Method

Joint imputation and deconvolution of gene expression across spatial transcriptomics platforms

Hongyu Zheng,[1,3] Hirak Sarkar,[1,2,3] and Benjamin J. Raphael[1]

[1]*Department of Computer Science, Princeton University, Princeton, New Jersey 08540, USA;* [2]*Ludwig Cancer Institute, Princeton Branch, Princeton University, Princeton, New Jersey 08540, USA*

Spatially resolved transcriptomics (SRT) technologies measure gene expression across thousands of spatial locations within a tissue slice. Multiple SRT technologies are currently available and others are in active development, with each technology having varying spatial resolution (subcellular, single-cell, or multicellular regions), gene coverage (targeted vs. whole-transcriptome), and sequencing depth per location. For example, the widely used 10x Genomics Visium platform measures whole transcriptomes from multiple-cell-sized spots, whereas the 10x Genomics Xenium platform measures a few hundred genes at subcellular resolution. A number of studies apply multiple SRT technologies to slices that originate from the same biological tissue. Integration of data from different SRT technologies can overcome limitations of the individual technologies, enabling the imputation of expression from unmeasured genes in targeted technologies and/or the deconvolution of admixed expression from technologies with lower spatial resolution. Here, we introduce Spatial Integration for Imputation and Deconvolution (SIID), an algorithm to reconstruct a latent spatial gene expression matrix from a pair of observations from different SRT technologies. SIID leverages a spatial alignment and uses a joint nonnegative factorization model to accurately impute missing gene expression and infer gene expression signatures of cell types from admixed SRT data. In simulations involving paired SRT data sets from different technologies (e.g., Xenium and Visium), SIID shows superior performance in reconstructing spot-to-cell-type assignments, recovering cell type–specific gene expression and imputing missing data compared to contemporary tools. When applied to real-world 10x Xenium-Visium pairs from human breast and colon cancer tissues, SIID achieves highest performance in imputing holdout gene expression.

[Supplemental material is available for this article.]

Spatially resolved transcriptomics (SRT) technologies have transformed the study of tissue biology by enabling the simultaneous measurement of gene expression at thousands to hundreds of thousands of locations within a tissue section, along with the spatial coordinates of each location. SRT technologies allow researchers to study complex spatial gene expression patterns and intricate cellular organization, providing a closer look at the tissue microenvironment and spatial context for a given disease (Rao et al. 2021; Palla et al. 2022). There are multiple SRT technologies currently in use with varying spatial resolution and breadth/depth of gene expression (Rozenblatt-Rosen et al. 2020; Moses and Pachter 2022). For example, in situ capture-based SRT such as Slide-seq (Ståhl et al. 2016; Rodriques et al. 2019) and 10x Genomics Visium measure thousands of genes using barcoded beads (of radius 10–50 μm) on a slide. Here, a single bead captures mRNA molecules from multiple spatially nearby cells and thus the gene expression measurement is for a mixture of multiple cells. In contrast, in situ sequencing- (Ke et al. 2013; Wang et al. 2018) and in situ hybridization-based SRT such as 10x Genomics Xenium (Janesick et al. 2023; Oliveira et al. 2025), merFISH (Chen et al. 2015), CosMx (NanoString), and MERSCOPE (Vizgen) measure the expression for a subset of preselected genes at cellular or subcellular resolution. The number of mRNA molecules measured for each gene varies considerably across different platforms.

Given the varying properties of individual SRT platforms, it is advantageous to integrate information from two or more platforms (e.g., combining information from whole-transcriptome platforms with lower spatial resolution with platforms that measure expression of a limited number of genes at high spatial resolution). Such integration could assist in two tasks: (1) predicting the expression of genes that are missing in the high resolution SRT data sets by referencing its expression in the low resolution SRT data set, a process known as imputation; and (2) inferring the mixture proportion of different cell types in a spot from the low resolution SRT data set by using the high resolution SRT data set as a reference, commonly referred to as deconvolution.

Multiple methods have been introduced to impute gene expression across single-cell sequencing technologies including scRNA-seq and single-cell sequencing assay for transposase accessible chromatin (scATAC-seq) (Lopez et al. 2018; Korsunsky et al. 2019; Stuart et al. 2019; Wu et al. 2021; Cohen Kalafut et al. 2023; Cao et al. 2024). These tools commonly integrate multiple single-cell modalities into a shared latent space, allowing them to impute gene expression in cells on one modality by identifying nearby cells from a complementary modality within the latent space. A common way of modeling the latent space is by finding a low-dimensional factorization of scRNA-seq gene expression matrices (Yang and Michailidis 2016; Argelaguet et al. 2018, 2020; Townes et al. 2019; Liu et al. 2020; Qian et al. 2022).

Similarly, in the spatial transcriptomics domain, the SRT gene expression is traditionally imputed by integrating with a reference scRNA-seq data set, often disregarding the spatial information

[3]**These authors contributed equally to this work. Order determined by a coin flip.**

Corresponding author: braphael@princeton.edu

Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.280555.125. Freely available online through the *Genome Research* Open Access option.

inherent in the SRT data set, treating it as a nonspatial modality (Lopez et al. 2019). Other methods such as Tangram (Biancalani et al. 2021), SpaGE (Abdelaal et al. 2020), and SpaOTsc (Cang and Nie 2020) impute genes in a SRT data set by learning a mapping between each spatial location and the reference single cells. On the other hand, akin to the modeling of single-cell data sets, several recent methods use low-dimensional factorization (Chidester et al. 2023; Townes and Engelhardt 2023) to model SRT data, taking into account the spatial information. However, these methods are only applicable to a single data set and therefore are not suitable for imputation across multiple SRT modalities.

All of the above mentioned tools used for imputation implicitly assume at least one of the modalities is scRNA-seq. As a result, applying them to impute gene expression across two SRT data sets faces two significant limitations. First, the spatial information in the data sets is ignored by the algorithm, and second, the reference data set is assumed to have single-cell resolution. These limitations have been partially addressed by recently developed spatial alignment based tools, SLAT (Xia et al. 2023) and SANTO (Li et al. 2024), which impute missing gene expression in one of the SRT data sets based on a learned spatial alignment. These tools do not handle platform-specific differences and overlook the need of deconvolution in case of low spatial resolution. To address the issue of admixed gene expression in technologies with multicellular spatial resolution, a number of methods have been developed to deconvolve admixed SRT spots (Andersson et al. 2020; Miller et al. 2022; Vahid et al. 2023). However, these tools are not designed for imputation across two SRT data sets, and furthermore, they assume perfectly matched gene sets across the data sets, making them unusable for imputing gene expression in one SRT data from another.

We introduce **S**patial **I**ntegration for **I**mputation and **D**econvolution (SIID), an algorithm that simultaneously imputes missing genes in a targeted single-cell (or subcellular) resolution SRT data set (e.g., 10x Genomics Xenium) and deconvolves a supercellular resolution whole-transcriptome SRT data set (e.g., 10x Genomics Visium). SIID uses nonnegative matrix factorization (NMF) to construct a latent gene expression matrix from the views obtained by the two SRT technologies. We demonstrate the effectiveness of SIID on both simulated and two paired Xenium-Visium data sets by evaluating the imputed gene expression and the inferred cell type mixtures. On paired 10x Genomics Xenium and Visium data sets from a breast tumor and colorectal tumor, we demonstrate that SIID enables fine-grained cell typing and better characterization of tumor microenvironments.

Results

Spatial Integration for Imputation and Deconvolution

SIID performs joint imputation and deconvolution of a single-cell spatially resolved transcriptomics data set (e.g., 10x Genomics Xenium or MERFISH) measuring a targeted gene panel, and a lower resolution spatially resolved transcriptomics data set measuring a large number of genes (e.g., 10x Genomics Visium, Slide-seq, or DBiT-seq). Given an alignment of the two SRT slices, SIID performs a spatially regularized nonnegative matrix factorization, with components shared across slices to capture common biological information. For the sake of clarity and to match data sets where we apply SIID below, we denote these data sets as the Xenium and Visium data sets for the remainder of this manuscript.

Given the observed Xenium expression matrix A_X with gene panel G_X and the observed Visium expression matrix A_V with gene panel G_V ($G_X \subset G_V$), along with the spatial mapping Γ from Xenium spots to Visium spots, SIID constructs Q, a shared gene expression matrix for latent factors with a gene panel of G_V, and matrix P containing the Xenium spot-to-cell-type assignments. The Xenium data are modeled as $A_X \approx PQ_T$, where Q_T is the subset of Q for genes in G_X. Similarly, the Visium data are modeled as $A_V \approx (\Gamma^T P)Q$ (Fig. 1). This yields a shared nonnegative matrix factorization of the observations with latent dimension equal to the number of cell types. By constraining the number of latent factors (rows of Q or columns of P) reflecting underlying cell types, SIID jointly estimates P and Q, enabling imputation of the absent Xenium genes and deconvolution of the Visium spots.

Compared with prior NMF-based methods, SIID differs in two key aspects. First, most existing methods either work with a single modality or fit two NMFs sharing one of the components. In our model, the NMFs of two SRT data sets share both components via the spatial mapping. Second, by sharing both NMF components, our model learns the cell type assignments for admixed spots in a way that respects spatial information.

Setup and evaluation

For simulation and evaluation, we use two publicly available data sets of paired Xenium and Visium SRT from adjacent tissue sections: one from a breast cancer (BRCA) pair (Janesick et al. 2023) and another from a colorectal cancer (CRC) pair (Oliveira et al. 2025). Xenium, Visium, and scRNA-seq data for both data sets (and cell type annotation for BRCA) are downloaded and aligned as described in Supplemental Methods C.1. With aligned coordinates, Γ is generated by matching each Xenium spot to its closest Visium spot up to a distance of 100 μm. Furthermore, Supplemental Methods C.2 and Supplemental Table S1 contain detailed data set statistics.

Imputation setup and holdout evaluation

For both the BRCA and CRC data sets, we divide Xenium panel genes into 10 equal-sized folds. For each fold, we remove the expression of genes in fold from the Xenium data set, then run the imputation algorithm to obtain estimates of these holdout genes. To evaluate the imputation results, for each gene in the fold, we compute the R^2 score between estimated expression for each Xenium cell and the observed counts (which were held out during training). The R^2 score between two vectors (x, y), also called the coefficient of determination, equals $\text{cov}^2(x, y)/\sigma_x^2\sigma_y^2$, where cov is the covariance and σ^2 is the variance. The R^2 score for a given model is the average gene-wise holdout R^2 across all 10 folds, where each gene is a holdout exactly once across 10 folds.

SIID recovers the cell type expression in simulated data set

We evaluate the performance of SIID on a simulated paired Xenium and Visium data set generated from the BRCA data set (Janesick et al. 2023). The evaluation focuses on three aspects: imputation of genes not present in the Xenium data; recovery of the spatial location to cell type assignments; and accuracy of cell type gene expression profiles.

To simulate a realistic expression profile, we use the gene expression for the scRNA-seq data set and the spatial coordinates from the Xenium and Visium data. For this simulation, we used clusters (see Supplemental Methods B.1) obtained from

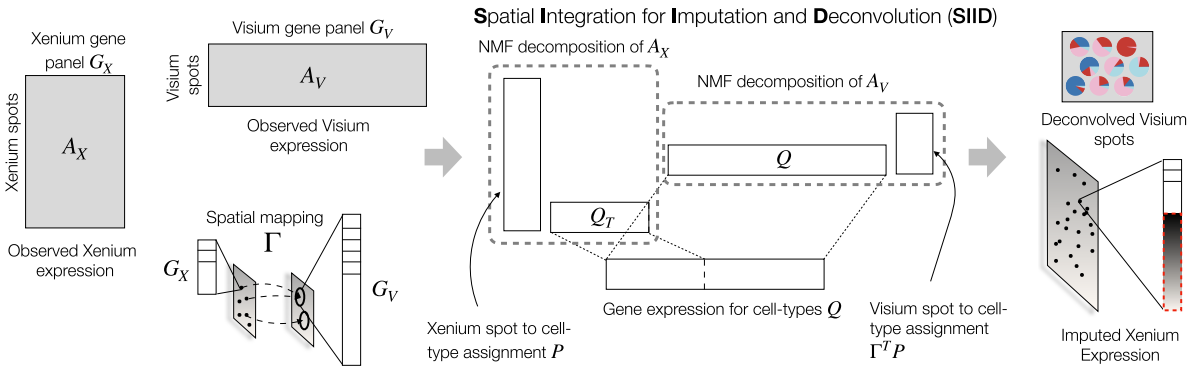

Figure 1. Overview of Spatial Integration for Imputation and Deconvolution (SIID). Given Xenium and Visium expression matrices A_X and A_V, respectively, with corresponding gene panels G_X and G_V (with $G_X \subset G_V$), and a Xenium to Visium spatial mapping Γ, SIID finds a NMF decomposition of a latent gene count matrix A_U, with corresponding factorizations of A_X and A_V, into location to latent factor assignment matrix P and the gene expression matrix Q for latent factors. The estimated parameters are used to impute the absent gene expression in Xenium data and predict the cell type mixture proportions for each Visium spot.

unsupervised clustering of the scRNA-seq data and define them as cell types. Each Xenium spot is assigned to a cell type using a checkerboard spatial pattern (as described in Fig. 2A; Jackson et al. 2024), where each grid on the checkerboard has a different cell type from its neighbors. We use the average gene expression for individual cell type from the scRNA-seq data set to simulate the gene counts for Xenium spots according to their cell type. Expression data for Visium spots are generated by applying the spatial mapping from Xenium to Visium data (see detailed steps in Supplemental Methods B.1).

We created a variety of simulation data sets by varying (i) the number of grids, l on each side of the checkerboard with $l \in \{10, 20\}$, (ii) gene counts generated for each Xenium spots, and (iii) the number of cell types h, where $h \in \{4, 8, 16\}$. To simulate Xenium gene expression, we began with the actual vector of UMI counts N_X from the BRCA Xenium data set. This was then scaled using a coverage fraction ρ (which we refer to as *coverage*) to create a UMI count vector $N'_X = \rho N_X$, with $\rho \in \{0.25, 0.5, 1, 2\}$, which is used for simulating gene expression for Xenium spots. By varying l, h, ρ, we created in total 24 distinct configurations, each representing varying levels of difficulty. We profiled and compared the run time and memory usage of SIID and Tangram as described in Supplemental Methods B.4 and Supplemental Figure S7. We also studied the effect of additional factors in simulation, namely, spatially variable coverage, noise in spatial mapping matrix, and using platform scaling factor in the imputation results (Supplemental Methods B.3, B.5, B.6; Supplemental Figs. S2–S4, S11).

We evaluated the accuracy of SIID for gene expression imputation and cell type deconvolution in simulations, comparing its performance to Tangram, STdeconvolve, and SpiceMix. For the configuration with $l = 20$ grids and $h = 8$ cell types, SIID achieves superior R^2 scores when compared to Tangram (Biancalani et al. 2021) across different coverage levels (Fig. 2B). On the same simulation configuration, SIID outperforms STdeconvolve and SpiceMix by achieving the lowest average Jensen-Shanon (JS) divergence (Fig. 2C) between the predicted cell type mixture proportion (see Supplemental Methods B.2) and the ground truth. On other simulation configurations (Supplemental Fig. S1), SIID consistently receives the best R^2 scores when compared to Tangram and comparable JS divergence when compared to STdeconvolve, and SpiceMix (see Supplemental Methods B.2). Additionally, we created a challenging simulation scenario by setting grid size $l = 100$ (Supplemental Fig. S16A), where we observed that SIID achieves

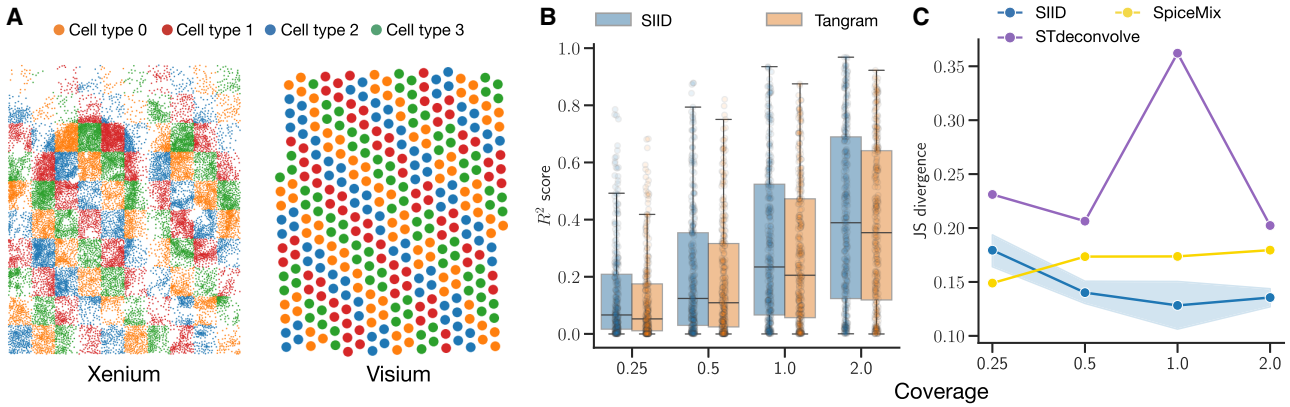

Figure 2. Evaluating SIID on simulated data. (*A*) A simulated Xenium and corresponding Visium data set with gene expression of cell types obtained from a matching scRNA-seq data set (Janesick et al. 2023). (*B*) R^2 of imputed gene expression for holdout genes from SIID and Tangram stratified by coverage. (*C*) Average Jensen-Shannon (JS) divergence between ground truth and predicted cell type mixture proportions predicted by SIID, STdeconvolve, and SpiceMix for simulated Visium data over all spots stratified by coverage.

significant improvement in R^2 score when compared to Tangram (Supplemental Fig. S16B).

Imputing missing genes in cancer SRT data

To perform benchmarking on BRCA and CRC data sets (Results, "Imputation setup and holdout evaluation"), we ran SIID with number of latent factors $h = 20$ for BRCA and 40 for CRC data sets, entropy regularization with $\lambda = 1000$, platform scaling enabled, 5000 training epochs and three restarts. For training, we used genes that are available in both Xenium and Visium data sets, excluding the ones in the holdout set. We document the details of hyperparameters in Supplemental Methods C.3.

We benchmarked four existing methods (Li et al. 2022) for evaluating the imputation results: Tangram (Biancalani et al. 2021) (in cell and cluster modes), gimVI (Lopez et al. 2019), SLAT (Xia et al. 2023), and SANTO (Li et al. 2024) (with precomputed mapping Γ). However, we do not present results for gimVI (Lopez et al. 2019) due to run time errors. In addition, we present four baseline approaches by using the spatial mapping Γ: Baseline A imputes gene expression for each Xenium cell by simply using the gene expression from the corresponding Visium spot based on the spatial mapping; Baseline B is a variant of Baseline A that takes total count per Xenium cell into account; and Baselines C and D employ *k*-nearest-neighbor based smoothing based on Baselines *A* and *B*, respectively. Detailed procedures for benchmarking are in Supplemental Methods C.4, and run times for benchmarking are in Supplemental Methods C.6 and Supplemental Table S2. In addition, we show that SIID is scalable and runs efficiently with a larger number of latent factors, genes, Xenium spots, or Visium spots (Supplemental Methods C.6; Supplemental Fig. S5).

SIID accurately imputes genes in the holdout experiments

Compared to other methods, SIID achieves the best imputation performance on both BRCA and CRC data sets, as measured by average holdout R^2 scores across all 10-folds (Table 1). Among the competing methods, Tangram performs reasonably well when run in cell mode (referred to as Tangram [cell] in Table 1). In addition to R^2 scores, SIID also consistently outperforms other benchmarked methods on four other correlation metrics (Supplemental Methods D.2; Supplemental Table S5; Li et al. 2022). Additionally, SIID shows robust performance in the presence of noise in the spatial mapping matrix (Supplemental Methods C.5; Supplemental Fig. S6).

SIID also achieves a higher R^2 score for most of the individual genes compared to the closest competitor (Fig. 3A). For this evaluation, we compared SIID with Tangram (in cell mode) by evaluating the R^2 scores of both methods for each gene in the Xenium panel from the BRCA data set (see Supplemental Fig. S8A for CRC) along with the total UMI counts. We observed that SIID outperforms Tangram on most of the genes, whereas Tangram performs better for a small number of genes. We further visualized the difference of R^2 scores and its relationship to expression level in Supplemental Figure S9 and concluded that there is no clear correlation between them. As SIID is a generative model, we also validated the model by showing its ability to predict the sparsity of holdout genes (Supplemental Fig. S10).

SIID recovers spatial pattern of Xenium gene expression

SIID is superior to Tangram in recovering spatial patterns for certain marker genes (Fig. 3B–D; Supplemental Fig. S12; Janesick et al. 2023). In particular, we found that the SIID-imputed expression for *ZEB2*, a well-known oncogenic driver implicated in epithelial-mesenchymal transition (EMT) in breast cancer (di Gennaro et al. 2018; De Coninck et al. 2019), closely mirrors the true tissue structure ($R^2 = 0.48$) (Fig. 3C), whereas Tangram's imputed expression is more uniformly distributed across the slice ($R^2 = 0.29$) (Fig. 3D). We present similar results for other marker genes in Supplemental Figure S12.

Comparison to imputing with paired scRNA-seq

On imputing Xenium genes, SIID with Visium data outperforms Tangram paired with a scRNA-seq reference (Table 2). SLAT and SANTO are only applicable to SRT data sets and therefore are not included in this comparison. gimVI failed to finish on several runs (details in Supplemental Methods C.4) for both modes and both data sets (Supplemental Table S6). SIID is not specifically designed to impute Xenium gene expression with a nonspatial reference, but Tangram can be run in both cell and cluster mode for this setup. In order to run Tangram on the CRC data set in a reasonable time, we downsampled the data set as described in Supplemental Methods C.4. To emphasize, despite not using the high resolution scRNA-seq data, SIID outperforms Tangram with access to scRNA-seq data, although with a smaller lead.

Deconvolving cell types in cancer SRT data

We also evaluated SIID's performance in deconvolving the admixed gene expression in Visium data using the paired Xenium data. For deconvolving Visium spots without annotation, we ran the same setup as described in Results, "Imputing missing genes in cancer SRT data" and in Supplemental Methods C.3, with no holdout genes and $\lambda = 500$ for entropy regularization. Here, we evaluated only with the BRCA data set, as this data set includes annotated cell types for both the Xenium and scRNA-seq data provided by 10x Genomics.

SIID recovers most annotated major cell types

SIID recovers most annotated cell types with more than 1000 cells by mapping them to one or a few latent factors (Fig. 4A). To benchmark this, we trained the model on the BRCA data set. Because the

Table 1. Comparison of average holdout R^2 scores across methods

		Tangram				Baselines			
Data set	SIID	Cell mode	Cluster mode	SLAT	SANTO	A	B	C	D
BRCA	**0.2527**	0.1874	0.1556	0.0688	0.1042	0.0373	0.0807	0.0766	0.0664
CRC	**0.2248**	0.1789	0.1382	0.0608	0.0727	0.0325	0.0573	0.0443	0.0412

Bold indicates the best performer.

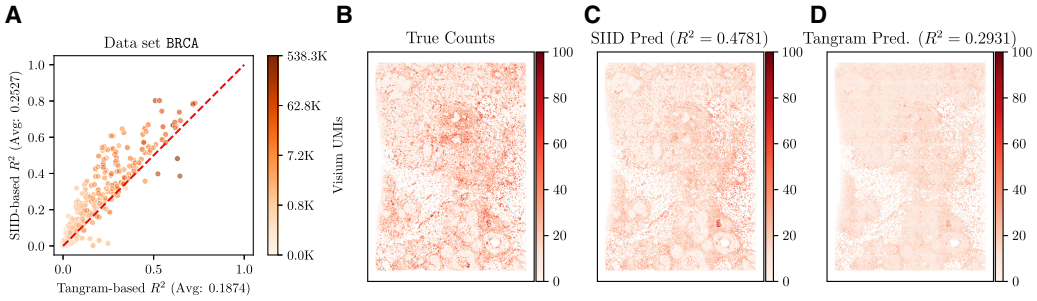

Figure 3. Analysis of SIID imputation on breast cancer (BRCA) data. (*A*) Comparison of holdout R^2 score of SIID (*y*-axis, correlation score averaged over five runs) and Tangram in cell mode (*x*-axis) for each gene in the Xenium panel for the BRCA data set. Each point on the plot corresponds to a single gene, whose color corresponds to the number of Visium UMIs mapped to the gene on a log-scale. Genes *above* the red $y = x$ line have higher imputation performance for SIID compared to Tangram. (*B*) Expression of gene *ZEB2* in Xenium (ground truth for evaluating imputation), (C) SIID prediction of *ZEB2* expression when the gene is held out ($R^2 = 0.4781$ against the ground truth). (*D*) Tangram prediction of *ZEB2* expression ($R^2 = 0.2931$ against ground truth). Total counts of each gene over all cells are normalized to 1,000,000, and normalized counts are shown on the same color scale in *B*, *C*, and *D*.

reference annotation assigns each cell to exactly one type, we similarly assigned each Xenium cell to one of 20 latent factors by finding which latent factor contributes the most to its inferred expression; that is, Xenium cell i is assigned to latent factor $\arg\max_j P[i, j]$. We calculated the cosine similarity (Fig. 4A) of this clustering of cells to the cell type annotation provided by 10x Genomics, which contain 19 distinct cell types (besides "Unlabeled") and some annotations are more granular than others. In addition, we found that the clustering of Xenium cells by SIID is more similar to the 10x Genomics annotation (adjusted Rand index [ARI] = 0.460) than the Leiden clustering (with the same number of clusters) is to the 10x annotations (ARI = 0.347).

SIID recovers spatial distribution of deconvolved cell types

SIID's unsupervised deconvolution of Visium spots are visually similar to the supervised deconvolution performed by RCTD (Cable et al. 2022), a leading method for deconvolving admixed SRT data (Fig. 4B,C). For SIID, we trained the model on the BRCA data set (with Xenium and Visium data) and obtained the unsupervised latent factor deconvolution for each Visium spot by computing $M^T P$ and normalizing by each spot. We ran RCTD (Cable et al. 2022) with Visium and scRNA-seq data annotated with cell types to obtain a supervised cell type deconvolution of Visium spots (cosine similarity plot in Supplemental Fig. S8B). We observed that the spatial distribution of deconvolved DCIS_1 cell type from RCTD (Fig. 4C) and latent factor 18 from SIID (Fig. 4B) have strong spatial agreement. We emphasize that other imputation tools such as Tangram are not capable of determining cell types for spatial location.

SIID imputation leads to stromal cell subtype discovery in breast cancer Xenium data

We used SIID to impute gene expression in the breast cancer Xenium data set and used the imputed expression data to decipher cell subtypes within the subpopulation of 41,422 cells (out of 167,780 total cells in the data set) that were annotated as stromal cells by Janesick et al. (2023). The imputed gene expression from SIID improves clustering (Supplemental Methods C.7; Supplemental Fig. S13) and led to identification of four stromal cell subtypes which we labeled as inflammatory cancer associated fibroblasts (iCAF), myofibroblast-like CAF (myCAF), immune-stromal niche, and adipocyte-rich stroma (Fig. 5A). These annotations were inferred from chosen gene signatures that were previously associated with the corresponding cell type. Specifically, we used curated gene signatures (Wolf et al. 2018) to score each Xenium spot and assigned a label to the cluster that showed the strongest agreement with the signature (Fig. 5B; Supplemental Fig. S15).

The first cluster exhibits upregulation of *CXCL12, PTGDS, KCNN3*, suggesting *CXCL12*-high fibroblasts in the immune infiltrated stroma within the tumor microenvironment (TME). CAF-secreted *CXCL12* (Ahirwar et al. 2018) has been shown to be a major driver within the TME that interacts with the endothelial cells and promotes vascularization. *PTGDS* (prostaglandin D2 synthase) (Jiang et al. 2020; Forsthuber et al. 2024) and *KCNN3*, encoding a potassium channel that regulates the interactions between the immune system with the CAFs (Mohr et al. 2019), are consistent with these immune-interactive stromal cells. Consequently, we refer to this cluster as inflammatory CAFs (iCAFs). The second cluster exhibits upregulation of *ACTA2* (alpha smooth muscle actin), which is an indication of CAFs (Hu et al. 2022). More specifically, the presence of *TAGLN* (transgelin), and *COL13A1* (collagens) points to matrix-remodeling of CAF subtype (Santi et al. 2018), and we therefore annotated this cluster as myfibroblast-like CAF or myCAF. The third cluster shows a *RORC*-driven (Cinnamon et al. 2024), interferon-inducible program with upregulated *MX1, KLF9* (Zhang et al. 2020), and *AQP3* (Charlestin et al. 2022), suggesting an active Th17 (a subset of helper T cells) effector program. The coexistence of *CXCL12*-high CAFs and immune-rich stroma indicates these spots represent a mixture rather than pure stromal cell types. Due to this hybrid nature, we refer to this cluster as an immune–stromal niche. Finally, spots in the fourth cluster coexpress adipocyte markers (*ADIPOQ, PLIN1, CIDEC*) reflecting adipocyte-rich stroma that lead to microvasculature in breast cancer (Zhu et al. 2022). Thus, we refer to this cluster as adipocyte-rich stroma. Spatial colocalization (Fig. 5A) of these cell subtypes shows a clear pattern of stromal subtypes where myCAF cells localize in the interior of iCAFs that constitute the majority of the tissue.

Table 2. Comparison of average holdout R^2 scores using scRNA-Xenium pairing

Data set	SIID (Visium)	Tangram (scRNA, cell)	Tangram (scRNA, cluster)
BRCA	**0.2527**	0.2195	0.2203
CRC	**0.2248**	0.196	0.1822

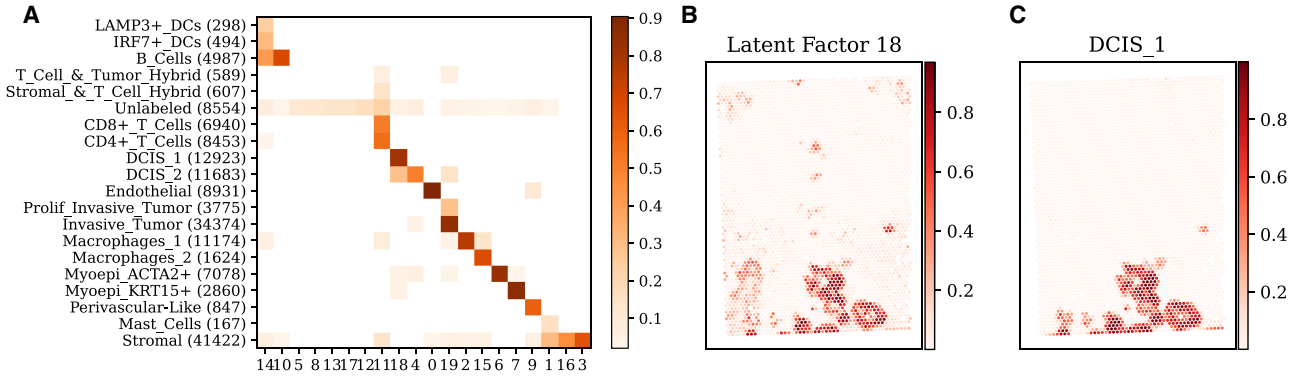

Figure 4. Analysis of SIID deconvolution on breast cancer (BRCA) data. (*A*) Comparison of the clustering of Xenium cells obtained by SIID on the BRCA data set with the cell type annotation provided by 10x Genomics. Each row corresponds to an annotated cell type, with the number of Xenium cells belonging to that type in parentheses. Each column corresponds to a latent factor derived by SIID. Intensity of each grid point is the cosine similarity between the assignments, with similarity scores below 0.02 excluded for visual clarity. (*B*) Spatial distribution of a latent factor inferred by SIID. (*C*) Spatial distribution of deconvolved DCIS_1 cell type from RCTD.

Further, we analyzed spatial colocalization of these stromal subtypes with the stromal and T cell hybrid cluster (Supplemental Fig. S13) and observed consistent spatial co-occurrence (Supplemental Methods C.7; Supplemental Fig. S14).

Discussion

SRT technologies continue to evolve, with new techniques to measure gene expression in physical space with varying spatial resolution, gene panel size, sequencing depth, processing time, and cost. Selecting the most appropriate SRT protocol for profiling a target sample can be challenging. One alternative is to use multiple SRT methods to profile adjacent or nearby slices from the same tissue and combine gene expression information across slices. This approach has the potential to mitigate the limitations of individual modalities. However, most existing methods for integrating data across SRT slices either focus on integration without spatial information—essentially treating SRT data sets as scRNA-seq data sets—or are designed for integrating SRT slices from the same modality and largely ignore potential differences between modalities.

SIID is a new approach of integrating SRT data sets that uses spatial information during integration and simultaneously performs imputation on a high spatial resolution targeted SRT data set and deconvolution on a low spatial resolution SRT data set. We demonstrate SIID on 10x Genomics Xenium and Visium data sets where we infer a latent single-cell whole transcriptome SRT data set that simultaneously imputes the Xenium data to whole transcriptome, recovers gene expression of the cell types, and deconvolves the Visium spots into cell type proportions. We evaluate SIID on both simulated and paired Xenium-Visium data sets where it shows strong performance in both imputation and deconvolution.

SIID will be useful in tissue atlas projects that aim to produce high-resolution multimodal reconstructions of tissues. Due to current costs of SRT platforms, multimodal tissue atlases are mostly large consortium projects. For example, the Human Tumor Atlas Network (HTAN) (de Bruijn et al. 2025) has been profiling tumors with multiple SRT technologies to evaluate tumor heterogeneity in 3D. Multiple SRT technologies are being used in this project; however, no single SRT technology provides both whole-transcriptome coverage and high spatial resolution. SIID enables the integration of two SRT data sets from adjacent tissue slices without using a nonspatial modality such as single-cell RNA sequencing. Thus, both deconvolution and imputation are performed on biologically similar transcriptomes, rather than relying on a transcriptome that may have been altered during single-cell RNA-seq preparation steps (Piwecka et al. 2023). SIID enables downstream analyses such as identifying differentially expressed genes across tissue slices measured with different SRT technologies. As costs of SRT technologies decrease, smaller projects may benefit from SIID's integration of complementary SRT data (such as 10x Visium and 10x Xenium) from the same tissue.

There are several limitations of SIID which are also directions for further improvement. First, SIID models platform-specific differences in expression solely through the platform scaling factors. Thus, our current analyses did not use out-of-Xenium-panel genes

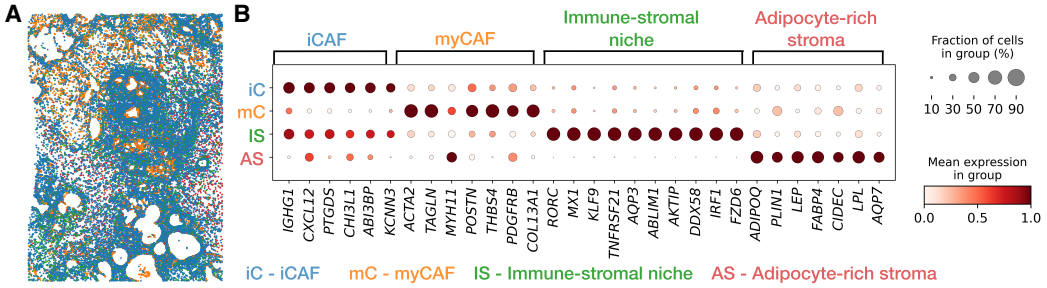

Figure 5. Stromal cell subtype analysis. (*A*) Spatial distribution of stromal cells in Xenium data with four annotated cell subtypes. (*B*) Marker gene expression for the annotated cell subtypes within the stromal cell type in Xenium data.

that are present in Visium data sets, effectively discarding a substantial fraction of Visium counts (Supplemental Methods C.2). Extending the loss function in SIID to account for the expression differences would enable better utilization of out-of-panel gene expression during inference. Second, our current method is reliant on an accurate spatial mapping matrix Γ, which is not always available. We further discuss the alignment procedure and potential sources of misalignment in Supplemental Methods C.1 and effects of misalignment in both simulation (Supplemental Methods B.3) and BRCA/CRC (Supplemental Methods C.5) data sets. Jointly inferring the spatial alignment Γ by extending existing SRT alignment algorithms (Zeira et al. 2022; Clifton et al. 2023; Liu et al. 2023) might yield better performance. Third, our method currently uses a Poisson count model but can easily be extended to other probability distributions. For example, a very common practice in modeling single-cell RNA-seq data is to assume overdispersion of observed counts. Towards this end, in Supplemental Methods C.8 we describe an extension of SIID that models Visium counts using negative binomial distribution with gene-specific overdispersion. An alternative approach is to model dropouts with zero-inflated distributions (Ridout et al. 2001; Yau et al. 2003); however, careful implementation is required to avoid overfitting the data given the large number of parameters. Fourth, SIID contains a number of hyperparameters whose values need to be selected. We selected the number of hidden dimensions h through exploratory data analysis with unsupervised clustering but found that SIID is robust with larger values of h (Supplemental Methods C.3, D.1). Analysis of hyperparameter values showed small variation in performance for varying regularization parameter λ and number of epochs. Multiple restarts provide a boost to performance and are always recommended (Supplemental Tables S3, S4). In future extensions, we will explore venues of automatically inferring these hyperparameters depending on the use case. Finally, SIID has thus far been tested only on 10x Genomics Xenium and Visium paired data due to the lack of publicly available paired data from other spatial platforms. Whereas the underlying algorithm of SIID is applicable to other imaging and sequencing-based spatial platforms, the performance of SIID on other platforms is not yet known.

SIID is a robust and interpretable method that is scalable and flexible to handle new SRT technologies, such as the recent Xenium 5K and Visium HD (Oliveira et al. 2025) that measure a large panel of genes at subcellular resolution. These and other high resolution data sets are typically more sparse than low resolution or small-panel data sets. Integration methods such as SIID can be adapted to improve coverage of these sparse data sets, allowing for accurate characterization of the transcriptome of the target tissue.

Methods

Representing paired spatially resolved transcriptomics data

We represent a spatially resolved transcriptomics data set by a pair (A, S), where $A \in \mathbb{N}^{|S|\times|G|}$ is a gene count matrix and $S \in \mathbb{R}^{|S|\times 2}$ is the two-dimensional physical coordinates for each spatial location and where G represents the set of genes measured in the SRT experiment. Suppose we are given two SRT tissue slices (A_X, S_X) and (A_V, S_V) from the same tissue measured with two different protocols which we denote by X and V. We assume that X has higher spatial resolution and a limited set of genes (small-panel), whereas V has lower spatial resolution but contains a superset of measured genes (large-panel), meaning $|S_X| \gg |S_V|$ and $G_X \subset G_V$. In this manuscript, X and V originate from 10x Xenium and 10x Visium platforms, respectively.

We further assume that there exists an alignment between (A_X, S_X) and (A_V, S_V); that is, there is a spot-spot correspondence between two slices as computed from any method that aligns two SRT data sets (Zeira et al. 2022; Clifton et al. 2023; Jones et al. 2023; Liu et al. 2023; Li et al. 2024; Tang et al. 2024). Given the spot-spot correspondence, the coordinates S_X and S_V can be transformed to represent locations in a shared coordinate system between the SRT data sets. After such spatial transformation, the spot-to-spot correspondence matrix is represented with a binary mapping matrix $\Gamma \in \{0, 1\}^{|S_X|\times|S_V|}$, where $\Gamma[i, j] = 1$ if and only if the Xenium spot i is mapped to the Visium spot j. Our proposed framework works with any arbitrary Γ.

A shared cell type model between paired SRTs

We assume that A_X and A_V are sampled from a latent gene count matrix A_U based on the following assumptions:

- There exists a latent SRT data set U represented with (A_U, S_X), where $A_U \in \mathbb{N}^{|S_X|\times|G_V|}$ is the latent gene expression matrix. By construction, U has the same spatial coordinates as X and the same gene set as V.
- A_U follows a Poisson distribution with mean $\overline{A_U}$: $A_U \sim \text{Pois}(\overline{A_U})$. As A_X is a submatrix of A_U, it naturally follows that $A_X \sim \text{Pois}(\overline{A_X})$, where $\overline{A_X}$ is a submatrix of $\overline{A_U}$ with columns restricted to G_X.
- A_V follows a Poisson distribution with mean $\overline{A_V} = K^T\overline{A_U}$, where $K \in \mathbb{R}^{|S_X|\times|S_V|}_{\geq 0}$ and $K \circ (1-\Gamma) = 0$, where $\circ$ represents the Hadamard product. $K[i, j]$ represents the weight of contribution of a Xenium spot i to construct the Visium spot j and is zero wherever $\Gamma[i, j] = 0$.
- $\overline{A_U}$ has a low-dimensional nonnegative matrix factorization, which we denote as $\overline{A_U} = WH$, with $W \in \mathbb{R}^{|S_X|\times h}_{\geq 0}$ and $H \in \mathbb{R}^{h\times|G_V|}_{\geq 0}$, where h is the number of latent factors. This implies A_X has Poisson mean $\overline{A_X} = WH_X$, where H_X is a submatrix of H with columns restricted to G_X, and A_V has Poisson mean $\overline{A_V} = K^TWH$.

With these assumptions, we formally state our inference problem as follows.

Paired NMF Inference Problem

Given a pair of SRT slices $(A_X \in \mathbb{N}^{|S_X|\times|G_X|}, S_X \in \mathbb{R}^{|S_X|\times 2})$, $(A_V \in \mathbb{N}^{|S_V|\times|G_V|}, S_V \in \mathbb{R}^{|S_V|\times 2})$, and spatial mapping matrix $\Gamma \in \{0, 1\}^{|S_X|\times|S_V|}$, find nonnegative matrices $W \in \mathbb{R}^{|S_X|\times h}_{\geq 0}$ $H \in \mathbb{R}^{h\times|G_V|}_{\geq 0}$, and $K \in \mathbb{R}^{|S_X|\times|S_V|}_{\geq 0}$ that solve the following problem:

$$\begin{aligned} &\min_{W,H,K} \text{PoiLoss}(A_X; WH_X) + \text{PoiLoss}(A_V; K^TWH) \\ &\text{subject to } W, H, K \geq 0 \text{ and } K \circ (1-\Gamma) = 0. \end{aligned} \tag{1}$$

Here, K corresponds to mixture weights matrix, where $K[i, j]$ represents the contribution of Xenium spot i to the Visium spot j and $\text{PoiLoss}(Y; Z) = \sum(Z - Y\log Z)$ is the negative log-likelihood for observing $Y \sim \text{Pois}(Z)$.

Count-scaled reparameterization

For numerical stability and better interpretation, we solve a reparameterized version of the Paired NMF Inference Problem. Recall $\overline{A_U}$ has a low-dimensional NMF $\overline{A_U} = WH$, from which we derive estimated expression $\overline{A_X}$ and $\overline{A_V}$. We rewrite this factorization with an alternative formulation where $\overline{A_U} = \text{diag}(N)PQ$ and where

- $N \in \mathbb{R}_{\geq 0}^{|S_X|}$ is a vector where $N[i]$ is the inferred total gene counts for spot i in U, and $\text{diag}(N) \in \mathbb{R}_{\geq 0}^{|S_X| \times |S_X|}$ is a matrix whose diagonal elements are N,
- $P \in \mathbb{R}_{\geq 0}^{|S_X| \times h}$, where each row $P[i]$ is the normalized latent factor composition of spot i in U, that is, $P\mathbb{1}_h = \mathbb{1}_{|S_X|}$,
- $Q \in \mathbb{R}_{\geq 0}^{h \times |G_V|}$, where each row $Q[j]$ is the normalized expression of latent factor j in G_V, that is, $Q\mathbb{1}_{|G_V|} = \mathbb{1}_h$.

We first show that the reparameterized problem is equivalent to the original.

Lemma 1. Given $\overline{A_U} = WH$, there exists P, Q, N as defined above such that $\overline{A_U} = \text{diag}(N)PQ$, and vice versa.

Proof (see Supplemental Methods A).

From $\overline{A_U} = \text{diag}(N)PQ$, we have $\overline{A_X} = \text{diag}(N)PQ_X$, where Q_X is a submatrix of Q with columns restricted to G_X . We next reparameterize K, the mixture weight. We define $M = \text{diag}(N)K$, and thus, $\overline{A_V} = K^T\text{diag}(N)PQ = M^TPQ$. M replaces K as the variable for inference and has identical constraints of K: $M \geq 0$, $M \circ (1 - \Gamma) = 0$. The reparameterized version is formally stated as follows.

Reparameterized Paired NMF Inference

Given a pair of SRT slices $(A_X \in \mathbb{N}^{|S_X| \times |G_X|}, S_X \in \mathbb{R}^{|S_X| \times 2})$, $(A_V \in \mathbb{N}^{|S_V| \times |G_V|}, S_V \in \mathbb{R}^{|S_V| \times 2})$, and spatial mapping matrix $\Gamma \in \{0, 1\}^{|S_X| \times |S_V|}$, find nonnegative matrices $P \in \mathbb{R}_{\geq 0}^{|S_X| \times h}$, $Q \in \mathbb{R}_{\geq 0}^{h \times |G_V|}$, $N \in \mathbb{R}_{\geq 0}^{|S_X|}$ and $M \in \mathbb{R}_{\geq 0}^{|S_X| \times |S_V|}$ that solve the following problem:

$$\min_{P,Q,N,M} \text{PoiLoss}(A_X;\ \text{diag}(N)PQ_X) + \text{PoiLoss}(A_V; M^TPQ)$$
$$\text{subject to } P, Q, N, M \geq 0,\ P\mathbb{1}_h = \mathbb{1}_{|S_X|},\ Q\mathbb{1}_{|G_V|} = \mathbb{1}_h,\ M \circ (1 - \Gamma) = 0 \quad (2)$$

M corresponds to the mixture weights, with $M[i, j]$ being the contribution of gene counts from Xenium spot i to Visium spot j, and, as above, $\text{PoiLoss}(Y; Z) = \sum(Z - Y\log Z)$ is the negative log-likelihood for observing $Y \sim \text{Pois}(Z)$.

The reparameterized problem allows easier interpretation of the inferred model.

Implementation, parameter inference, and imputing missing genes

We implement SIID to solve the Reparameterized Paired NMF Inference problem (Eq. 2). SIID solves the optimization problem in PyTorch (https://pytorch.org/) using gradient descent to optimize the model parameters. We use Adam optimizer and train for 5000 epochs, with a learning rate of 0.05.

Parameters of the model and ℓ_2 regularization

Because P and Q are row-normalized, we represent them as softmaxed matrices. Similarly, N, M are represented as exponentiated matrices. For numerical stability reasons, we place a ℓ_2 regularization with weight 10^{-5} on the parameters of the model (P, Q before softmax, N, M before exponentiation) (Biancalani et al. 2021).

Platform scaling

Our model assumes existence of a latent gene expression matrix A_U that, in turn, generates observations A_X and A_V. In practice, if the two SRT data sets are generated from different SRT platforms, we expect platform-specific effects that should be modeled on top of the latent gene expression. Observing that the same gene could be expressed at different rates in different SRT data sets (Cable et al. 2022), we introduce a gene-wise scaling $\phi \in \mathbb{R}_{\geq 0}^{|G_V|}$, as a multiplier to the columns (corresponding to genes) of $\overline{A_V}$ (to reduce nonidentifiability in practice, we fix $\phi[j] = 1$ if $j \notin G_X$). More specifically, when platform scaling is enabled, the loss function of the Reparameterized Paired NMF Inference (Eq. 2) is updated to $\text{PoiLoss}(A_X; \text{diag}(N)PQ_X) + \text{PoiLoss}(A_V; M^TPQ\text{diag}(\phi))$.

Entropy regularization

To assign the spots in X into distinct cell types, we optionally add an entropy regularization term of $\mathcal{H} = -\omega\sum(P\log P)$ to the loss function (2) with increasing weight $\omega = \exp(k/\lambda)$ across training epochs, where k is the current epoch and λ is a hyperparameter. Lower entropy encourages each Xenium spot to be predominantly assigned to a single latent factor or cell type, rather than being evenly distributed across multiple types. In general, when the focus is imputing missing genes, we suggest a larger value of λ such that the Poisson losses are the dominant terms. When the focus is the deconvolution of admixed spots, we suggest a smaller value of λ such that entropy regularization becomes more dominant near the end of the training process. We use $\lambda = 500$ in simulation and deconvolution on BRCA and CRC data sets (Results, "SIID recovers the cell type expression in simulated data set"; Results, "Deconvolving cell types in cancer SRT data") and $\lambda = 1000$ for imputation on BRCA and CRC data sets (Results, "Imputing missing genes in cancer SRT data"; Results, "SIID imputation leads to stromal cell subtype discovery in breast cancer Xenium data").

Random restarts

As NMFs are known to be sensitive to initializations (Gaujoux and Seoighe 2010; Fathi Hafshejani and Moaberfard 2023), we always restart our model three times and use the one with the best loss value.

Imputing missing genes

To impute the expression of a missing or holdout gene g on X, the model is trained with $g \notin G_X$ and $g \in G_V$. With trained parameters P, Q, and N, the imputed expression for g is $\text{diag}(N)PQ_g$, where $Q_g \in \mathbb{R}_{\geq 0}^h$ is the column of Q corresponding to g. Multiple genes can be imputed in the same run.

Software availability

A PyTorch implementation of SIID is available at GitHub (https://github.com/raphael-group/siid). The repository also includes installation guides and several example Jupyter notebooks for potential users to get started. In addition, a snapshot of this repository is available as Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This research is supported by National Cancer Institute (NCI) grants U24CA248453 and U24CA264027 to B.J.R. and by the Princeton Ludwig Branch.

Author contributions: H.Z. and B.J.R. conceived the project and developed the initial approach. H.Z and H.S. implemented the computational methods, designed the benchmarks, and conducted the experiments. All authors wrote and reviewed the manuscript.

References

Abdelaal T, Mourragui S, Mahfouz A, Reinders MJ. 2020. SpaGE: spatial gene enhancement using scRNA-seq. *Nucleic Acids Res* **48:** e107. doi:10.1093/nar/gkaa740

Ahirwar DK, Nasser MW, Ouseph MM, Elbaz M, Cuitiño MC, Kladney RD, Varikuti S, Kaul K, Satoskar AR, Ramaswamy B, et al. 2018. Fibroblast-derived CXCL12 promotes breast cancer metastasis by facilitating tumor cell intravasation. *Oncogene* **37:** 4428–4442. doi:10.1038/s41388-018-0263-7

Andersson A, Bergenstråhle J, Asp M, Bergenstråhle L, Jurek A, Fernández Navarro J, Lundeberg J. 2020. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun Biol* **3:** 565. doi:10.1038/s42003-020-01247-y

Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O. 2018. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* **14:** e8124. doi:10.15252/msb.20178124

Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, Stegle O. 2020. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol* **21:** 111. doi:10.1186/s13059-020-02015-1

Biancalani T, Scalia G, Buffoni L, Avasthi R, Lu Z, Sanger A, Tokcan N, Vanderburg CR, Segerstolpe Å, Zhang M, et al. 2021. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat Methods* **18:** 1352–1362. doi:10.1038/s41592-021-01264-7

Cable DM, Murray E, Zou LS, Goeva A, Macosko EZ, Chen F, Irizarry RA. 2022. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat Biotechnol* **40:** 517–526. doi:10.1038/s41587-021-00830-w

Cang Z, Nie Q. 2020. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nat Commun* **11:** 2084. doi:10.1038/s41467-020-15968-5

Cao Y, Zhao X, Tang S, Jiang Q, Li S, Li S, Chen S. 2024. ScButterfly: a versatile single-cell cross-modality translation method via dual-aligned variational autoencoders. *Nat Commun* **15:** 2973. doi:10.1038/s41467-024-47418-x

Charlestin V, Fulkerson D, Arias Matus CE, Walker ZT, Carthy K, Littlepage LE. 2022. Aquaporins: new players in breast cancer progression and treatment response. *Front Oncol* **12:** 988119. doi:10.3389/fonc.2022.988119

Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. 2015. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348:** aaa6090. doi:10.1126/science.aaa6090

Chidester B, Zhou T, Alam S, Ma J. 2023. SpiceMix enables integrative single-cell spatial modeling of cell identity. *Nat Genet* **55:** 78–88. doi:10.1038/s41588-022-01256-z

Cinnamon E, Stein I, Zino E, Rabinovich S, Shovman Y, Schlesinger Y, Salame T-M, Reich-Zeliger S, Albrecht T, Roessler S, et al. 2024. RORc-expressing immune cells negatively regulate tertiary lymphoid structure formation and support their pro-tumorigenic functions. *J Hepatol* **82:** 1050–1067. doi:10.1016/j.jhep.2024.12.015

Clifton K, Anant M, Aihara G, Atta L, Aimiuwu OK, Kebschull JM, Miller MI, Tward D, Fan J. 2023. STalign: alignment of spatial transcriptomics data using diffeomorphic metric mapping. *Nat Commun* **14:** 8123. doi:10.1038/s41467-023-43915-7

Cohen Kalafut N, Huang X, Wang D. 2023. Joint variational autoencoders for multimodal imputation and embedding. *Nat Mach Intell* **5:** 631–642. doi:10.1038/s42256-023-00663-z

de Bruijn I, Nikolov M, Lau C, Clayton A, Gibbs DL, Mitraka E, Pozhidayeva D, Lash A, Sumer SO, Altreuter J, et al. 2025. Sharing data from the Human Tumor Atlas Network through standards, infrastructure and community engagement. *Nat Methods* **22:** 664–671. doi:10.1038/s41592-025-02643-0

De Coninck S, Berx G, Taghon T, Van Vlierberghe P, Goossens S. 2019. ZEB2 in T-cells and T-ALL. *Adv Biol Regul* **74:** 100639. doi:10.1016/j.jbior.2019.100639

di Gennaro A, Damiano V, Brisotto G, Armellin M, Perin T, Zucchetto A, Guardascione M, Spaink HP, Doglioni C, Snaar-Jagalska BE, et al. 2018. A p53/miR-30a/ZEB2 axis controls triple negative breast cancer aggressiveness. *Cell Death Differ* **25:** 2165–2180. doi:10.1038/s41418-018-0103-x

Fathi Hafshejani S, Moaberfard Z. 2023. Initialization for non-negative matrix factorization: a comprehensive review. *Int J Data Sci Anal* **16:** 119–134. doi:10.1007/s41060-022-00370-9

Forsthuber A, Aschenbrenner B, Korosec A, Jacob T, Annusver K, Krajic N, Kholodniuk D, Frech S, Zhu S, Purkhauser K, et al. 2024. Cancer-associated fibroblast subtypes modulate the tumor-immune microenvironment and are associated with skin cancer malignancy. *Nat Commun* **15:** 9678. doi:10.1038/s41467-024-53908-9

Gaujoux R, Seoighe C. 2010. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11:** 367. doi:10.1186/1471-2105-11-367

Hu D, Li Z, Zheng B, Lin X, Pan Y, Gong P, Zhuo W, Hu Y, Chen C, Chen L, et al. 2022. Cancer-associated fibroblasts in breast cancer: challenges and opportunities. *Cancer Commun* **42:** 401–434. doi:10.1002/cac2.12291

Jackson KC, Booeshaghi AS, Gálvez-Merchán Á, Moses L, Chari T, Kim A, Pachter L. 2024. Identification of spatial homogeneous regions in tissues with concordex. bioRxiv doi:10.1101/2023.06.28.546949

Janesick A, Shelansky R, Gottscho AD, Wagner F, Williams SR, Rouault M, Beliakoff G, Morrison CA, Oliveira MF, Sicherman JT, et al. 2023. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nat Commun* **14:** 8353. doi:10.1038/s41467-023-43458-x

Jiang J, Pan W, Xu Y, Ni C, Xue D, Chen Z, Chen W, Huang J. 2020. Tumour-infiltrating immune cell-based subtyping and signature gene analysis in breast cancer based on gene expression profiles. *J Cancer* **11:** 1568–1583. doi:10.7150/jca.37637

Jones A, Townes FW, Li D, Engelhardt BE. 2023. Alignment of spatial genomics data using deep Gaussian processes. *Nat Methods* **20:** 1379–1387. doi:10.1038/s41592-023-01972-2

Ke R, Mignardi M, Pacureanu A, Svedlund J, Botling J, Wählby C, Nilsson M. 2013. In situ sequencing for RNA analysis in preserved tissue and cells. *Nat Methods* **10:** 857–860. doi:10.1038/nmeth.2563

Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh P-R, Raychaudhuri S. 2019. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16:** 1289–1296. doi:10.1038/s41592-019-0619-0

Li B, Zhang W, Guo C, Xu H, Li L, Fang M, Hu Y, Zhang X, Yao X, Tang M, et al. 2022. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nat Methods* **19:** 662–670. doi:10.1038/s41592-022-01480-9

Li H, Lin Y, He W, Han W, Xu X, Xu C, Gao E, Zhao H, Gao X. 2024. SANTO: a coarse-to-fine alignment and stitching method for spatial omics. *Nat Commun* **15:** 6048. doi:10.1038/s41467-024-50308-x

Liu J, Gao C, Sodicoff J, Kozareva V, Macosko EZ, Welch JD. 2020. Jointly defining cell types from multiple single-cell datasets using LIGER. *Nat Protoc* **15:** 3632–3662. doi:10.1038/s41596-020-0391-8

Liu X, Zeira R, Raphael BJ. 2023. Partial alignment of multislice spatially resolved transcriptomics data. *Genome Res* **33:** 1124–1132. doi:10.1101/gr.277670.123

Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. 2018. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15:** 1053–1058. doi:10.1038/s41592-018-0229-2

Lopez R, Nazaret A, Langevin M, Samaran J, Regier J, Jordan MI, Yosef N. 2019. A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. arXiv:1905.02269 [cs.LG]. doi:10.48550/arXiv.1905.02269

Miller BF, Huang F, Atta L, Sahoo A, Fan J. 2022. Reference-free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data. *Nat Commun* **13:** 2339. doi:10.1038/s41467-022-30033-z

Mohr CJ, Steudel FA, Gross D, Ruth P, Lo W-Y, Hoppe R, Schroth W, Brauch H, Huber SM, Lukowski R. 2019. Cancer-associated intermediate conductance Ca^{2+}-activated K^+ channel KCa3.1. *Cancers (Basel)* **11:** 109. doi:10.3390/cancers11010109

Moses L, Pachter L. 2022. Museum of spatial transcriptomics. *Nat Methods* **19:** 534–546. doi:10.1038/s41592-022-01409-2

Oliveira MF, Romero JP, Chung M, Williams S, Gottscho AD, Gupta A, Pilipauskas SE, Mohabbat S, Raman N, Sukovich D, et al. 2025. High-definition spatial transcriptomic profiling of immune cell populations in colorectal cancer. *Nat Genet* **57:** 1512–1523. doi:10.1038/s41588-025-02193-3

Palla G, Fischer DS, Regev A, Theis FJ. 2022. Spatial components of molecular tissue biology. *Nat Biotechnol* **40:** 308–318. doi:10.1038/s41587-021-01182-1

Piwecka M, Rajewsky N, Rybak-Wolf A. 2023. Single-cell and spatial transcriptomics: deciphering brain complexity in health and disease. *Nat Rev Neurol* **19:** 346–362. doi:10.1038/s41582-023-00809-y

Qian K, Fu S, Li H, Li WV. 2022. scINSIGHT for interpreting single-cell gene expression from biologically heterogeneous data. *Genome Biol* **23:** 82. doi:10.1186/s13059-022-02649-3

Rao A, Barkley D, França GS, Yanai I. 2021. Exploring tissue architecture using spatial transcriptomics. *Nature* **596:** 211–220. doi:10.1038/s41586-021-03634-9

Ridout M, Hinde J, Demétrio CG. 2001. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics* **57:** 219–223. doi:10.1111/j.0006-341X.2001.00219.x

Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, Welch J, Chen LM, Chen F, Macosko EZ. 2019. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363:** 1463–1467. doi:10.1126/science.aaw1219

Rozenblatt-Rosen O, Regev A, Oberdoerffer P, Nawy T, Hupalowska A, Rood JE, Ashenberg O, Cerami E, Coffey RJ, Demir E, et al. 2020. The Human Tumor Atlas Network: charting tumor transitions across space and time at single-cell resolution. *Cell* **181:** 236–249. doi:10.1016/j.cell.2020.03.053

Santi A, Kugeratski FG, Zanivan S. 2018. Cancer associated fibroblasts: the architects of stroma remodeling. *Proteomics* **18:** 1700167. doi:10.1002/pmic.201700167

Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, Giacomello S, Asp M, Westholm JO, Huss M, et al. 2016. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353:** 78–82. doi:10.1126/science.aaf2403

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, Satija R. 2019. Comprehensive integration of single-cell data. *Cell* **177:** 1888–1902.e21. doi:10.1016/j.cell.2019.05.031

Tang Z, Luo S, Zeng H, Huang J, Sui X, Wu M, Wang X. 2024. Search and match across spatial omics samples at single-cell resolution. *Nat Methods* **21:** 1818–1829. doi:10.1038/s41592-024-02410-7

Townes FW, Engelhardt BE. 2023. Nonnegative spatial factorization applied to spatial genomics. *Nat Methods* **20:** 229–238. doi:10.1038/s41592-022-01687-w

Townes FW, Hicks SC, Aryee MJ, Irizarry RA. 2019. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol* **20:** 295. doi:10.1186/s13059-019-1861-6

Vahid MR, Brown EL, Steen CB, Zhang W, Jeon HS, Kang M, Gentles AJ, Newman AM. 2023. High-resolution alignment of single-cell and spatial transcriptomes with CytoSPACE. *Nat Biotechnol* **41:** 1543–1548. doi:10.1038/s41587-023-01697-9

Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, Evans K, Liu C, Ramakrishnan C, Liu J, et al. 2018. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361:** eaat5691. doi:10.1126/science.aat5691

Wolf FA, Angerer P, Theis FJ. 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19:** 15. doi:10.1186/s13059-017-1382-0

Wu KE, Yost KE, Chang HY, Zou J. 2021. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. *Proc Natl Acad Sci* **118:** e2023070118. doi:10.1073/pnas.2023070118

Xia C-R, Cao Z-J, Tu X-M, Gao G. 2023. Spatial-linked alignment tool (SLAT) for aligning heterogenous slices. *Nat Commun* **14:** 7236. doi:10.1038/s41467-023-43105-5

Yang Z, Michailidis G. 2016. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* **32:** 1–8. doi:10.1093/bioinformatics/btv544

Yau KK, Wang K, Lee AH. 2003. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biomet J* **45:** 437–452. doi:10.1002/bimj.200390024

Zeira R, Land M, Strzalkowski A, Raphael BJ. 2022. Alignment and integration of spatial transcriptomics data. *Nat Methods* **19:** 567–575. doi:10.1038/s41592-022-01459-6

Zhang J, Li G, Feng L, Lu H, Wang X. 2020. Krüppel-like factors in breast cancer: function, regulation and clinical relevance. *Biomed Pharmacother* **123:** 109778. doi:10.1016/j.biopha.2019.109778

Zhu Q, Zhu Y, Hepler C, Zhang Q, Park J, Gliniak C, Henry GH, Crewe C, Bu D, Zhang Z, et al. 2022. Adipocyte mesenchymal transition contributes to mammary tumor progression. *Cell Rep* **40:** 111362. doi:10.1016/j.celrep.2022.111362

Received February 19, 2025; accepted in revised form October 8, 2025.

Joint imputation and deconvolution of gene expression across spatial transcriptomics platforms

Hongyu Zheng, Hirak Sarkar and Benjamin J. Raphael

Genome Res. published online November 17, 2025
Access the most recent version at doi:10.1101/gr.280555.125