it is necessary to eliminate unwanted variation (i.e., batch effects) within the combined expression matrix. When multiple target datasets have been annotated using the CAESAR suite, these batch effects can be mitigated by leveraging a set of housekeeping genes as negative controls. These genes remain unaffected by other biological influences, allowing for the precise removal of unwanted variation [64]. In this study, mouse/human housekeeping gene sets obtained from the Housekeeping and Reference Transcript Atlas were employed [65]. First, we performed PCA of the gene expression matrices of housekeeping genes present in each target dataset t, obtaining the top ten principal components (PCs), $\hat{\mathbf{m}}_t$, which can be treated as the unwanted variation factors. The weighted average distance matrix $\hat{\mathbf{h}}_t \in \mathbb{R}^{S\times K}$, whose (s,k)-th element is $d(\mathbf{h}_s, \Gamma_k)$, reflects the specificity of location s to cell/domain types and is suitable for explaining biological variation between cell/domain types. Finally, we used a linear model to remove unwanted variation from the normalized gene expression matrix:

$$\mathbf{X}_t = \hat{\mathbf{h}}_t\boldsymbol{\alpha} + \hat{\mathbf{m}}_t\boldsymbol{\beta} + \boldsymbol{\epsilon_t}, \tag{3}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{K\times G}$ is the coefficient matrix for biological effects between cell/domain types and $\boldsymbol{\beta} \in \mathbb{R}^{10\times G}$ is the coefficient matrix for unwanted variations. After estimating the coefficients in Eqn. (3), unwanted variations can be removed from the original normalized gene expression matrix via

$$\hat{\mathbf{X}}_t = \mathbf{X}_t - \hat{\mathbf{m}}_t\hat{\boldsymbol{\beta}}.$$

A similar strategy can be used to remove unwanted biological conditions or other variations that the user wishes to eliminate by including such information in Equation (3).

Comparison of methods

We conducted extensive simulation studies and real data analyses to benchmark the CAESAR suite against Cell-ID, a tool implemented within the R package *CelliD* [21], focusing on annotation accuracy, dimension reduction capabilities, and co-embedding performance. Throughout these evaluations, both CAESAR and Cell-ID utilized an identical list of signature genes, derived via signature gene detection in reference data, as their input. Of note, Cell-ID was specifically designed for co-embedding scRNA-seq data while leveraging multiple correspondence analysis (MCA).

To assess the performance of the CAESAR suite in detecting pathway activity, we compared it against two competitors: Cell-ID and GSDensity [23]. GSDensity, implemented in the R package *gsdensity*, is a gene set scoring approach that leverages the MCA co-embedding generated by Cell-ID. During the implementation, we adhered to the default parameter settings outlined in the respective packages for both methods.

Evaluation metrics

We evaluated the methods' performances in annotation accuracy, dimension reduction, co-embedding performance, and pathway activity detection using the following metrics.

Classification accuracy. To assess annotation accuracy, we utilized classification accuracy (ACC) [26], the standard benchmark for evaluating classifier performance. We excluded spots with cell/domain types not present in the reference data from the ACC calculation. ACC is the ratio of correctly annotated spots to the total number of spots, with higher values indicating superior accuracy in predicting correct labels for the target data.

Average silhouette width. To evaluate the ability of the embeddings to distinguish between cell/domain types, we employed the average silhouette width (ASW) [25]. ASW ranges from -1 to 1, with higher scores indicating better preservation of biological signals.

Signature score. To assess co-embedding performance, we measured the specificity of agreement between cell-type-specific genes and cell types. For each cell type k, we identified its top 3 differentially expressed genes with the largest log-fold change as cell-type-specific genes. We then calculated the sum of ranks of these genes based on their average distance from spots of cell type k in descending order. The SigScore for cell type k was obtained by normalizing this sum using min-max normalization. An optimal tool would demonstrate high specificity for all cell types, reflected in a SigScore close to 1 for each cell type. The final SigScore is the weighted average of SigScores across all cell types, weighted by the proportion of spots per cell type.

Area under curve. We evaluated the effectiveness of spot-level pathway activity detection by assessing its capability to precisely identify correct cell types through the utilization of the pathway activity scores generated by the CAESAR suite and comparative methods. For each cell type k, we designated the gene set that included the top three differentially expressed genes with the greatest log-fold change as its fundamental enriched pathway. Using this pathway, we computed the pathway activity scores for all spots, employing both the CAESAR suite and comparative approaches. Next, we employed these activity scores to determine the area under the curve (AUC) for accurately distinguishing the correct cell type across various score thresholds. Specifically, we ordered the cells based on their pathway activity scores, resolving ties randomly, and calculated the recovery ratio at every feasible point. Consequently, for each cell-type-specific pathway, a superior method will achieve a higher AUC value. The final AUC was determined as the weighted average of the AUCs corresponding to all cell types in the dataset, where the weight is proportional to the ratio of spots belonging to each cell type.

Simulations

To evaluate the performance of the CAESAR suite under scenarios with different resolutions for spatial locations, we designed simulation studies based on a subcellular-resolution CosMx dataset for lung cancer [24].

Scenario 1. Same-resolution reference and target data. For this scenario, we used all fields of view (FOVs) from section Lung5 rep1 as reference datasets. The original annotation was treated as underlying truth, which included 14 cell types and assigned based on gene expression profiles similarity. The FOVs from three other sections (Lung5 rep2, Lung13, and Lung12) were adopted as the target datasets. Therefore, the heterogeneity between the reference and target datasets was naturally considered in our scenario, with Lung5 rep1 and Lung5 rep2 from two consecutive sections of the same donor exhibiting strong similarity.

Scenario 2. High-resolution reference data and low-resolution target data. For this scenario, the reference datasets were same as Scenario 1, and we binned 5 cells as a location in the target datasets to generate low resolution target data. Specifically, we divided each target dataset into grids of equal length and width according to the spatial coordinates, so that each grid contained 5 spots on average. Then, we added the gene expression of the spots located in a grid as the gene expression of the new location, spatial coordinates of which are defined as the grid center and the domain type is defined as the domain cell type in the grid with ties resolved with random select.

Scenario 3. Low-resolution reference data and high-resolution target data. For this scenario, the target datasets were same as in Scenario 1, and the low-resolution reference datasets were generated via the same binned method as in scenario 2.

Real data analyses

All real datasets utilized in this study are comprehensively detailed in the Supplementary Notes. Through rigorous quality control measures, we excluded genes displaying zero expression across multiple spots, those exclusively present in either the reference or target dataset, and spots where numerous genes exhibited no expression. In our analyses, we performed log normalization and identified the top 2000 variable genes using Seurat4 [66]. We treated all genes as variable genes for Xenium and CosMx data analyses, since the number of available genes was less than 2000. For Pixel-seq data analysis, the top 3000 variable genes were calculated due to the high sparsity of Pixel-seq data. The final variable genes used for co-embedding were the intersection of variable genes in the reference and target data. However, we used the variable genes from the reference data to co-embed the spatial ATAC-seq data, as its data consists of gene scores.

Data availability

All datasets used in this study are publicly available. These include the four human non-small-cell lung cancer CosMx data (`https://nanostring.com/products/cosmx-spatial-molecular-imager/ffpe-dataset/nsclc-ffpe-dataset/`); the four human breast cancer Xenium datasets (`https://www.dropbox.com/s/t05w7ccufh1v0h8/xenium_prerelease_jul12_hBreast_replicates.tar?dl=0` and `https://www.10xgenomics.com/products/xenium-in-situ/preview-dataset-human-breast`) as well as its reference data (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE176078`); four human hepatocellular carcinoma Visium datasets (Raw FASTQ data are available at `https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=858545`, and H&E images, which are available at `https://doi.org/10.6084/m9.figshare.21280569.v1` and `https://doi.org/10.6084/m9.figshare.21061990.v1`), as well as its scRNA-seq human reference data (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE125449`) and its scRNA-seq mouse reference data (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE181515`); mouse olfactory bulb ST dataset (`https://www.spatialresearch.org/`) and Pixel-seq dataset (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE186097`), as well as their reference datasets (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111672`; and `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE121891`),

and the mouse spatial ATAC-seq dataset (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM5238385) as well as its accompanying scRNA-seq reference data (https://oncoscape.v3.sttrcancer.org/atlas.gs.washington.edu.mouse.rna/downloads). All other relevant data supporting the key findings of this study are available within the article and its Supplementary Information files or from the corresponding author upon reasonable request.

Code availability

The CAESAR suite was implemented in an open-source, publicly available R package [67] that is available at https://cran.r-project.org/package=CAESAR.Suite and https://github.com/XiaoZhangryy/CAESAR.Suite. Code for reproducing the analysis can be found at https://github.com/XiaoZhangryy/CAESAR.Suite_Analysis.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China (grant #12371283), the University Development Fund from the Chinese University of Hong Kong, Shenzhen (grant # UDF01003033), the Guangdong Provincial Key Laboratory of Mathematical Foundations for Artificial Intelligence (grant # 2023B1212010001), and Shenzhen Key Laboratory of Cross-Modal Cognitive Computing (grant # ZDSYS20230626091302006). Fig. 3a was by Figdraw. Fig. 1b was modified from scidraw (https://scidraw.io/), licensed under a Creative Common Attribution 3.0 Generic License.(https://creativecommons.org/licenses/by/3.0/).

Author contributions

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

Competing Interests

The authors declare no competing interests.

References

[1] Rao, A., Barkley, D., Fran**c**ca, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).

[2] Moses, L. & Pachter, L. Museum of spatial transcriptomics. *Nature methods* **19**, 534–546 (2022).

[3] Tian, L., Chen, F. & Macosko, E. Z. The expanding vistas of spatial transcriptomics. *Nature Biotechnology* **41**, 773–782 (2023).

[4] Deng, Y. *et al.* Spatial profiling of chromatin accessibility in mouse and human tissues. *Nature* **609**, 375–383 (2022).

[5] Liu, Y. *et al.* High-plex protein and whole transcriptome co-mapping at cellular resolution with spatial cite-seq. *Nature Biotechnology* **41**, 1405–1409 (2023).

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

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

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

[9] Liu, W. *et al.* Joint dimension reduction and clustering analysis of single-cell rna-seq and spatial transcriptomics data. *Nucleic acids research* **50**, e72–e72 (2022).

[10] Yang, Y. *et al.* Sc-meb: spatial clustering with hidden markov random field using empirical bayes. *Briefings in bioinformatics* **23**, bbab466 (2022).

[11] Farrell, J. A. *et al.* Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, eaar3131 (2018).

[12] Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics* **20**, 273–282 (2019).

[13] Barkley, D., Rao, A., Pour, M., Fran**c**ca, G. S. & Yanai, I. Cancer cell states and emergent properties of the dynamic tumor system. *Genome research* **31**, 1719–1727 (2021).

[14] Barkley, D. *et al.* Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. *Nature genetics* **54**, 1192–1201 (2022).

[15] Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell rna-seq data across data sets. *Nature methods* **15**, 359–362 (2018).

[16] Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology* **20**, 163–172 (2019).

[17] Tan, Y. & Cahan, P. Singlecellnet: a computational tool to classify single cell rna-seq data across platforms and across species. *Cell systems* **9**, 207–213 (2019).

[18] Zhang, A. W. *et al.* Probabilistic cell-type assignment of single-cell rna-seq for tumor microenvironment profiling. *Nature methods* **16**, 1007–1015 (2019).

[19] Guo, H. & Li, J. scsorter: assigning cells to known cell types according to marker genes. *Genome biology* **22**, 69 (2021).

[20] Shi, X. *et al.* Probabilistic cell/domain-type assignment of spatial transcriptomics data with spatialanno. *Nucleic Acids Research* **51**, e115–e115 (2023).

[21] Cortal, A., Martignetti, L., Six, E. & Rausell, A. Gene signature extraction and cell identity recognition at the single-cell level with cell-id. *Nature biotechnology* **39**, 1095–1102 (2021).

[22] Chen, H., Ryu, J., Vinyard, M. E., Lerer, A. & Pinello, L. Simba: single-cell embedding along with features. *Nature Methods* **21**, 1003–1013 (2024).

[23] Liang, Q., Huang, Y., He, S. & Chen, K. Pathway centric analysis for single-cell rna-seq and spatial transcriptomics data with gsdensity. *Nature communications* **14**, 8416 (2023).

[24] He, S. *et al.* High-plex imaging of rna and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nature Biotechnology* **40**, 1794–1806 (2022).

[25] Zhang, K., Zemke, N. R., Armand, E. J. & Ren, B. A fast, scalable and versatile tool for analysis of single-cell omics data. *Nature methods* **21**, 217–227 (2024).

[26] Tharwat, A. Classification assessment methods. *Applied computing and informatics* **17**, 168–192 (2021).

[27] Wu, S. Z. *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nature genetics* **53**, 1334–1347 (2021).

[28] Janesick, A. *et al.* High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature Communications* **14**, 8353 (2023).

[29] Wu, W. *et al.* Multiple signaling roles of cd3ε and its application in car-t cell therapy. *Cell* **182**, 855–871 (2020).

[30] Zhu, Z. *et al.* Comprehensive analysis reveals a prognostic and therapeutic biomarker cd3d in the breast carcinoma microenvironment. *Bioscience reports* **41**, BSR20202898 (2021).

[31] Tedder, T., Boyd, A., Freedman, A., Nadler, L. & Schlossman, S. The b cell surface molecule b1 is functionally linked with b cell activation and differentiation. *Journal of immunology (Baltimore, Md.: 1950)* **135**, 973–979 (1985).

[32] Yokoyama, K. *et al.* Bank regulates bcr-induced calcium mobilization by promoting tyrosine phosphorylation of ip3 receptor. *The EMBO journal* (2002).

[33] Roifman, C. M., Zhang, J., Chitayat, D. & Sharfe, N. A partial deficiency of interleukin-7rα is sufficient to abrogate t-cell development and cause severe combined immunodeficiency. *Blood, The Journal of the American Society of Hematology* **96**, 2803–2807 (2000).

[34] Gaud, G., Lesourne, R. & Love, P. E. Regulatory mechanisms in t cell receptor signalling. *Nature Reviews Immunology* **18**, 485–497 (2018).

[35] Muro, R., Takayanagi, H. & Nitta, T. T cell receptor signaling for $\gamma\delta$t cell development. *Inflammation and Regeneration* **39**, 1–11 (2019).

[36] Fletcher, G. *et al.* hag-2 and hag-3, human homologues of genes involved in differentiation, are associated with oestrogen receptor-positive breast tumours and interact with metastasis gene c4. 4a and dystroglycan. *British journal of cancer* **88**, 579–585 (2003).

[37] Campa, D. *et al.* Genetic variation in genes of the fatty acid synthesis pathway and breast cancer risk. *Breast cancer research and treatment* **118**, 565–574 (2009).

[38] Nakshatri, H. & Badve, S. Foxa1 in breast cancer. *Expert reviews in molecular medicine* **11**, e8 (2009).

[39] Liu, W. *et al.* Probabilistic embedding, clustering, and alignment for integrating spatial transcriptomics data with precast. *Nature communications* **14**, 296 (2023).

[40] Ma, L. *et al.* Tumor cell biodiversity drives microenvironmental reprogramming in liver cancer. *Cancer cell* **36**, 418–430 (2019).

[41] Zhou, L. *et al.* Lineage tracing and single-cell analysis reveal proliferative prom1+ tumour-propagating cells and their dynamic cellular transition during liver cancer progression. *Gut* **71**, 1656–1668 (2022).

[42] Wu, K. *et al.* Hepatic transforming growth factor beta gives rise to tumor-initiating cells and promotes liver cancer development. *Hepatology* **56**, 2255–2267 (2012).

[43] Tummala, K. S. *et al.* Hepatocellular carcinomas originate predominantly from hepatocytes and benign lesions from hepatic progenitor cells. *Cell reports* **19**, 584–600 (2017).

[44] Lee, R. D. *et al.* Single-cell analysis identifies dynamic gene expression networks that govern b cell development and transformation. *Nature communications* **12**, 6843 (2021).

[45] Bai, X.-S. *et al.* Rnf128 promotes malignant behaviors via egfr/mek/erk pathway in hepatocellular carcinoma. *OncoTargets and therapy* 10129–10141 (2020).

[46] Fukuda, M. *et al.* Genetic polymorphisms of hepatic abc-transporter in patients with hepatocellular carcinoma. *J Cancer Ther* **1**, 114–123 (2010).

[47] Li, C. *et al.* Srebp-1 has a prognostic role and contributes to invasion and metastasis in human hepatocellular carcinoma. *International journal of molecular sciences* **15**, 7124–7138 (2014).

[48] Zhao, Q., Lin, X. & Wang, G. Targeting srebp-1-mediated lipogenesis as potential strategies for cancer. *Frontiers in oncology* **12**, 952371 (2022).

[49] Su, F. & Koeberle, A. Regulation and targeting of srebp-1 in hepatocellular carcinoma. *Cancer and Metastasis Reviews* **43**, 673–708 (2024).

[50] Goel, S. *et al.* Normalization of the vasculature for treatment of cancer and other diseases. *Physiological reviews* **91**, 1071–1121 (2011).

[51] Forster, J. C., Harriss-Phillips, W. M., Douglass, M. J. & Bezak, E. A review of the development of tumor vasculature and its effects on the tumor microenvironment. *Hypoxia* 21–32 (2017).

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

[53] Tepe, B. *et al.* Single-cell rna-seq of mouse olfactory bulb reveals cellular heterogeneity and activity-dependent molecular census of adult-born neurons. *Cell reports* **25**, 2689–2703 (2018).

[54] Birkner, K. *et al.* Neuronal icam-5 plays a neuroprotective role in progressive neurodegeneration. *Frontiers in neurology* **10**, 205 (2019).

[55] Kostyuk, P. G. Calcium channels in the neuronal membrane. *Biochimica et Biophysica Acta (BBA)-Reviews on Biomembranes* **650**, 128–150 (1981).

[56] Urrutia, D. N. *et al.* Comparative study of the neural differentiation capacity of mesenchymal stromal cells from different tissue sources: An approach for their use in neural regeneration therapies. *PloS one* **14**, e0213032 (2019).

[57] Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).

[58] Pressey, J. C., de Saint-Rome, M., Raveendran, V. A. & Woodin, M. A. Chloride transporters controlling neuronal excitability. *Physiological Reviews* **103**, 1095–1135 (2023).

[59] Wang, J., Peng, J., Chen, Y., Nasser, M. & Qin, H. The role of stromal cells in epithelial–mesenchymal plasticity and its therapeutic potential. *Discover Oncology* **15**, 13 (2024).

[60] Baumann, R. & Meuer, H.-J. Blood oxygen transport in the early avian embryo. *Physiological reviews* **72**, 941–965 (1992).

[61] Farnsworth, R. H., Lackmann, M., Achen, M. G. & Stacker, S. A. Vascular remodeling in cancer. *Oncogene* **33**, 3496–3505 (2014).

[62] Bai, Y. & Chu, L. A robust framework for graph-based two-sample tests using weights. *arXiv preprint arXiv:2307.12325* (2023).

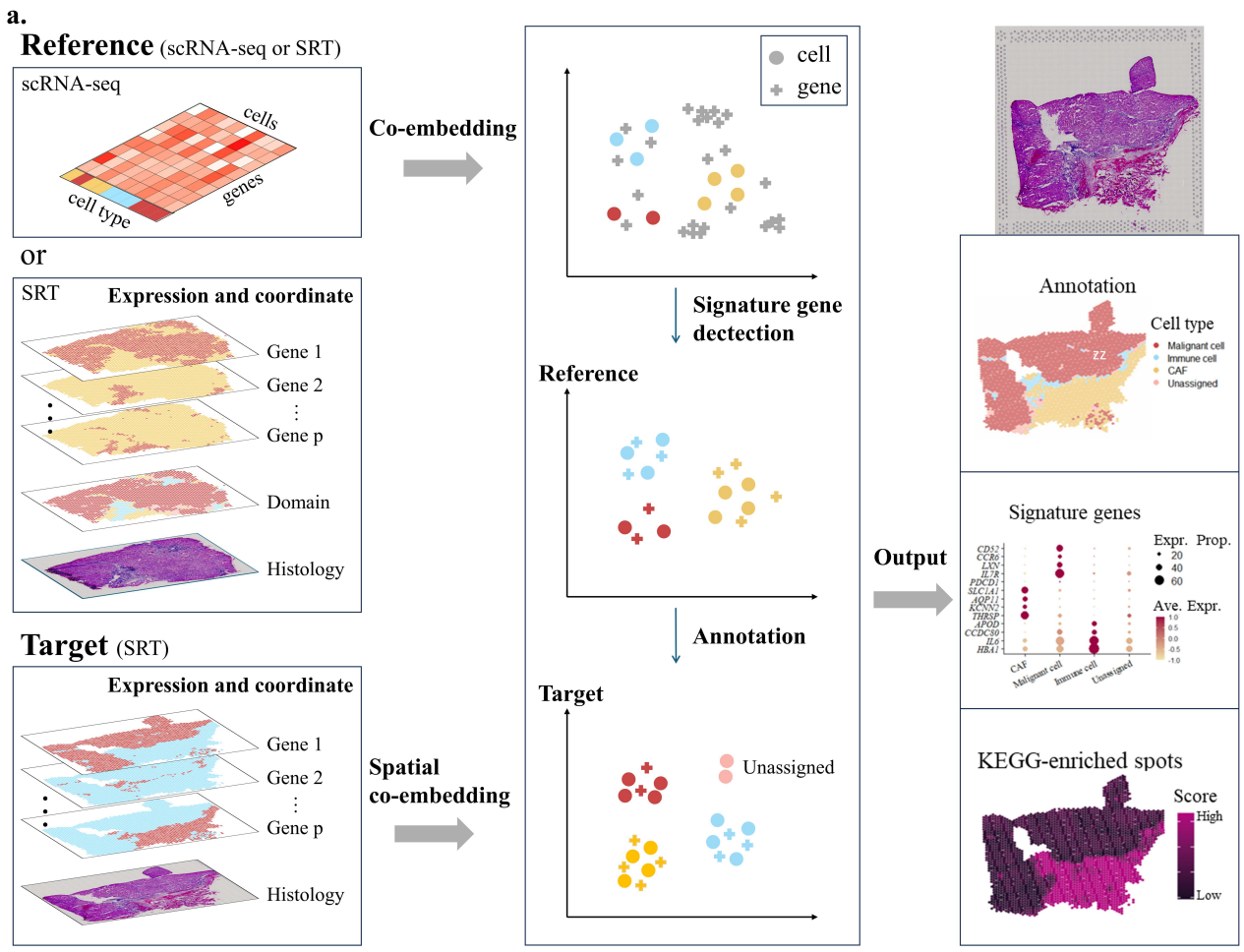
[63] Liu, Y. & Xie, J. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association* **115**, 393–402 (2020).

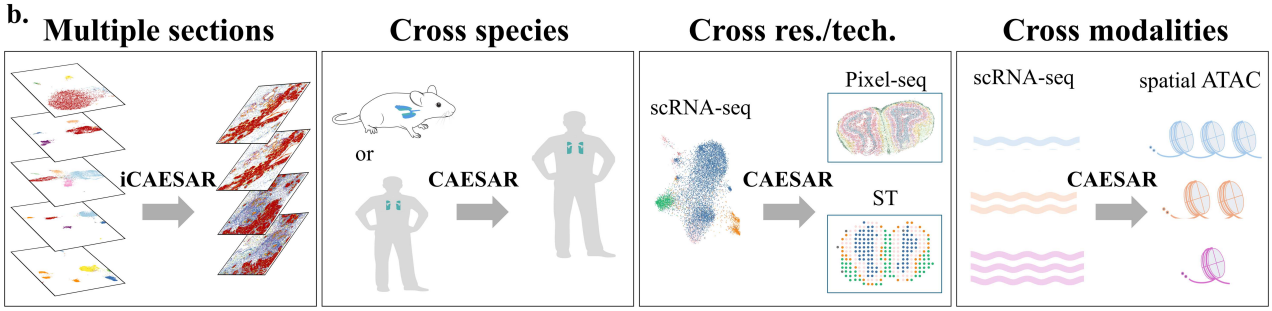
[64] Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of rna-seq data using factor analysis of control genes or samples. *Nature biotechnology* **32**, 896–902 (2014).

[65] Hounkpe, B. W., Chenou, F., de Lima, F. & De Paula, E. V. Hrt atlas v1. 0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive rna-seq datasets. *Nucleic acids research* **49**, D947–D955 (2021).

[66] Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* (2021). URL `https://doi.org/10.1016/j.cell.2021.04.048`.

[67] Zhang, X., Liu, W. & Liu, J. *CAESAR.Suite: CAESAR: a Cross-Technology and Cross-Resolution Framework for Spatial Omics Annotation* (2024). URL `https://github.com/XiaoZhangryy/CAESAR.Suite`. R package version 0.1.0.


a.
Reference (scRNA-seq or SRT)
scRNA-seq
cells
genes
cell type
or
SRT
Expression and coordinate
Gene 1
Gene 2
Gene p
Domain
Histology
Target (SRT)
Expression and coordinate
Gene 1
Gene 2
Gene p
Histology
Co-embedding
Spatial co-embedding
cell
gene
Signature gene dectection
Reference
Annotation
Target
Unassigned
Output
Annotation
Cell type
Malignant cell
Immune cell
CAF
Unassigned
Signature genes
Expr. Prop.
20
40
60
Ave. Expr.
CAF
Malignant cell
Immune cell
Unassigned
KEGG-enriched spots
Score
High
Low


b.
Multiple sections
iCAESAR
Cross species
or
CAESAR
Cross res./tech.
scRNA-seq
Pixel-seq
CAESAR
ST
Cross modalities
scRNA-seq
spatial ATAC
CAESAR

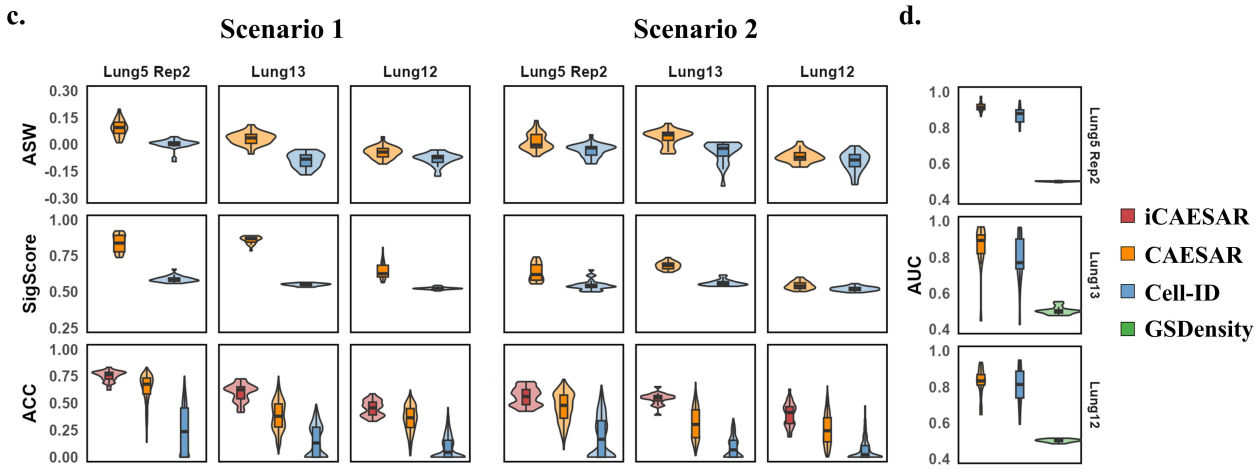

c.
Scenario 1
Scenario 2
Lung5 Rep2
Lung13
Lung12
ASW
SigScore
ACC
d.
AUC
iCAESAR
CAESAR
Cell-ID
GSDensity

Figure 1: Schematic overview of CAESAR suite and simulation results. (a) Left panel: CAESAR suite takes labelled scRNA-seq or spatial transcriptomics sections as reference data and unlabelled spatial transcriptomics sections as target data. Middle panel: For each section, the model projects the cells and genes into a common embedding space, where the gene-cell distance reflects their specificity. Spatial co-embedding integrate morphological or histology images and spatial location information into low-dimensional space to better characterize the gene-cell relationship. The signature genes for a cell/domain type are the top-ranked genes based on their average distance to cells of that cell/domain type. These signature gene sets can be independently extracted from a collection of reference datasets for downstream annotation procedure. Right panel: CAESAR suite performs cell/domain type annotation by evaluating cell signatures against (multiple) cell/domain-type markers from reference datasets. Once the target data is annotated, its signature genes can be detected. When a pathway is provided, CAESAR suite can detect gradual activation of the pathway among locations. (b) The CAESAR suite is capable of flexibly performing annotations for spatial omics datasets with heterogeneous reference datasets, across species, resolutions, technologies, and modalities. (c) Model validation using CosMx data. We used all fields of view (FOVs) from sample Lung5 rep1 as the reference dataset (30 FOVs, 3,109 spots on median), and all FOVs from samples Lung5 rep2, Lung13, and Lung12 as target data (29, 28, and 20 FOVs; 3,530, 2,524, and 4,099 spots on median, respectively) to evaluate performance under different conditions (scenario 1). We binned 5 cells per location in the target datasets to create low-resolution target datasets (706, 495, and 810 spots on median, respectively), which used to evaluate performance with a low-resolution target dataset (Scenario 2). We evaluated performance in terms of cell embedding, co-embedding, and annotation, using average silhouette width (ASW), signature score (SigScore) and classification accuracy (ACC). (d) We used differentially expressed gene sets for each cell type as pathways to evaluate performance on pathway enrichment, which was assessed by the area under the curve (AUC).

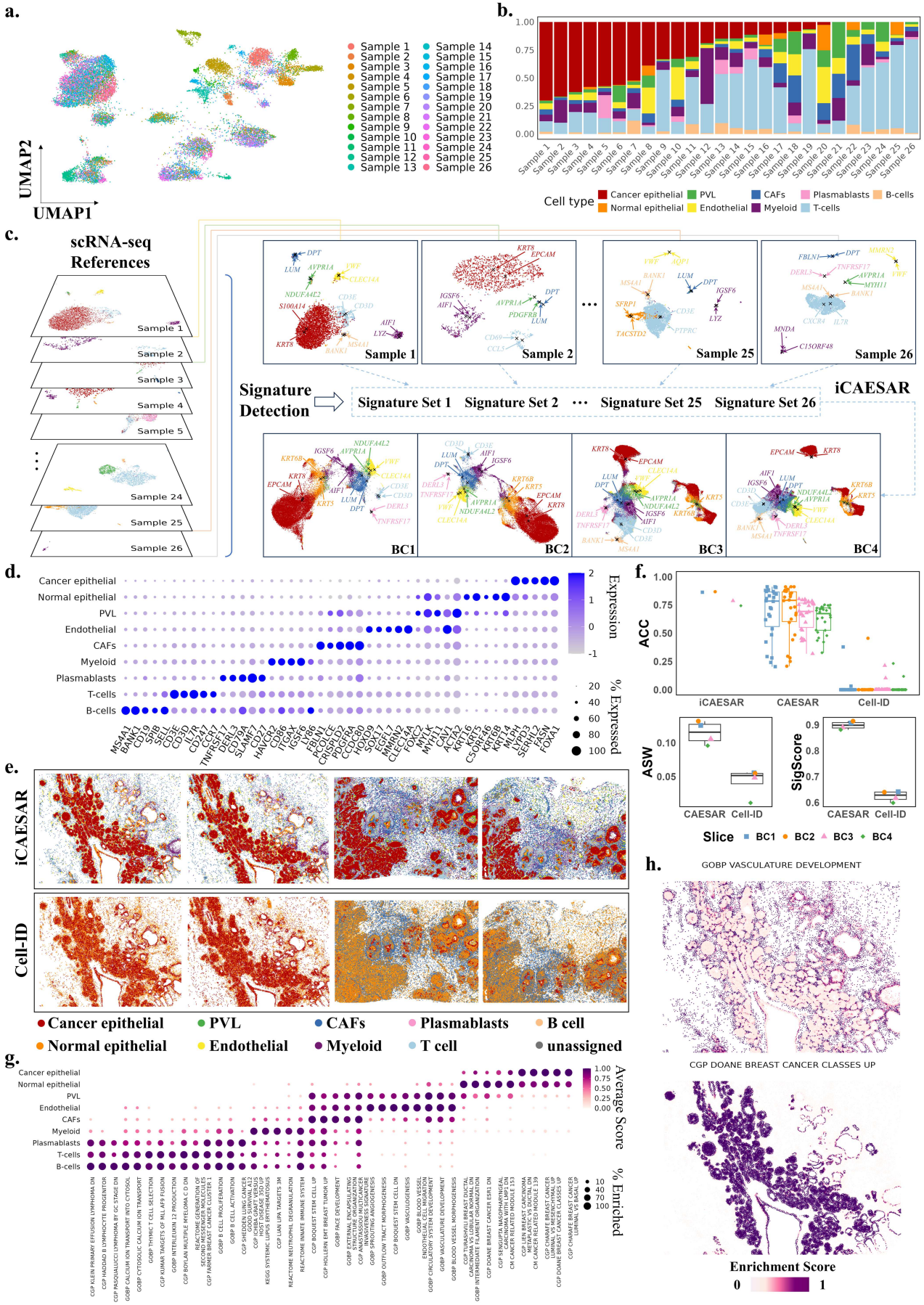

a.
UMAP2
UMAP1
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
Sample 9
Sample 10
Sample 11
Sample 12
Sample 13
Sample 14
Sample 15
Sample 16
Sample 17
Sample 18
Sample 19
Sample 20
Sample 21
Sample 22
Sample 23
Sample 24
Sample 25
Sample 26
b.
Cell type
Cancer epithelial
Normal epithelial
PVL
Endothelial
CAFs
Myeloid
Plasmablasts
T-cells
B-cells
c.
scRNA-seq References
Signature Detection
Signature Set 1
Signature Set 2
Signature Set 25
Signature Set 26
iCAESAR
BC1
BC2
BC3
BC4
d.
Expression
% Expressed
e.
iCAESAR
Cell-ID
Cancer epithelial
Normal epithelial
PVL
Endothelial
CAFs
Myeloid
Plasmablasts
T cell
B cell
unassigned
f.
ACC
ASW
SigScore
iCAESAR
CAESAR
Cell-ID
Slice
g.
Average Score
% Enriched
h.
GOBP VASCULATURE DEVELOPMENT
CGP DOANE BREAST CANCER CLASSES UP
Enrichment Score

Figure 2: Analysis of human breast cancer Xenium data. (a) UMAP plot for 26 reference datasets, colored by the reference identities. (b) Stacked barplot for the cell type proportions from manual annotations in each reference dataset, where CAFs represents cancer-associated fibroblasts and PVL represents perivascular-like cells. (c) Schematic representation of the CAESAR suite's spatial annotations process utilizing multiple single-cell RNA sequencing (scRNA-seq) reference datasets. (d) Dot plot of top five signature genes identified by CAESAR suite for the transferred annotations on four BC sections, where "% expressed" means the percentage of cells that expressed this gene. (e) Spatial heatmaps for annotations obtained by CAESAR suite and Cell-ID. (f) Visual representations of the ASW for assessing the performance of location embeddings, the SigScore for evaluating the efficacy of signature gene detection, and the ACC metric for spatial annotation performance, are presented through boxplots by comparing CAESAR suite and Cell-ID. (g) Dot plot of the top five cell type specific pathways for each transferred cell types by CAESAR suite of four BC sections, where "% enriched" means the percentage of cells in which this pathway was enriched. (h) Spatial heatmaps of enrichment scores for cell-type-specific pathways: for PVL cells, the pathway "vasculature development" from the GOBP database, and for Cancer Epithelial cells, the pathway "Doane Breast Cancer Classes Up" from the CGP database.

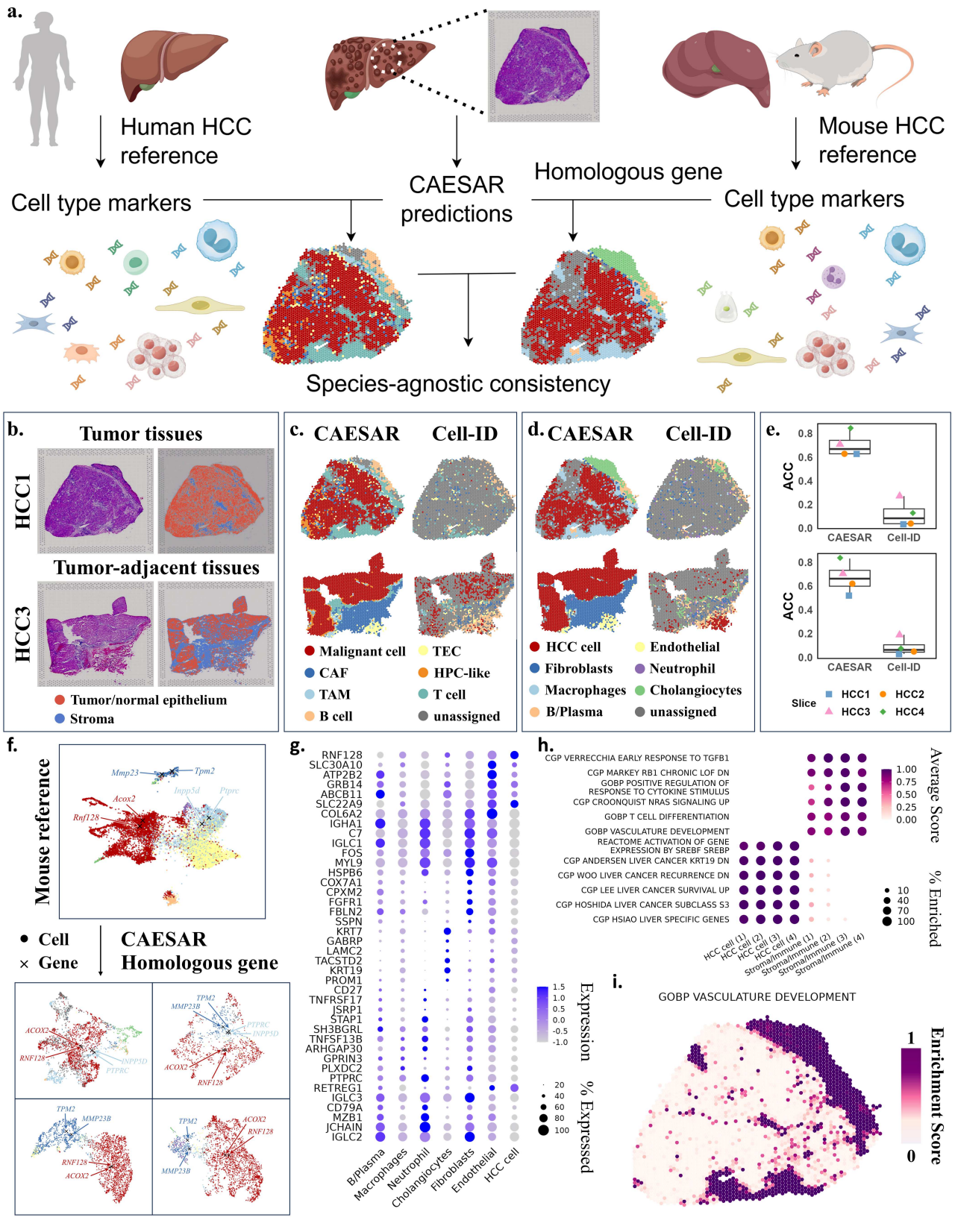

a.
Human HCC reference
Mouse HCC reference
Cell type markers
CAESAR predictions
Homologous gene
Cell type markers
Species-agnostic consistency
b.
Tumor tissues
HCC1
Tumor-adjacent tissues
HCC3
Tumor/normal epithelium
Stroma
c.
CAESAR
Cell-ID
Malignant cell
CAF
TAM
B cell
TEC
HPC-like
T cell
unassigned
d.
CAESAR
Cell-ID
HCC cell
Fibroblasts
Macrophages
B/Plasma
Endothelial
Neutrophil
Cholangiocytes
unassigned
e.
ACC
CAESAR
Cell-ID
Slice
HCC1
HCC2
HCC3
HCC4
f.
Mouse reference
Cell
Gene
CAESAR
Homologous gene
g.
RNF128
SLC30A10
ATP2B2
GRB14
ABCB11
SLC22A9
COL6A2
IGHA1
C7
IGLC1
FOS
MYL9
HSPB6
COX7A1
CPXM2
FGFR1
FBLN2
SSPN
KRT7
GABRP
LAMC2
TACSTD2
KRT19
PROM1
CD27
TNFRSF17
JSRP1
STAP1
SH3BGRL
TNFSF13B
ARHGAP30
GPRIN3
PLXDC2
PTPRC
RETREG1
IGLC3
CD79A
MZB1
JCHAIN
IGLC2
B/Plasma
Macrophages
Neutrophil
Cholangiocytes
Fibroblasts
Endothelial
HCC cell
Expression
% Expressed
h.
CGP VERRECCHIA EARLY RESPONSE TO TGFB1
CGP MARKEY RB1 CHRONIC LOF DN
GOBP POSITIVE REGULATION OF RESPONSE TO CYTOKINE STIMULUS
CGP CROONQUIST NRAS SIGNALING UP
GOBP T CELL DIFFERENTIATION
GOBP VASCULATURE DEVELOPMENT
REACTOME ACTIVATION OF GENE EXPRESSION BY SREBF SREBP
CGP ANDERSEN LIVER CANCER KRT19 DN
CGP WOO LIVER CANCER RECURRENCE DN
CGP LEE LIVER CANCER SURVIVAL UP
CGP HOSHIDA LIVER CANCER SUBCLASS S3
CGP HSIAO LIVER SPECIFIC GENES
HCC cell (1)
HCC cell (2)
HCC cell (3)
HCC cell (4)
Stroma/Immune (1)
Stroma/Immune (2)
Stroma/Immune (3)
Stroma/Immune (4)
Average Score
% Enriched
i.
GOBP VASCULATURE DEVELOPMENT
Enrichment Score

Figure 3: Analysis of human HCC Visium data. (a) Schematic representation of the CAESAR suite's spatial annotations process transferred from human and mouse reference. For each reference, CAESAR co-embeds the locations and genes in a common space, and detect signature genes for each cell type based on their distance. Then, those signature genes are used as cell type markers. The signature genes from mouse were transferred to their homologous human genes. CAESAR's annotation results using references from different species show species-agnostic consistency. (b) H&E image and manual annotations by a pathologist for HCC1 and HCC3. (c) Spatial heatmaps of spatial annotations for HCC1 and HCC3 transferred by CAESAR suite and Cell-ID based on a human HCC scRNA-seq reference. (d) Spatial heatmaps of spatial annotations for HCC1 and HCC3 transferred by CAESAR suite and Cell-ID based on a mouse HCC scRNA-seq reference. (e) Boxplots of annotation accuracy of CAESAR suite and Cell-ID based on human reference (upper panel) and mouse reference (bottom panel). (f) UMAP plots of cells/spots and partial overlapped signature genes between mouse HCC reference data and four target SRT sections. (g) Dot plot of top six signature genes for each transferred cell types by CAESAR suite based on mouse HCC reference. (h) Dot plot of average enrichment scores for cell-type specific pathways. (i) Spatial heatmap of enrichment scores of vasculature development in GOBP database.

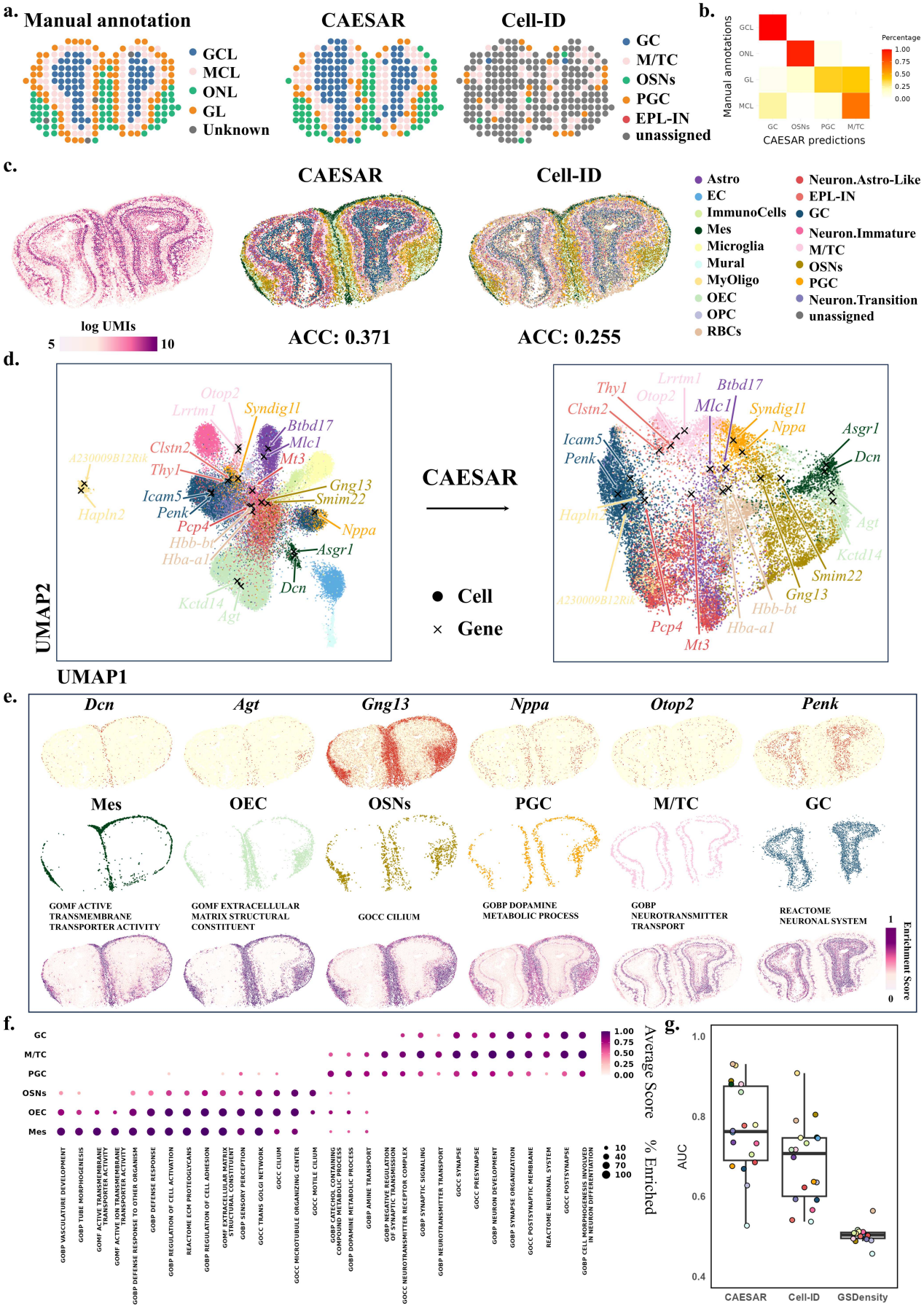
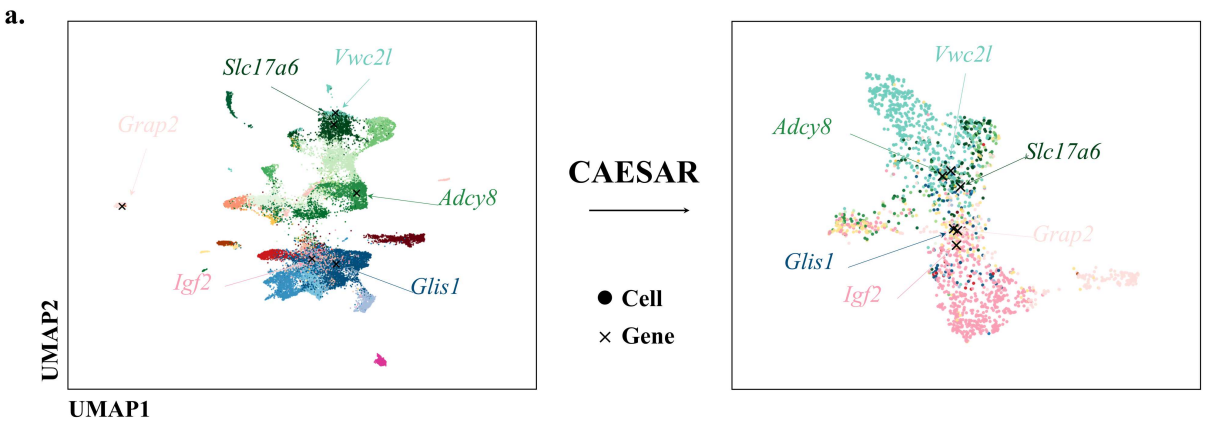
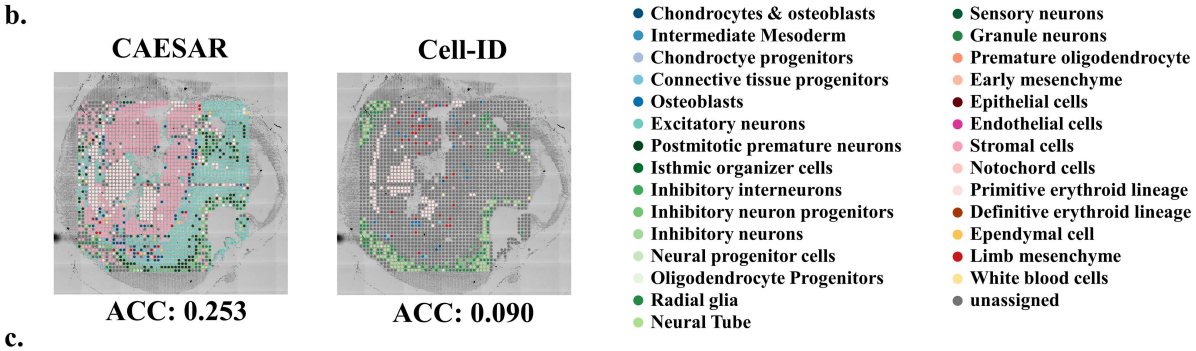

a.
Manual annotation
CAESAR
Cell-ID
GCL
MCL
ONL
GL
Unknown
GC
M/TC
OSNs
PGC
EPL-IN
unassigned
b.
Manual annotations
CAESAR predictions
Percentage
c.
CAESAR
Cell-ID
log UMIs
ACC: 0.371
ACC: 0.255
Astro
EC
ImmunoCells
Mes
Microglia
Mural
MyOligo
OEC
OPC
RBCs
Neuron.Astro-Like
EPL-IN
GC
Neuron.Immature
M/TC
OSNs
PGC
Neuron.Transition
unassigned
d.
UMAP2
UMAP1
CAESAR
Cell
Gene
e.
Dcn
Agt
Gng13
Nppa
Otop2
Penk
Mes
OEC
OSNs
PGC
M/TC
GC
GOMF ACTIVE TRANSMEMBRANE TRANSPORTER ACTIVITY
GOMF EXTRACELLULAR MATRIX STRUCTURAL CONSTITUENT
GOCC CILIUM
GOBP DOPAMINE METABOLIC PROCESS
GOBP NEUROTRANSMITTER TRANSPORT
REACTOME NEURONAL SYSTEM
Enrichment Score
f.
Average Score
% Enriched
g.
AUC
CAESAR
Cell-ID
GSDensity

Figure 4: Analysis of MOB ST and Pixel-seq data. (a) Spatial heatmaps of manual annotations (left panel), annotations obtained by CAESAR suite and Cell-ID for MOB ST data, where GCL, the granule cell layer; MCL, the mitral cell layer; ONL, the nerve layer; GL, the glomerular layer; GC, granule cell; M/TC, mitral and tufted cell; OSNs, Olfactory sensory neurons; PGC, periglomerular cell; EPL-IN, external plexiform layer interneuron. (b) Heatmap of confusion matrix between manual annotations and the predicted cell types obtained by CAESAR suite. (c) Spatial heatmaps of logarithm of UMIs, and annotations obtained by CAESAR suite and Cell-ID for MOB Pixel-seq data, the cell types include: Astro, astrocyte; EC, endothelial cell; ImmunoCells, monocyte and macrophage; Mes, mesenchymal cell; Microglia, microglia; Mural, mural cell; MyOligo, myelinating oligodendrocyte; OEC, olfactory ensheathing cell; OPC, oligodendrocyte precursor; RBCs, red blood cells; Neuron.Astro-Like, astrocyte like neuron; EPL-IN; GC; Neuron.Immature, immature neuron; M/TC; OSNs; PGC; Neuron.Transition, transitional neuron. (d) UMAP plots of embeddings for cells/spots and two overlapped signature genes between MOB scRNA-seq reference and the MOB Pixel-seq data. (e) Spatial heatmaps of expression levels of the cell type specific genes, cell types and enrichment scores of cell type differentially enriched pathways. (f) Dot plot of average enrichment scores for cell-type differentially enriched pathways in MOB Pixel-seq data. (g) Boxplot of AUC obtained by CAESAR suite, Cell-ID and GSDensity for assessing the pathway enrichment performance.


a.
Slc17a6
Vwc2l
Grap2
Adcy8
Igf2
Glis1
UMAP2
UMAP1
CAESAR
Cell
Gene
Vwc2l
Adcy8
Slc17a6
Glis1
Grap2
Igf2


b.
CAESAR
Cell-ID
ACC: 0.253
ACC: 0.090
Chondrocytes & osteoblasts
Intermediate Mesoderm
Chondroctye progenitors
Connective tissue progenitors
Osteoblasts
Excitatory neurons
Postmitotic premature neurons
Isthmic organizer cells
Inhibitory interneurons
Inhibitory neuron progenitors
Inhibitory neurons
Neural progenitor cells
Oligodendrocyte Progenitors
Radial glia
Neural Tube
Sensory neurons
Granule neurons
Premature oligodendrocyte
Early mesenchyme
Epithelial cells
Endothelial cells
Stromal cells
Notochord cells
Primitive erythroid lineage
Definitive erythroid lineage
Ependymal cell
Limb mesenchyme
White blood cells
unassigned


c.
GOCC CULLIN RING UBIQUITIN LIGASE COMPLEX
GOBP REGULATION OF TRANSCRIPTION REGULATORY REGION DNA BINDING
GOBP EMBRYONIC CAMERA TYPE EYE MORPHOGENESIS
CGP ZHOU PANCREATIC ENDOCRINE PROGENITOR
CGP BENPORATH ES CORE NINE CORRELATED
GOBP REPLACEMENT OSSIFICATION
GOBP POSITIVE REGULATION OF CARTILAGE DEVELOPMENT
CGP VERRECCHIA RESPONSE TO TGFB1 C2
CGP SANSOM APC MYC TARGETS
CGP VERRECCHIA EARLY RESPONSE TO TGFB1
GOMF ANTIOXIDANT ACTIVITY
CGP VANDESLUIS NORMAL EMBRYOS DN
CGP WELCH GATA1 TARGETS
GOMF OXIDOREDUCTASE ACTIVITY ACTING ON PEROXIDE AS ACCEPTOR
GOBP GAS TRANSPORT
CGP BLANCO MELO HUMAN PARAINFLUENZA VIRUS 3 INFECTION A594 CELLS DN
GOBP NEGATIVE REGULATION OF MUSCLE TISSUE DEVELOPMENT
GOBP CELL SURFACE RECEPTOR SIGNALING PATHWAY INVOLVED IN HEART DEVELOPMENT
GTRD ZNF746 TARGET GENES
GOBP EPITHELIAL TO MESENCHYMAL TRANSITION INVOLVED IN ENDOCARDIAL CUSHION FORMATION
MIRDB MIR3162 5P
REACTOME COMPLEMENT CASCADE
CGP SHEDDEN LUNG CANCER GOOD SURVIVAL A12
GOCC NODE OF RANVIER
CGP MCCLUNG DELTA FOSB TARGETS 2WK
MIRDB MIR11399
GOBP INORGANIC ANION TRANSMEMBRANE TRANSPORT
GOBP CHLORIDE TRANSMEMBRANE TRANSPORT
MIRDB MIR3074 3P
CGP ZHANG GATA6 TARGETS UP
Excitatory neurons
Postmitotic premature neurons
Stromal cells
Primitive erythroid lineage
Chondrocytes & osteoblasts
Radial glia
Average Score
1.00
0.75
0.50
0.25
0.00
Percent Enriched
10
40
70
100

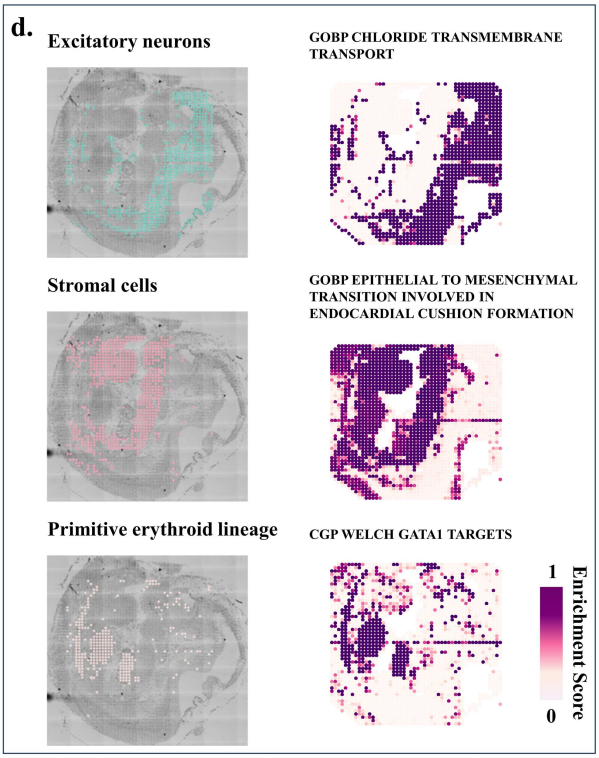

d.
Excitatory neurons
GOBP CHLORIDE TRANSMEMBRANE TRANSPORT
Stromal cells
GOBP EPITHELIAL TO MESENCHYMAL TRANSITION INVOLVED IN ENDOCARDIAL CUSHION FORMATION
Primitive erythroid lineage
CGP WELCH GATA1 TARGETS
1
0
Enrichment Score

Figure 5: Analysis of mouse Embryo 11 spATAC-seq data. (a) UMAP plots of co-embeddings for cells/spots and overlapped signature genes between mouse embryo scRNA-seq reference and the mouse Embryo 11 spATAC-seq data. (b) Spatial heatmaps for cell-type assignment of CAESAR and Cell-ID. (c) Dot plot of the top five differentially enriched pathways for each of top six domain cell types. (d) Spatial heatmaps for top three domain cell types and the corresponding enriched pathway.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementrayData1.xlsx
- CAESARsupplementary.pdf