



METHOD

Open Access



SDePER: a hybrid machine learning and regression method for cell-type deconvolution of spatial barcoding-based transcriptomic data

Yunqing Liu^{1†}, Ningshan Li^{1,2,3†}, Ji Qi¹, Gang Xu^{1,4}, Jiayi Zhao¹, Nating Wang¹, Xiayuan Huang¹, Wenhao Jiang¹, Huanhuan Wei^{1,5}, Aurélien Justet^{5,6}, Taylor S. Adams⁵, Robert Homer⁷, Amei Amei⁴, Ivan O. Rosas⁸, Naftali Kaminski⁵, Zuoheng Wang^{1,9*}  and Xiting Yan^{1,5*} 

[†]Yunqing Liu and Ningshan Li contributed equally to this work.

*Correspondence: zuoheng.wang@yale.edu; xiting.yan@yale.edu

¹ Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA

⁵ Section of Pulmonary, Critical Care and Sleep Medicine, Yale School of Medicine, New Haven, CT, USA

Full list of author information is available at the end of the article

Abstract

Spatial barcoding-based transcriptomic (ST) data require deconvolution for cellular-level downstream analysis. Here we present SDePER, a hybrid machine learning and regression method to deconvolve ST data using reference single-cell RNA sequencing (scRNA-seq) data. SDePER tackles platform effects between ST and scRNA-seq data, ensuring a linear relationship between them while addressing sparsity and spatial correlations in cell types across capture spots. SDePER estimates cell-type proportions, enabling enhanced resolution tissue mapping by imputing cell-type compositions and gene expressions at unmeasured locations. Applications to simulated data and four real datasets showed SDePER's superior accuracy and robustness over existing methods.

Background

Spatial transcriptomic technologies enabled measuring gene expression and physical locations of spots and/or cells simultaneously in intact tissues of various types in an unbiased and high-throughput way [1–4], providing unprecedented information to understand disease-associated changes. Specifically, the spatial barcoding-based (ST) technologies, such as Slide-seq [5], HDST [6], ST [4], and 10 × Genomics Visium, divide tissue into small capture spots and measure high-throughput gene expression levels unbiasedly for each spot with known physical location [4–10]. Depending on the size of capture spots, the measured expression profile is an average expression profile of cells of unknown types. Therefore, the corresponding data lacks single-cell resolution [11] and requires cell-type deconvolution to understand the cell-type composition and cell-type-specific gene expression in each spot.



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

One common way to deconvolve ST data is to use cell-type-specific expression profile from existing single-cell RNA sequencing (scRNA-seq) data of the same tissue type. Many methods have been developed [12–28], which can be divided into four categories: machine learning-based [20–22], regression-based [23–28], statistical modeling-based [12–15], and data mapping-based methods [16]. Benchmarking studies have been conducted to compare the performance of these methods [29–31].

Despite the success of current methods, the following three challenges have not been well addressed and, more importantly, no method addresses them simultaneously. First, systematic difference exists between scRNA-seq and ST data [12–15, 22, 24, 25] due to various technical factors, such as differences in protocols, reagents, platforms, or simply sequencing depths. This systematic difference, termed as platform effects [12], makes the relationship between ST data and cell-type-specific expression profiles from the reference scRNA-seq data non-linear and varying across different technologies. A few statistical model-based methods [12–15] consider the platform effects as multiplicative random or fixed effect. However, these methods were shown in a previous benchmarking study [29] to have comparable performance to methods that do not address platform effects, leaving it unclear whether platform effects were adequately addressed. DSTG and some of the data mapping-based methods implicitly addressed platform effects by embedding scRNA-seq or scRNA-seq-derived pseudo-spot data and real ST data into a common latent space. Second, among all cell types existed in the tissue, only a few cell types are present in each spot. For example, 38 different cell types were found in the scRNA-seq data of whole lung tissues (the IPF dataset in real data analyses). However, capture spots of the $10 \times$ Genomics Visium platform with a size of $\sim 55 \mu\text{m}$ contained only 2–10 cells per spot, demonstrating a sparse presentation of all cell types existed in the tissue. This sparsity was considered by RCTD, SPOTlight, DestVI, and SpatialDWLS but using subjective hard thresholding. Lastly, previous studies [24, 32, 33] have shown that cell-type composition of spots that are physically close in the tissue tend to be similar or correlated. Only CARD explicitly considered the across-spot spatial correlation of cell-type compositions.

To address all the aforementioned challenges, we propose a two-step hybrid machine learning and regression method, SDePER, that considers platform effects removal, spatial correlation, and sparsity (Fig. 1). In the first step, a conditional variational autoencoder (CVAE) [34] is used to adjust the ST and reference scRNA-seq data for platform effects removal. In the second step, a graph Laplacian regularized model (GLRM) is fitted to the adjusted ST data with consideration of the spatial correlation of cell-type compositions between neighboring spots and sparsity of present cell types per spot. Based on the estimated cell-type compositions, a random walk is performed to impute cell-type compositions and gene expression at unmeasured locations in a tissue map with enhanced resolution. We demonstrate the advantage of SDePER through extensive simulations and applications to four real datasets from various tissues, species, and technologies.

Results

SDePER—efficiently corrects for platform effects

We conducted simulations to evaluate the performance of SDePER and compared it to seven other deconvolution methods with the best performance based on previous benchmarking studies [11, 29–31]: RCTD [12], SpatialDWLS [26], cell2location

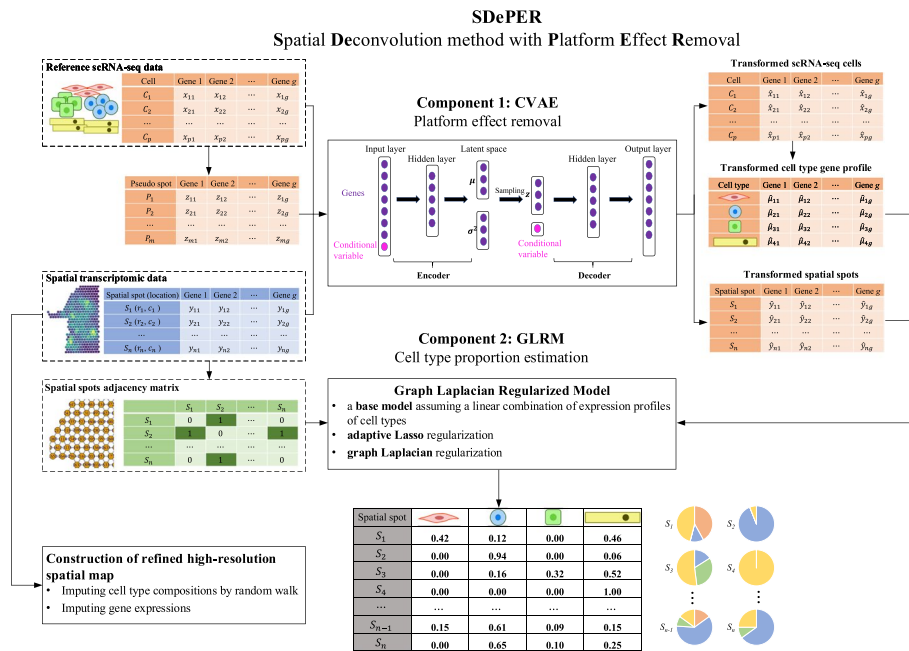


Fig. 1 Schematic overview of SDePER. SDePER performs cell-type deconvolution of ST data in a two-step fashion. In the first step, conditional variational autoencoder (CVAE) takes three datasets as input: real ST data, reference scRNA-seq data, and pseudo-spot data generated using the reference scRNA-seq data. Using the trained encoder and decoder under the two conditions (ST and scRNA-seq), real ST data is transformed into the same space as scRNA-seq data and pseudo-spot data. The transformed real ST data and cell type-specific expression profiles are then used to fit the graph Laplacian regularized model (GLRM) with penalties for sparsity and across-spot spatial correction in cell-type compositions. The estimated cell-type compositions from GLRM can be further used to impute for cell-type compositions and gene expression at unmeasured locations in the original spatial map to construct new spatial map at arbitrarily higher resolution

[15], SONAR [28], SPOTlight [25], CARD [24], and DestVI [14]. ST data with 581 spots was simulated by coarse-graining a real spatial transcriptomic data with single-cell resolution (Fig. 2A) generated using the STARmap technology [35]. The true cell-type composition at each simulated spot is calculated and serves as the ground truth. To demonstrate the impact of platform effects [12] on the method performance, each method was applied using both external and internal reference data, representing situations with and without platform effects. Moreover, to demonstrate the effectiveness of CVAE on removing platform effects, we ran SDePER with the CVAE component deactivated, which was named GLRM.

Performance comparison based on the median RMSE, Pearson correlation, and JSD showed that SDePER achieved the highest estimation accuracy regardless of the existence of platform effects (Fig. 2B, Additional File 1: Fig. S1). Visualization of the ground truth and estimated proportion of L2/L3 excitatory neurons (Fig. 2C) and other cell types (Additional File 1: Fig. S2) using an external reference further confirmed the highest accuracy of SDePER results (Pearson correlation=0.872). Furthermore, the accuracy of all methods was lower for external reference compared to internal reference, indicating that platform effects have a complicated form that cannot be efficiently addressed using a random effect. SDePER and DestVI had the smallest accuracy difference between internal and external reference, suggesting their best

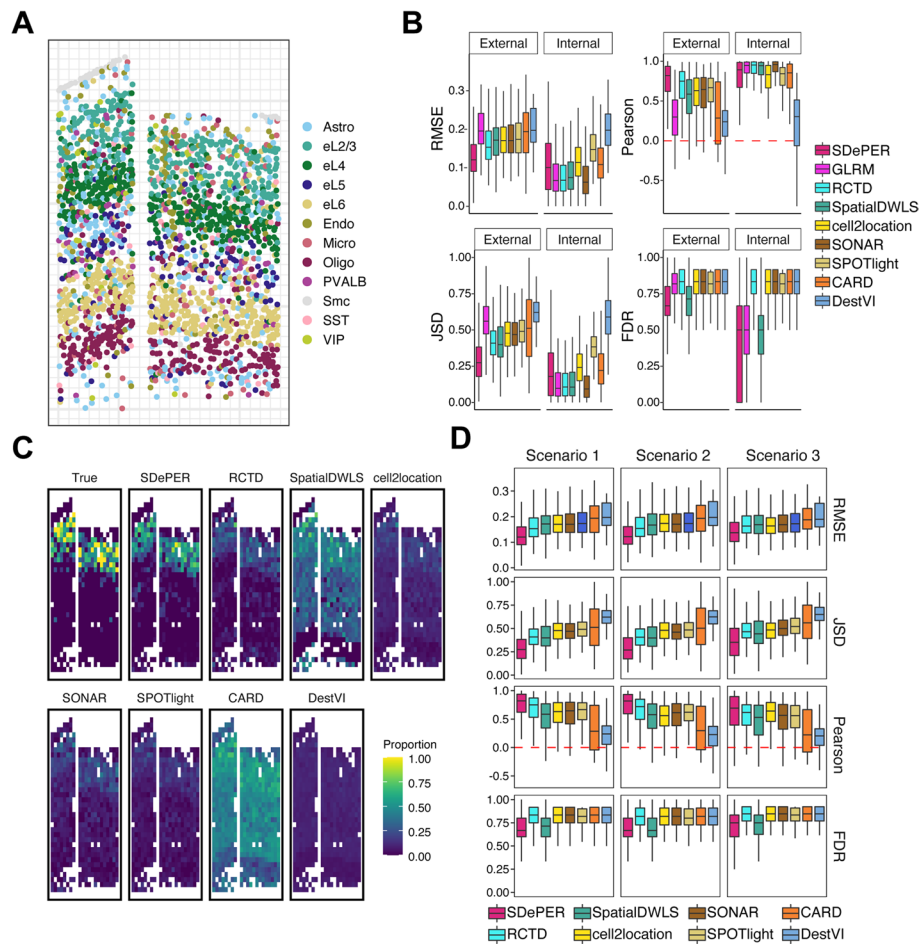


Fig. 2 Performance evaluation and comparison using simulation studies. **A** Coarse-graining procedure to simulate ST data (581 spots) with ground truth. **B** Demonstration of the impact of platform effects on method performance: boxplots show the median (center line), interquartile range (hinges), and 1.5 times the interquartile (whiskers) of RMSE, JSD, Pearson's correlation, and FDR across all 581 spots using external scRNA-seq reference and internal single-cell level spatial reference. **C** The proportion of L2/3 excitatory neurons in the simulated spots. **D** Boxplots show the median (center line), interquartile range (hinges), and 1.5 times the interquartile (whiskers) of RMSE, JSD, Pearson's correlation, and FDR across 581 spots using different scRNA-seq reference: scenario 1: scRNA-seq reference with matched cell type; scenario 2: one missing cell type in scRNA-seq reference; scenario 3: one added irrelevant cell type in scRNA-seq reference

robustness to platform effects (Fig. 2B, Additional File 1: Fig. S1). Lastly, when using internal reference without platform effects, SDePER had slightly worse performance than GLRM with an increase of 0.034 and 0.082 in the median RMSE and JSD, respectively, and a decrease of 0.056 in the median correlation, indicating the potential noise introduced by the CVAE component. But when platform effects were present (external reference), SDePER had a much better performance than GLRM (39%, 51%, 174%, 19% improvement in RMSE, JSD, Pearson's correlation, and FDR, respectively), and this increase was much larger than the decrease in performance when using internal reference (Additional File 1: Table S1). All these demonstrated that SDePER achieved the best performance in both estimating cell-type compositions and removing platform effects.

To demonstrate the performance of SDePER in datasets with both reference and ST data being purely sequencing-based for small platform effects, we generated another sequencing-based simulated ST data. We retained the spatial location of cells in the STARmap data but replaced the expression profiles of each cell with those of a randomly chosen cell of the same type from an independent scRNA-seq data. All methods achieved better performance than the STARmap-based simulated data as expected, and SDePER remained to have the best performance (Additional File 1: Fig. S3).

We noticed that certain tissues, like the solid cancer tissues, may have higher cell density than the STARmap data. Therefore, we conducted simulations for high cell density. Based on all performance criteria, SDePER had robust performance across all cell density settings while GLRM had decreasing performance when the cell density increased (Additional File 1: Fig. S4) even when there were no platform effects (internal reference). This suggests that ST data with higher cell density have larger systematic difference from the reference data, no matter whether there are platform effects or not, potentially because ST data with higher cell density tend to have more variation caused by the heterogeneity across cells from the same type. CVAE successfully addressed for this difference and made SDePER robust to cell density.

Ablation test

To understand the contribution of different components in SDePER, we conducted ablation tests by disabling each component. The results for external reference when using the STARmap-based simulated dataset (Additional File 1: Fig. S5) showed that CVAE had the most contribution and pseudo-spot inclusion in the CVAE training had the second largest contribution to the performance. Both the adaptive LASSO penalty and graph Laplacian penalty had negligible contribution to the RMSE but did contribute to lower the false discovery rate with the adaptive LASSO penalty having a slightly larger contribution. This reduction in FDR by the adaptive LASSO was also observed in simulation with only five cell types included in both ST and scRNA-seq data (Additional File 1: Fig. S6) indicating the necessity to include the LASSO penalty even when there are small number of cell types. For internal reference, i.e., when there are no platform effects, all components did not have noticeable contribution except that CVAE introduced noises and led to larger RMSE. When the sequencing-based simulated dataset was used (Additional File 1: Fig. S7), similar contribution was observed for each component but at a much smaller scale when the number of cells per spot is the same ($1 \times$) because the sequencing-based simulated data was expected to have smaller platform effects than the STARmap-based simulated data. But when the number of cells per spot increased, the contribution of CVAE and pseudo-spot inclusion significantly increased. This observation was consistent between the external and internal references.

Taken together, the ablation test showed that the most contributing component in SDePER is the CVAE component, which removes platform effects. It also showed that although this component did not help when there were no platform effects, it would have a large contribution when cell density is high.

Robustness of methods to mismatching cell types

To demonstrate the robustness of methods to mismatching cell types between the reference and ST data, we conducted deconvolution under three scenarios representing perfect match, one missing cell type, and one extra cell type in the external reference data compared to the ST data. The performance rankings of all methods were consistent across these three scenarios with SDePER consistently achieving the best accuracy (Fig. 2D). In scenario 1, the improvements in RMSE of SDePER compared to RCTD, SpatialDWLS, cell2location, SONAR, SPOTlight, CARD, and DestVI were 22%, 30%, 30%, 30%, 31%, 38%, and 39%, respectively. In scenarios 2 and 3, compared to the other methods, SDePER achieved 21–38% and 15–27% improvement in RMSE, respectively. Compared to scenario 1, SDePER also had an increase of 0.002 and 0.018 in the median RMSE in scenarios 2 and 3, respectively. These results showed that SDePER had the best robustness to the mismatching cell types between the ST data and reference scRNA-seq data.

Robustness of SDePER to rare cell types

Rare cell types in both reference scRNA-seq and ST data pose challenges to the task of deconvolution. To assess the robustness of SDePER to rare cell types, we conducted two simulation analyses for rare cell types. In the first analysis to simulate rare cell types in the reference data, the performance of SDePER on Oligodendrocytes was evaluated (Additional File 1: Fig. S8) using RMSE, false negative rate (FNR), and false discovery rate (FDR). For external reference, the performance is robust to the number of Oligodendrocytes in the down-sampled reference data. For internal reference, when there were more Oligodendrocytes, the median RMSE remained unchanged with shorter interquartile range suggesting a more stable result when there were more Oligodendrocytes. In the second simulation analysis for rare cell types in the ST data, within each group of spots with the same total number of cells, both the relative absolute error (RAE) and the false negative rate (FNR) decreased when the number of oligodendrocytes per spot increased (Additional File 1: Fig. S9). Specifically, when there were at least three cells in the spot, the rare cell type could be always identified as present (FNR=0). When there were only two cells in the spot, SDePER had over 87% chance to identify the rare cell type as present (FNR=0.125). This trend is consistent across spot groups with different total number of cells and between external and internal references. In summary, these results suggested that the performance of SDePER is robust to rare cell types in the reference scRNA-seq data but worse for the rare cell types in the ST data, especially when there are less than two cells in the spot.

Mouse olfactory bulb data

To demonstrate the efficacy of SDePER on real data, we first applied SDePER and the other seven methods to a ST data of mouse olfactory bulb (MOB) [4] with well-defined anatomic layers organized in a well-characterized spatial architecture. We took an independent scRNA-seq data of the same tissue type profiled using the $10 \times$ Genomics Chromium platform as reference data for the deconvolution [36] (Additional File 1: Fig. S10). Based on the H&E staining, four major tissue layers were identified from inside to outside with each dominantly composed of one cell type: the granule cell layer (GCL),

mitral cell layer (MCL), glomerular layer (GL), and olfactory nerve layer (ONL) dominated by GC, M/TC, PGC, and OSNs, respectively (Fig. 3A) [4, 24]. Expression maps of marker genes for these four dominant cell types were consistent with the four annotated layers (Fig. 3A). Expression maps of marker genes for other cell types can be found in Additional File 1: Fig. S11.

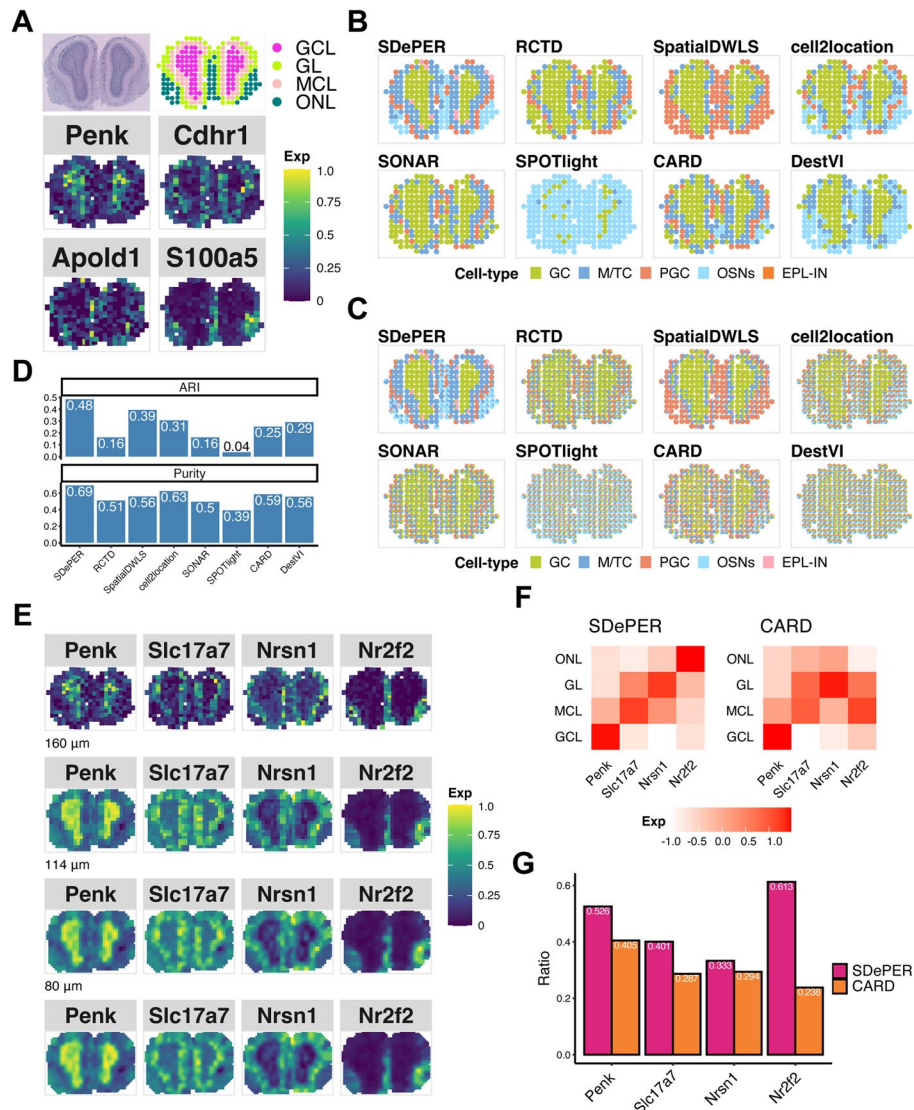


Fig. 3 Performance evaluation and comparison using MOB dataset. **A** H&E staining of MOB (top-left), annotated regions (top-right GCL: granule cell layer; MCL: mitral cell layer; GL: glomerular layer; ONL: olfactory nerve layer) and expression pattern of cell-type-specific marker genes for dominant cell types (bottom, *Penk* for GC, *Cdhr1* for mitral and tufted cell (M/TC), *Apold1* for periglomerular cell (PGC), and *S100a5* for olfactory sensory neurons (OSNs)). **B** Visualization of inferred dominant cell type in each spot (EPL-IN: external plexiform layer interneuron). **C** Spatial scatter pie chart of estimated cell-type composition within each spot. **D** Comparing deconvolution methods using ARI (left) and purity (right). **E** Expression patterns of the corresponding layer-specific marker genes and imputed expression at three different resolution levels: 160 μm (about 64% of original size), 114 μm (about 32% of original size), 80 μm (about 16% of original size). **F** Heatmap showing average imputed expression of region-specific marker genes at 80 μm level within each annotated region for SDePER and CARD. **G** Bar plot showing the ratio of average layer-specific marker gene expression in the corresponding layer among all layers

The H&E staining image-based annotation and expression maps of the four dominant cell-type marker genes were considered as ground truth. The predicted dominant cell type by SDePER showed remarkable similarity with the ground truth (Fig. 3A, B). RCTD and SONAR mislabeled ONL as GCL. SpatialDWLS and DestVI did not separate ONL and GL. CARD and cell2location showed blurry layer boundaries and did not find ONL accurately. SPOTlight failed to identify the annotated regions and identified almost all spots to be dominantly OSN, potentially due to the randomness and bias introduced by its cell down-sampling procedure. Quantitative assessment of the similarity between the predicted dominant cell type and H&E staining image-based annotated layers using ARI and purity (Fig. 3D) confirmed the best performance of SDePER. In addition, when comparing the predicted dominant cell type (Fig. 3B) to the predicted cell compositions in the pie chart (Fig. 3C) for each method, SDePER showed the highest similarity between the two plots indicating less non-specific cell-type detection potentially due to its sparsity regularization.

To demonstrate the imputation results, we selected four layer-specific marker genes, one gene for each layer, from the ST data. Visualization of the original and imputed layer-specific marker gene expression on the original spatial map and three spatial maps with higher resolution (Fig. 3E) showed an expression enrichment of each layer marker gene in its corresponding layer. We compared the imputed cell-type proportion and layer-specific marker gene expression on various resolutions with CARD (Additional File 1: Fig. S12-13). To quantitatively assess the expression enrichment, we calculated the average imputed expression levels of each layer marker gene in its corresponding layer at 80 μm resolution. The average imputed expression by SDePER (Fig. 3F) displayed higher diagonal values and lower off-diagonal values, indicating a better separation between different layers based on the imputed expression than CARD. We further calculated the ratio of the average expression of each layer-marker gene in its corresponding layer to that across all layers for a quantitative assessment of the imputed expression-based layer separation (Fig. 3G). SDePER achieved a higher ratio for all four layer-marker genes than CARD, demonstrating a higher accuracy in imputed expression.

Stage III cutaneous malignant melanoma data

The second real data we analyzed investigated the cutaneous malignant melanoma sample from the lymph nodes [7]. Manual annotation of the tissue slide using H&E staining and clustering analysis of the ST data (Fig. 4A) identified regions of melanoma, stroma, and lymphoid tissue with expected cell types [7, 33]. For each expected cell type in each region, we selected its marker genes from existing literature [37], which included *PMEL* for malignant cells in melanoma regions, *COL1A1* for fibroblast in stroma regions, *MS4A1* for B cells and *CD14* for macrophage in lymphoid tissues [37]. The expression map of these marker genes in ST data (Fig. 4A) confirmed the prevalence of fibroblasts in stroma regions, B cells in the right-top lymphoid tissue 1, and macrophages in the lymphoid tissue 2 surrounding the melanoma region. Expression maps of marker genes for other cell types can be found in Additional File 1: Fig. S14. We used an independent scRNA-seq data of untreated metastatic melanoma samples from human lymph nodes [38] profiled using the inDrop technology as the reference data for deconvolution (Additional File 1: Fig. S15).

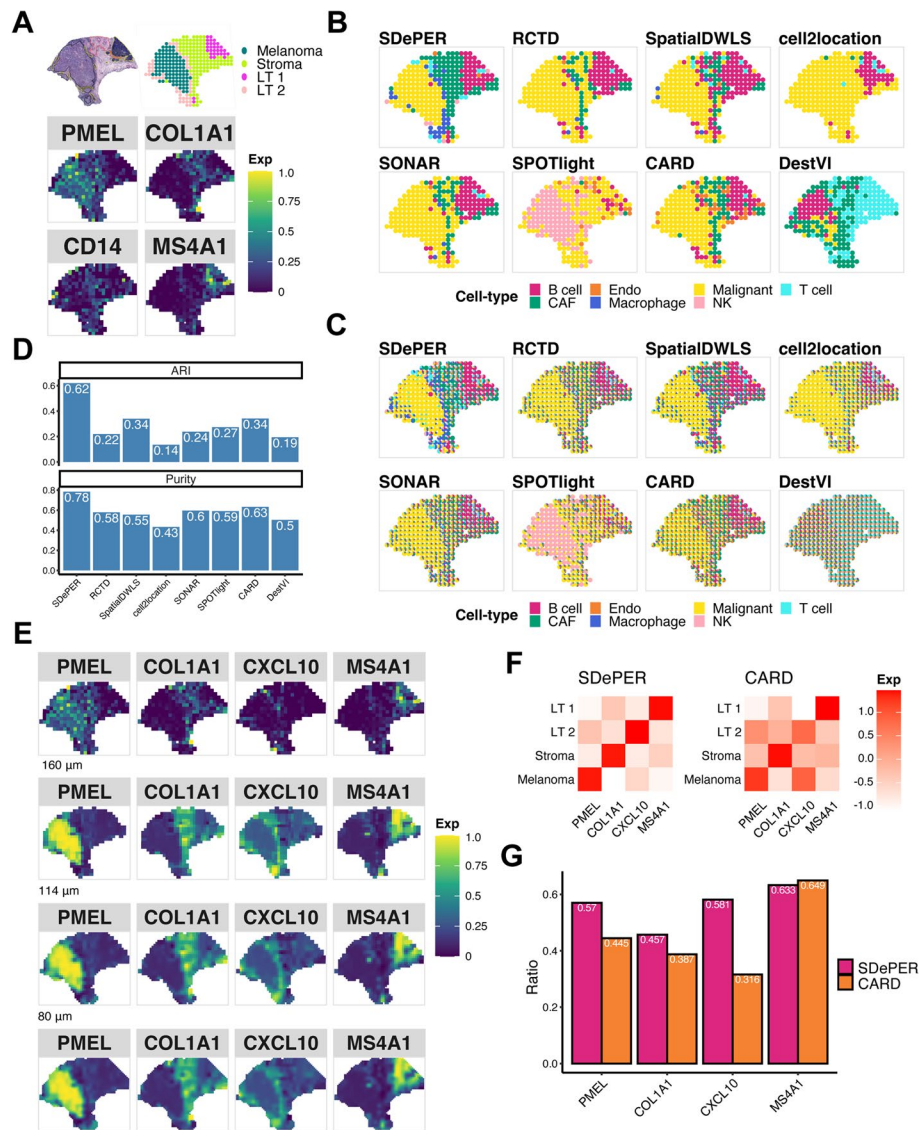


Fig. 4 Performance evaluation and comparison using melanoma dataset. **A** H&E staining of melanoma (top left), melanoma (black), stroma (red), lymphoid tissue (yellow), annotated regions (top right, LT lymphoid tissue) based on BayesSpace and expression pattern of cell-type-specific marker genes for dominant cell types (bottom, *PMEL* for malignant melanoma regions, *COL1A1* for fibroblast in stroma regions, *CD14* for macrophage, and *MS4A1* for B cells). **B** Visualization of inferred dominant cell type in each spot (CAF cancer-associated fibroblasts, Endo endothelial, NK natural killer). **C** Spatial scatter pie chart of estimated cell-type composition within each spot. **D** Comparing deconvolution methods using ARI and purity. **E** Expression patterns of the corresponding region-specific marker genes and its imputed expression at three different resolution levels: 160 μm (about 64% of original size), 114 μm (about 32% of original size), 80 μm (about 16% of original size). **F** Heatmap showing average imputed expression of region-specific marker genes at 80 μm level within each annotated region for SDePER and CARD. **G** Bar plot showing the ratio of average layer-specific marker gene expression in the corresponding layer among all layers

Like the results of MOB data, the dominant cell type predicted by SDePER highly matched the H&E staining image-based annotation (Fig. 4B). In contrast, other methods failed to identify a clear boundary between regions (Fig. 4B). SDePER also achieved the highest ARI and purity that are 1.82 and 1.24 times, respectively, as high

as the second-best method (Fig. 4D). The least non-specific cell-type detection by SDePER was again observed (Fig. 4B, C).

Four region-specific marker genes (*PMEL* for melanoma region, *COL1A1* for stroma region, *CXCL10* for lymphoid tissue 2, and *MS4A1* for lymphoid tissue 1) were identified from the ST data (Fig. 4E) to demonstrate the accuracy of imputed expression. As expected, the imputed expression map of each region marker gene by SDePER showed increased expression in the correct region (Fig. 4E). Compared to CARD, the SDePER recovered the cell-type proportion at a higher resolution (Additional File 1: Fig. S16) and imputed *CXCL10* expression remarkably better resembled the lymphoid tissue 2 on the periphery of the tumor (Additional File 1: Fig. S17). The average imputed expression of each region marker gene by SDePER had a better enrichment for the correct region (Fig. 4F), confirmed by the higher expression ratio of SDePER for each region marker gene (Fig. 4G). All these results suggested that the SDePER-imputed gene expression was more accurate.

HER2-positive breast tumor data

Next, we analyzed the ST data from patients with HER2-positive breast tumor, which consists of various cell types arranged in spatial domains annotated by pathologists [8]. The annotated regions included two cancer regions (cancer in situ and invasive cancer), four named regions (adipose tissue, breast glands, connective tissue, and immune infiltrate), and undetermined regions (Fig. 5A). Although no expected cell type was provided in each region, we expect the cancer regions to enrich for cancer epithelial cells. The expression map of identified cell markers for each cell type confirmed our hypothesis (Additional File 1: Fig. S18). An external scRNA-seq dataset from 5 HER2-positive patients was used as the reference data [39] (Additional File 1: Fig. S19).

The SDePER predicted dominant cell type had the best resemblance to the boundaries between tumor and normal regions in the H&E staining image (Fig. 5B), while other methods failed to detect regions annotated in the staining image. SPOTlight failed to detect any cancer epithelial in the cancer regions, whereas SONAR, DestVI, and cell2location predicted almost all spots to be mainly cancer epithelial cells, which was inconsistent with the H&E staining image. RCTD, SpatialDWLS, and CARD had vague boundaries between tumor and normal regions. SDePER also achieved the highest ARI and purity, which were 2.08 and 1.16 times as high as the second-best method, respectively, confirming the best performance of SDePER (Fig. 5D). The least non-specific cell types were detected by SDePER (Fig. 5B, C).

For the imputation results, we identified four region-specific markers from the ST data for the four annotated regions: cancer region, breast glands, immune infiltrate, and adipose tissue. The imputed expression map of each gene by SDePER showed a more refined and accurate boundary for its corresponding region (Fig. 5E). Compared to CARD, the SDePER imputed cell-type proportion and expression of each region marker gene separated its corresponding region from the other regions better demonstrated by visualization (Additional File 1: Fig. S20-21). Quantitative measure also confirmed that SDePER had a higher enrichment with a larger expression ratio (Fig. 5F-G). Furthermore, imputed expression of *ERBB2* showed an increase in the cancer region matching previous literature [40]. The expression map of known plasma cell marker gene

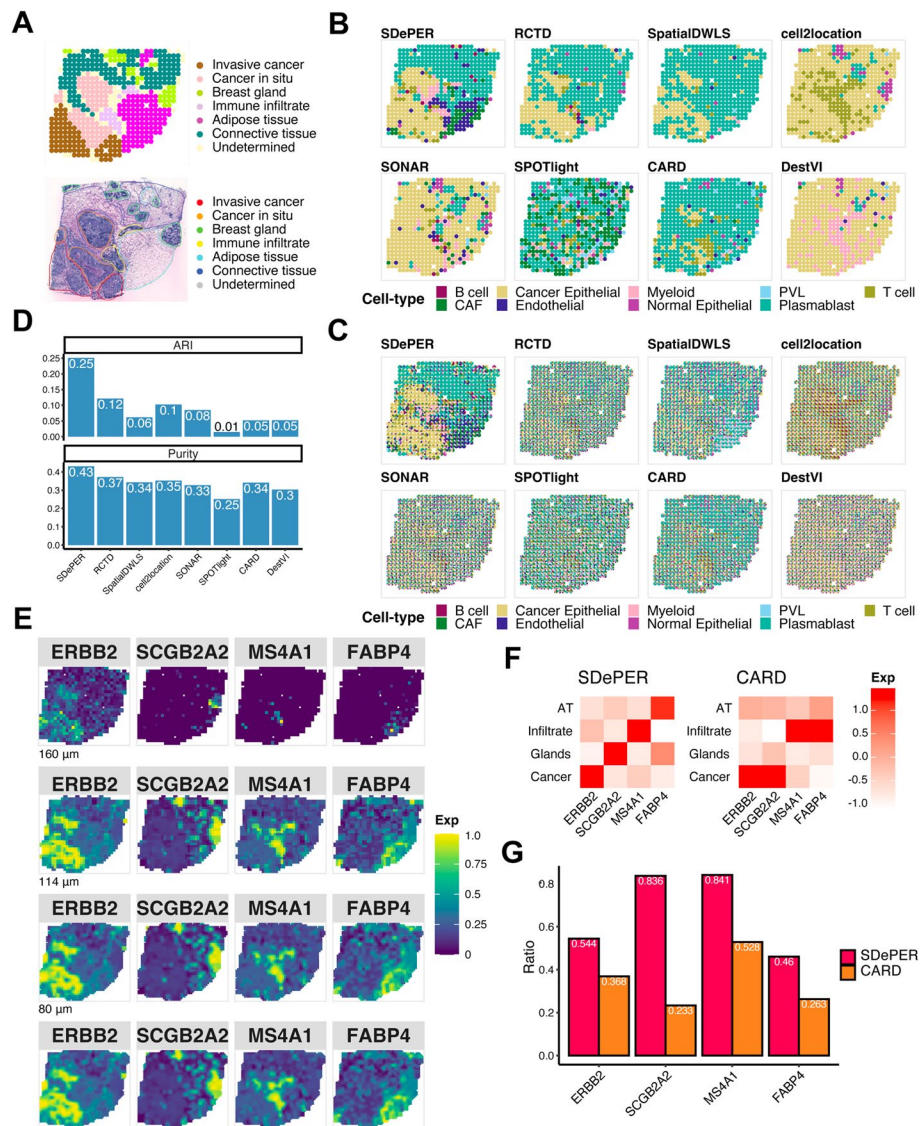


Fig. 5 Performance evaluation and comparison using breast cancer dataset. **A** H&E staining of breast cancer and annotated regions. **B** Visualization of inferred dominant cell type in each spot (CAF cancer-associated fibroblasts, PVL perivascular-like). **C** Spatial scatter pie chart of estimated cell-type composition within each spot. **D** Comparing deconvolution methods using ARI (left) and purity (right). **E** Expression patterns of the corresponding region-specific marker genes and its imputed expression at three different resolution levels: 160 μ m (about 64% of original size), 114 μ m (about 32% of original size), 80 μ m (about 16% of original size). **F** Heatmap showing average imputed expression of region-specific marker genes at locations in each annotated region for SDePER and CARD (AT, adipose tissue; Infiltrate, immune infiltrate; Glands, breast glands; Cancer, invasive cancer and cancer in situ). Imputation at 80 μ m level was used. A red diagonal indicates that each region-specific marker gene was imputed to have high expression in the region that it is the marker for and low expression in the other regions for which it is not a marker for. **G** Bar plot showing the ratio of the average imputed expression levels of the region-specific marker gene in the region that it is a maker for to the other regions. Higher ratio corresponds to more different imputed expression levels of the marker genes between its represented region and other regions

(Additional File 1: Fig. S18), *IGKC*, also ascertained the prevalence of plasma cells in breast glands and connective tissue as predicted by SDePER, rather than perivascular-like (PVL) cells predicted by RCTD. This was also confirmed by the original paper [8]. These further confirmed the higher accuracy of imputation by SDePER.

In the original publication for the breast cancer ST data [8], the co-localization of B cells and T cells was shown to be predictive of the tertiary lymphoid-like structure (TLS) presence in the tissue slice. Visualization of the cell-type proportion estimated by SDePER (Additional File 1: Fig. S22) also demonstrated the co-localization in the TLS regions. In addition, SDePER results also showed enrichment of myeloid cells in the TLS regions which is supported by previous literature [41–43].

Idiopathic pulmonary fibrosis lung data

Lastly, we generated the ST data of a frozen human explant lung sample with idiopathic pulmonary fibrosis (IPF), using the $10 \times$ Genomics Visium platform. IPF is a progressive and irreversible, scarring, and fibrotic lung disease that leads to a complete remodeling of the lung architecture. Due to the complexity and distortion of lung architecture in fibrotic frozen tissues, only the respiratory airway and blood vessels were confidently annotated by a lung fibrosis expert pathologist (Fig. 6A). For deconvolution, we utilized the scRNA-seq dataset of IPF distal lung parenchyma sample as the reference data [44] (Additional File 1: Fig. S23).

We demonstrated the results using four cell types: ciliated cells from the airway, smooth muscle cells (SMC) from the vascular and alveolar type 1 (AT1) and type 2 (AT2) cells from the alveoli. Expression maps of marker genes for ciliated cells and SMC match the annotations of airway and vascular, respectively (Fig. 6B). Marker gene expression maps of AT1 and AT2 cells also suggested their prevalence over the distal side of the lung where alveoli are located.

Visualization of the predicted cell-type proportions (Fig. 6C, Additional File 1: S24) showed that SDePER captured the location of four cell types accurately and precisely, which well-matched both the expression map of marker genes (Fig. 6B) and the pathological annotation (Fig. 6A). But other methods either lacked the specificity in the estimation or failed to identify cell types. RCTD, SpatialDWLS, and DestVI had excessive non-zero estimations in spots lacking the corresponding marker gene expression, especially for SMC in the vascular region and ciliated cells in the airway. For each method, the average expression of each marker gene across all spots weighted by the predicted proportion of its corresponding cell type was calculated to quantitatively measure the consistency between the estimated cell-type compositions and marker gene expression maps. SDePER achieved the highest weighted mean for SMC, ciliated cells, and AT2 cells (Fig. 6D). It had a comparable performance in AT1 cells. These quantitatively confirmed the highest estimation accuracy of SDePER. Moreover, SDePER results demonstrated co-localization of AT1 and AT2 cells on the margin of tissue slide with the highest pairwise correlation of estimated cell-type proportions, which is consistent with the cell-type marker gene expression map and anatomy of human lungs (Fig. 6E).

Furthermore, we examined the results of other important cell types (Additional File 1: Fig. S25), including aberrant basaloid cells, adventitial fibroblast, and airway

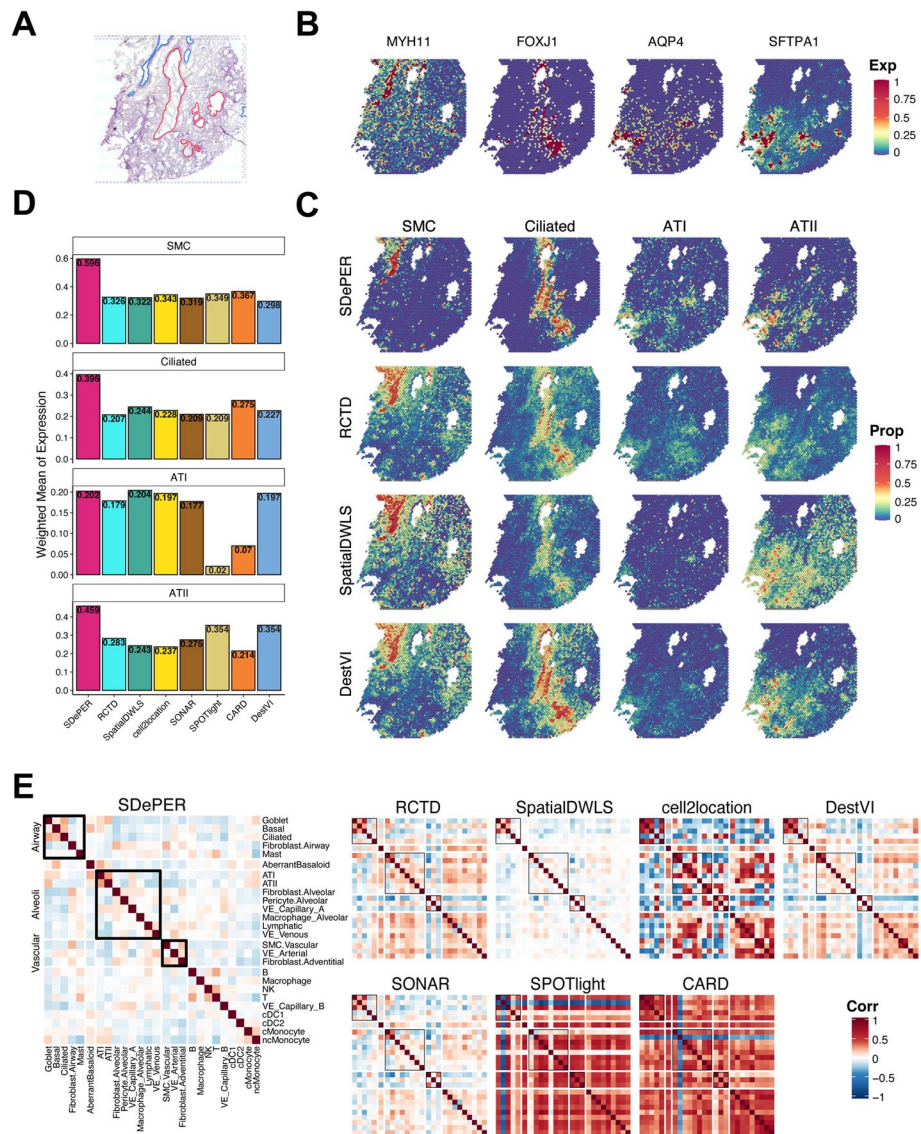


Fig. 6 Performance evaluation and comparison using idiopathic pulmonary fibrosis lung dataset. **A** H&E staining of breast cancer with annotated regions: respiratory airway (red) and blood vessels (blue). **B** Heatmaps of selected cell-type marker genes expression patterns for SMC (MYH11), ciliated cells (FOXJ1), AT1 (AQP4), and AT2 (SFTPA1) cells. **C** The estimated cell-type proportions on each location for SMC, ciliated cells, AT1, and AT2 cells inferred by SDePER, RCTD, SpatialDWLS, and DestVI. **D** Barplot of the average expression of marker genes among all spots weighted by estimated proportions of the corresponding cell type for each method. **E** Pairwise correlation of estimated cell-type proportions for each method

fibroblast. Aberrant basaloid cells seem to co-localize well with basal cells. They are also present in the alveoli together with AT I and AT II cells. Adventitial fibroblast was found to co-localize with the vascular smooth muscle cells, suggesting its presence in the vascular compartment. A recent spatial transcriptomic study [45] of IPF lung using the 10 × Genomics Xenium platform validated this finding. The airway fibroblasts were found to be present in the vascular compartment instead of in the airway, indicating that a further investigation of the location of these cells in human lungs is needed.

Overall, this is the first time that ST data from the human lung sample is used for the demonstration of cell-type deconvolution. The results showed that SDePER is a reliable method for complex tissue samples with vague structures and rare cell types.

Discussion

There are several directions of extensions of SDePER. First, SDePER removes the systematic differences between ST and scRNA-seq data in an unsupervised manner, which can be improved by utilizing the known cell-type compositions of reference scRNA-seq data and pseudo-spot data as supervision in the training of CVAE. Multi-task learning strategy can be used to integrate the unsupervised and supervised learning and leverage the information of cell-type compositions in the pseudo-spot data to guide the CVAE training. Second, we assumed that the distribution of embeddings in the CVAE latent space follows a standard normal distribution. This assumption can be relieved by introducing importance sampling [46, 47]. Third, the encoder and decoder in CVAE are multi-layer neural networks, which are generic to approximate any functions [48, 49], leading to its relatively high variance and sensitivity to the variation of model structure and initialization. This can be improved by using a negative binomial distribution [50] instead in the decoder. Fourth, the computational speed of SDePER may be a concern for larger-scale ST data with tens of thousands of spots. Based on the SDePER model and algorithm, the computational time should linearly increase with the total number of spots. We further evaluated how the number of genes used in the CVAE training and GLRM model fitting affected the computational speed (Additional File 1: Fig. S26), which demonstrated a linearly increasing computational time when the number of genes increased. For large-scale spatial transcriptomics data, by selecting limited but representative genes (~500 genes), we could finish the analysis of one $10 \times$ Visium tissue slide with about around 3500 spots in ~2.5 h. For data with even larger scale, it is possible to disable the graph Laplacian penalty in SDePER so that it can be run parallelly across different spots. In addition, the computational efficiency of SDePER can be further improved by caching the calculated log-likelihoods in GLRM fitting to avoid repetitive. Finally, results for the internal reference showed that CVAE may introduce noise when there were no platform effects. One potential way to assess the severity of platform effects is to examine the overlap between reference and real ST data in the UMAP before and after the CVAE process. This is similar to the integration analysis which improves overlap between cells of the same type across batches. Larger improvement in the overlap between the two platforms may suggest more severe platform effects.

Conclusions

We have developed a novel deconvolution method, SDePER, to deconvolute spatial barcoding-based transcriptomic data using reference scRNA-seq data, with considerations of platform effects, sparsity, and spatial correlation. Through simulations, we demonstrated the superior performance and robustness of SDePER to platform effects and mismatching cell types between ST and reference data. Applications to datasets from various tissue types, species, and platforms also showed a superior accuracy in the estimated cell-type compositions and imputed gene expression of SDePER.

Methods

SDePER method overview

SDePER is built upon the combination of a conditional variational autoencoder (CVAE) [34] and a graph Laplacian regularized regression model (GLRM). The CVAE component aims to remove platform effects and the GLRM component aims to estimate cell-type compositions at each spot based on cell-type-specific signatures from reference scRNA-seq data with considerations of sparsity and spatial correlation of cell-type compositions between neighboring spots in the tissue. Based on the estimated cell-type proportions, the imputation of cell-type compositions and gene expression at unmeasured locations in refined spatial maps with higher resolution is performed using a nearest neighbor random walk.

Conditional variational autoencoder for platform effect adjustment

The CVAE model [34] considers the two technology platforms, i.e., reference scRNA-seq and ST, as two conditions. The loss function of CVAE is defined as

$$\text{loss} = -KL(q_\phi(\mathbf{z}|\mathbf{x}, c)||p_\omega(\mathbf{z})) + \mathbb{E}_z(\log(p_\omega(\mathbf{x}|\mathbf{z}, c))),$$

where \mathbf{x} represents the gene expression profile, c is the conditional variable, \mathbf{z} is the latent embedding in the latent space, q_ϕ is the encoder parameterized by ϕ to embed samples into the latent space, $p_\omega(\mathbf{z})$ is the prior distribution of latent embedding \mathbf{z} defined as the standard Gaussian distribution $\mathcal{N}(0, \mathbf{I})$, $p_\omega(\mathbf{x}|\mathbf{z}, c)$ is the decoder parameterized by ω to generate gene expression data given the latent embedding \mathbf{z} and conditional variable c , and KL is the Kullback-Leibler divergence function. The CVAE loss function is optimized using Adam [51], which will learn $q_\phi(\mathbf{z}|\mathbf{x}, c)$, $p_\omega(\mathbf{x}|\mathbf{z}, c)$ and \mathbf{z} from the data.

Since CVAE assumes a Gaussian distribution for data under both conditions, it is critical for the data under the scRNA-seq condition to cover a similar spectrum of cell-type proportions as the ST data. However, the reference scRNA-seq data only has data with one single-cell type, whereas each spot in the ST data can have cells from multiple cell types. To make the training data under the two conditions have a similar spectrum of cell-type compositions, pseudo-spot data are generated from the reference scRNA-seq data to provide a wide spectrum of cell-type compositions for the input under the scRNA-seq condition. For each pseudo-spot, we randomly select a set of cells from reference scRNA-seq data and calculate the average normalized gene expression across cells as the expression profile for the pseudo-spot. The range of the number of selected cells per pseudo-spot is specified based on the cell density in real ST data. In total, the number of pseudo-spots generated is $\min(100 \times N \times K, 500,000)$, where N is the number of spots in the real ST data and K is the number of cell types in the reference scRNA-seq data. We train the CVAE model using 80% of the pseudo spots, the reference scRNA-seq data, and real ST data. The rest 20% pseudo-spots are used as validation data for learning rate decay and early stopping. Genes used in CVAE are the union of top highly variable genes and cell-type marker genes identified from the reference scRNA-seq data identified using Scanpy 1.9.1 [52]. The sizes of both gene lists can be tuned by users based on the properties of reference scRNA-seq data. Because the training of CVAE is sensitive to differences in data range across different genes, the normalized expression of each gene is further rescaled separately for the scRNA-seq and ST condition to be from 0 to

10 using min–max scaling. The conditional variable in CVAE represents which platform (scRNA-seq or ST) generated the data and is set to 0 or 10 by default. In the training data, the conditional variable for real ST data was set to 10 for the ST condition and the conditional variable for the pseudo-spot data and reference scRNA-seq data was set to 0 for the scRNA-seq condition.

In the CVAE training, we set the number of neurons in latent space as three times the number of cell types in the reference scRNA-seq data and use one hidden layer for both encoder and decoder under each condition, in which the number of neurons is the largest integer no more than the geometric mean of the number of neurons in the input layer and latent space. We use Adam [51] for optimization and the initial learning rate is set to 0.003 with decay specified based on the value of loss function of the validation dataset. The number of epochs is set to 1000 and the early training stopping criteria is when the value of loss function of the validation dataset increases.

After the CVAE training, the real ST data is embedded into the latent space for the ST condition (conditional variable=10) and then is decoded using the decoder for the scRNA-seq condition (conditional variable=0). The reference scRNA-seq data is encoded and decoded for denoising using the encoder and decoder for the scRNA-seq condition (conditional variable=0). The decoded gene expression levels are min–max scaled back using the scRNA-seq rescaling factors. For the real ST data, the rescaled values are further multiplied by 10,000 and rounded to the nearest integers. By using the same decoder, ST and scRNA-seq data are transformed into the same space to remove platform effects. Visualization of the transformed data showed that enough biological signals were retained to separate different cell types (Additional File 1: Fig. S27). The transformed real ST data and reference scRNA-seq data serve as input to the GLRM component. In addition, the mean–variance relationship was well preserved after the CVAE adjustment in all real datasets (Additional File 1: Fig. S28).

Multiple batch effect removal methods have been developed for scRNA-seq data including MNN [53], Harmony [54], Seurat Integration [55], and so on, which can be used to correct for platform effects as well. However, these methods strongly rely on the assumption that there are common cell types or shared biological cell states between batches. Specifically, the MNN-based approaches, such as the Seurat Data Integration method, identify pairs of cells from different batches and the difference between cells in each pair is utilized to estimate batch effects. When scRNA-seq and ST data are considered as two batches by these approaches, each spot from the ST data is a mixture of multiple cells with potentially multiple cell types while each cell from the scRNA-seq data has only one cell type. Unless there are many spots in the ST data that have only one cell type, the identified “cell pairs” do not have the same biological state and the difference between them include both platform effects and cell-type composition difference. This causes the estimated batch effects to be larger than platform effects, so these approaches cannot provide adequate adjustment for platform effects. CVAE is a deep generative model which learns the data distribution in a latent space and a generative process to generate new data points from the learned distribution. It assumes that scRNA-seq data and ST data share the same type of distribution in the embedding space, which is a weaker assumption than the batch correction methods. We accommodate for this assumption by adding pseudo-spots in the CVAE training data. The generated new

data will retain the original data distribution as well as the original biological meaning so the CVAE adjusted data is still gene expression data, which is critically important for the GLRM component because it assumes a linear additive relationship between the ST and reference data. To demonstrate the advantage of CVAE, we replaced the CVAE component in SDePER with Seurat Integration method and ran it on the STARmap-based simulated data with external reference. We compared the results to those of SDePER and GLRM (Additional File 1: Fig. S29). The comparison showed that Seurat Integration did correct for a certain part of the platform effects so it had a certain improvement over GLRM. But SDePER was able to achieve further and larger improvement over Seurat + GLRM, suggesting higher efficiency of CVAE in correcting for platform effects than Seurat.

Graph Laplacian regularized model for cell-type deconvolution

We fit a graph Laplacian regularized model to estimate cell-type compositions in each spot using the transformed ST and reference scRNA-seq data. Since biological signals can be lost by the CVAE adjustment, we identified cell-type marker genes from the transformed reference scRNA-seq data by comparing each cell type to every other cell type using the Wilcoxon Rank Sum test implemented in the FindMarker function in Seurat. Genes with false discovery rate less than 0.05, fold change ≥ 1.2 , pct.1 ≥ 0.3 and pct.2 ≤ 0.1 were kept and sorted based on the fold change. By default, the top 20 genes across all comparisons were merged and used to fit the GLRM model. The transformed ST data of each gene is assumed to follow a Poisson distribution with the log-transformed mean being a linear combination of its transformed across all cell types, which forms the base model. The transformed expression profile of each cell type is calculated as the average expression profiles across cells of the given cell type from the transformed reference scRNA-seq data. On top of the base model, we incorporate the spot location information using graph Laplacian regularization that encourages cell-type compositions of neighboring spots to be similar. We also enforce cell-type sparsity within each spot using adaptive LASSO regularization. Specifically, the GLRM component consists of the following three major components.

Base model

The transformed ST data is modeled using a Poisson-loglinear model which considers dispersion in the data (Additional File 1: Fig. S28). For each spot i and gene j , the transformed ST count Y_{ij} is assumed to follow a Poisson distribution:

$$Y_{ij} | \lambda_{ij} \sim \text{Poisson}(N_i \lambda_{ij}),$$

where N_i is the observed total UMI count of spot i and λ_{ij} represents the true underlying relative expression level of gene j in spot i . The rate parameter λ_{ij} is further modeled as a combination of expression profiles of all K cell types weighted by the cell-type proportions,

$$\log(\lambda_{ij}) = \alpha_i + \log\left(\sum_{k=1}^K \theta_{ik} \mu_{kj}\right) + \epsilon_{ij},$$

where α_i is a parameter representing spot-specific fixed effect, θ_{ik} is the proportion of cells from cell-type k in spot i , μ_{kj} is the mean expression level of gene j in cell-type k calculated from the transformed reference scRNA-seq data, and ϵ_{ij} is a random error that follows a normal distribution with mean 0 and variance σ^2 as defined in RCTD [12]. The distribution of ϵ_{ij} is relaxed to include a heavy tail using an approximation to a Cauchy–Gaussian mixture distribution, which is robust to outliers [56],

$$p(\epsilon) = \begin{cases} \frac{C}{\sqrt{2\pi}\sigma} e^{-\frac{\epsilon^2}{2\sigma^2}}, & |\epsilon| \leq 3\sigma \\ \frac{2\sqrt{2}C}{9(\epsilon\sigma - \frac{7}{3}\sigma^2)\sqrt{\pi}} e^{-\frac{9}{2}}, & |\epsilon| > 3\sigma \end{cases},$$

Where C is a normalizing constant which is chosen to make $p(\epsilon)$ integrate to 1. The model parameters θ_{ik} are subject to the constraint that $\sum_{k=1}^K \theta_{ik} = 1$ and $\theta_{ik} \geq 0$ for all i and k . We considered α_i because previous studies [37, 57–60] demonstrated that the gene expression profile of cells of the same type could vary depending on where they are located in the tissue, potentially due to the influences from neighboring cells or tissue microenvironment.

Adaptive LASSO regularization

To enforce the local sparsity of cell types in each spot, we penalize the likelihood function of the base model using the adaptive Lasso penalty [61]. For each spot i , we define the adaptive Lasso penalty as

$$r(\theta_i) = \sum_{k=1}^K q_{ik} |\theta_{ik}|,$$

where q_{ik} is the reciprocal of maximum likelihood estimation (MLE) of θ_{ik} from SDePER without the adaptive LASSO and graph Laplacian penalties (base model), serving as the weight of θ_{ik} in the adaptive Lasso penalty.

Laplacian regularization

To incorporate spatial information, we represent the physical proximity between spots using an adjacency matrix $A = [A_{ij}]_{I \times I}$. Although A can be calculated using the Euclidean distance between spots, for simplicity, we use unweighted adjacency matrix throughout this article, where A_{ij} is an indicator of whether spot i and j are neighbors on the tissue slide, and I is the total number of spots. The graph Laplacian is defined as $L = D - A$, where D is a diagonal matrix with $D_{ii} = \sum_j A_{ij}$, the degree of spot i . The graph Laplacian penalty is defined as

$$\mathcal{L}(\theta) = \frac{1}{2} \sum_{s,t=1}^K A_{st} \|\theta_s - \theta_t\|_2^2 = \text{tr}(\theta^T L \theta),$$

where $\text{tr}(\cdot)$ is the trace of a matrix. This penalty measures the aggregate deviation of θ between neighboring spots and therefore encourages θ to be similar across neighboring spots.

Taken together, we fit the GLRM model by minimizing the following objective function:

$$F(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\mu}, \sigma^2) = \sum_{i=1}^I \left[l_i(\boldsymbol{\theta}_i, \alpha_i, \boldsymbol{\mu}, \sigma^2) + \lambda_r r(\boldsymbol{\theta}_i) \right] + \lambda_l \mathcal{L}(\boldsymbol{\theta}).$$

The first term of the objective function is the negative log-likelihood of the base model. The second term is the local adaptive Lasso penalty that enforces cell types to be sparsely present within each spot. The third term is the graph Laplacian penalty to encourage smoothness of cell-type compositions across neighboring spots. λ_r and λ_l are two positive hyperparameters.

The objective function is minimized using a two-stage strategy. To provide initial values of all parameters ($\boldsymbol{\theta}$, $\boldsymbol{\alpha}$ and σ^2) for optimization, we calculate the gene expression profile of cell-type k using the average library size normalized expression levels of all identified cell-type marker genes across all cells of type k , denoted as $\hat{\mu}_k$. The maximum likelihood estimations ($\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\alpha}}$, and $\hat{\sigma}^2$) of the base model are obtained using the L-BFGS algorithm [62] in SciPy 1.8.1 [63] and serve as the initial values for optimization. In the first stage of optimization, we perform cell-type selection in each spot by minimizing the negative log-likelihood function of the base model with the adaptive LASSO penalty: $\sum_{i=1}^I [l_i(\boldsymbol{\theta}_i, \alpha_i | \hat{\boldsymbol{\mu}}, \hat{\sigma}^2) + \lambda_r r(\boldsymbol{\theta}_i)]$ using alternating direction method of multipliers (ADMM) [64]. A cutoff value of 0.001 is applied to the estimate $\hat{\boldsymbol{\theta}}$ to determine which cell types are present in each spot. In the second stage of optimization, we only include cell types selected from the first stage for each spot in the base model and minimize the negative log-likelihood function of the base model with graph Laplacian regularization using ADMM: $\sum_{i=1}^I l_i(\boldsymbol{\theta}_i, \alpha_i | \hat{\boldsymbol{\mu}}, \hat{\sigma}^2) + \lambda_l \mathcal{L}(\boldsymbol{\theta})$. The values of the two hyperparameters, λ_r and λ_l , are chosen using fivefold cross-validation. The cell-type marker genes identified from the transformed reference scRNA-seq data were randomly divided into five groups with equal size. Each group was considered as validation data and the rest groups were used as training datasets. For given values of λ_r and λ_l , the training data was used to fit the GLRM model, which was used to calculate the log-likelihood of the validation data using the base model. The log-likelihood of the five validation datasets was averaged and compared across different settings of λ_r and λ_l . The setting that achieved the largest average log likelihood was chosen.

Imputation on cell-type composition and gene expression

SDePER borrows information from neighboring spots to perform imputation of cell-type compositions at unmeasured locations in refined tissue map with arbitrary resolution by taking the nearest neighbor random walk on the spatial graph. We first use the “finding contour” function in opencv [65] to determine the contours of the tissue shape and distinguish outlines of the tissue and holes inside the tissue. The spatial spots closest to the outline and border of holes are set to be edge spots. We also develop a custom algorithm to calculate the missing spots in the holes. Suppose the spatial coordinates of the center for spot i is $c_i = (x_i, y_i)$ and the distance between centers of two neighboring spots is D . To construct a new spatial map at enhanced resolution, we first create the smallest rectangular region that covers all centers of original spots using $[\min_i(x_i), \max_i(x_i)] \times [\min_i(y_i), \max_i(y_i)]$. This rectangular is then gridded into

squares with side length d ($d < D$). Among all the squares, those with center located at $c_{i^*} = (x_{i^*}, y_{i^*})$ satisfying one of the following criteria are considered as spots in the new map and imputed:

$$\min_{\{i=1,\dots,N; i \text{ is an inner spot}\}} \|c_{i^*} - c_i\| \leq D \text{ or } \min_{\{i=1,\dots,N; i \text{ is an edge spot}\}} \|c_{i^*} - c_i\| \leq \frac{D-d}{2}.$$

These criteria filter out center locations that are far away from the edge and inner spot, outside the tissue slice or in biological holes. The new spatial map at enhanced resolution (square side length = d) consists of the spots with centers $\mathbf{C}^* = \{c_{i^*}, i^* = 1, \dots, N^*\}$.

Let θ_{i^*} denote the cell-type proportions in spot i^* in the new spatial map. To perform imputation, we first assign an initial value for θ_{i^*} by finding the nearest original spot(s) of spot i^* and set $\theta_{i^*}^0$ as the average cell-type proportion among its neighbors. We construct a Gaussian kernel \mathbf{W} as follows

$$W_{i^*j^*} = \begin{cases} 0, & r_{i^*j^*} > \phi \\ e^{-\frac{r_{i^*j^*}^2}{2\tau^2}}, & r_{i^*j^*} \leq \phi \end{cases} \quad i^*, j^* = 1, 2, \dots, N^*,$$

where $r_{i^*j^*} = \|C_{i^*} - C_{j^*}\|$ is the distance between two spots i^* and j^* in the new map, ϕ is a predefined neighborhood size within which spots contribute to the imputation of each other, and τ^2 is the variance. A nearest neighbor random walk matrix is constructed as $\mathbf{M} = \mathbf{D}^{-1}\mathbf{W}$, where \mathbf{D} is a diagonal matrix with $D_{i^*i^*} = \sum_{j^*} W_{i^*j^*}$. The imputed cell-type compositions are obtained by taking a one-step nearest neighbor random walk with the graph which can be written as

$$\theta_{\text{imputed}} = \mathbf{M}\theta^0$$

We further impute gene expression at enhanced resolution as $\mathbf{X}_{\text{imputed}} = \theta_{\text{imputed}}\theta^+\mathbf{X}$, where \mathbf{X} is the observed UMI counts from ST data, normalized by spot's sequence depth, and θ^+ is the Moore–Penrose inverse of θ with $\theta^+ = (\theta^T\theta)^{-1}\theta^T$.

The hyperparameters in the Gaussian kernel \mathbf{W} include ϕ and τ^2 . For a given real ST data, the hyperparameter tuning was conducted using the STARmap-based simulated data. We conducted coarse-graining procedures on the STARmap data to generate simulated ST dataset with different spot sizes varying from 100×100 to 1000×1000 with an interval of 100, which correspond to high to low resolution. In each simulated ST dataset, we know the true cell-type proportions in each spot. For a given setting of ϕ and τ^2 , we impute the cell-type proportions using the simulated dataset with spot size 1000×1000 to reconstruct spatial maps with different resolutions higher than 1000×1000 (smaller spot size). Then we compare the imputed cell-type proportions to the ground truth and calculated the average RMSE across the different higher resolution levels. The hyperparameter setting that achieved the smallest average RMSE was chosen. The search ranges for τ and ϕ are both 1–200 μm .

Other deconvolution methods for comparison

Seven state-to-art spatial deconvolution methods were chosen to compare with SDePER, including RCTD (version 2.0.1) [12], SpatialDWLS (implemented in the R package

Giotto, version 1.1.2) [26], SONAR (version 1.0.0) [28], SPOTlight (version 1.8.0) [25], cell2location (version 0.1.3) [15], DestVI (implemented in the python package scVI, version 1.1.3) [14], and CARD (version 1.0) [24]. We followed the tutorial on the GitHub repository of each method and used the recommended default parameter settings for the deconvolution analyses conducted in this article. When parameters are required to be set manually, we used the values suggested in the vignettes.

Simulation studies

To evaluate the method performance and demonstrate the impact of platform effects on all deconvolution methods, we simulate spot-level ST data in multiple different ways.

STARmap-based simulation

We first simulated ST data based on the adult mouse primary visual cortex STARmap data that has single-cell resolution [35]. We extract experiments “20,180,410-BY3_1kgenes” and “20180505_BY3_1kgenes” and manually put them in the same spatial map with enough space in between so that cells or simulated spots from different experiments are not considered as spatial neighbors. To simulate the ST data, we gridded the tissue slide into squares with a side length of $\sim 51.5 \mu\text{m}$ as capture spots, which generated 581 spots with 1 to 12 cells and an average of 3.6 cells present per spot. We only kept cells from cell types present in both STARmap data and external reference scRNA-seq data. In each spot, the proportion of cells from each cell type is calculated and serves as ground truth for performance evaluation. The simulated gene expression level of gene j for a given spot that contains cells $i = 1, \dots, n$ is calculated as $nUMI_j = \lceil \frac{\sum_{i=1}^n U_{ij}}{\sum_{i=1}^n U_{ij}} \times N \rceil$, where U_{ij} is the number of UMIs of gene j in cell i from the STARmap data and N is a fixed scaling factor set to be 1000.

Sequencing-based simulation

The STARmap technology is a hybrid technology of in situ hybridization and sequencing. The protocol enriches for transcripts using hybridization techniques and the final nUMI is generated based on sequencing. To simulate ST data that are purely sequencing-based, we utilized a scRNA-seq dataset from the mouse visual cortex [66] measured using the inDrops technique (GEO accession number: GSE102827). We modified the STARmap data by retaining the spatial location of each cell but replacing its expression profile with that of a randomly chosen cell of the same type from the inDrops data. Then the same coarse-graining procedure was applied to this modified STARmap data to simulate purely sequencing-based ST data.

External and internal reference

To demonstrate the impact of platform effects on the method performance, each method was applied to the simulated data using two different reference scRNA-seq datasets: internal reference and external reference. The internal reference data is the original ST data with single-cell resolution so there were no platform effects. For STARmap-based simulation, the internal reference data was the STARmap data. For sequencing-based simulation, the internal reference data is the inDrops data (GEO: GSE102827). The external reference data is an independent publicly available scRNA-seq dataset. For

both STARmap-based and sequencing-based simulations, the adult mouse visual cortex scRNA-seq dataset (GEO accession number: GSE115746) [67] was used as reference data. Under this case, significant platform effects were expected to exist because the simulated ST data and reference data were generated using the in situ sequencing and SMART-seq technologies, respectively. We selected 12 overlapping cell types between the external reference data and the STARmap data for deconvolution, which include astrocytes, excitatory neurons layer 2/3, excitatory neurons layer 4, excitatory neurons layer 5, excitatory neurons layer 6, endothelial, microglia, oligodendrocyte, *Pvalb*-positive cells, *Vip* inhibitory neurons, *Sst* neurons, and smooth muscle cells. In total, 2002 cells and 1020 genes were included in the STARmap data while 11,835 cells and 45,768 genes were in the external reference data.

Simulation for mismatching cell types

Deconvolutions using the overlapping 12 cell types between the STARmap data and external reference scRNA-seq data represent analysis scenario 1, under which cell types in the reference data match perfectly with those in the ST data. However, in practice, they can have mismatching cell types, so we modified the external reference data to demonstrate the robustness of all methods to mismatching cell types. In analysis scenario 2, we remove *Vip* inhibitory neurons ($n = 1690$) from the reference data. In analysis scenario 3, we added “high intronic” cells ($n = 182$), which are not present in the STARmap data, to the reference data.

Simulations for rare cell types

To assess the robustness of SDePER to rare cell types, we conducted the following two simulation analyses by choosing oligodendrocytes (“Oligo”) as the “rare cell type” for investigation. First, we down-sampled Oligo cells in the reference scRNA-seq data to a given number (5, 10, 20, and 50) for multiple times and used each down-sampled reference data to deconvolve the STARmap-based simulated ST data. In the second simulation analysis, we examined the performance of SDePER on Oligo cells using groups of spots stratified based on the number of Oligo cells per spot from the STARmap-based simulation. All the simulated spots were divided into groups based on the total number of cells (n) and the number of oligodendrocytes per spot. Within each group of spots with the same total number of cells, both the relative absolute error (RAE) and the false negative rate (FNR) were calculated.

Simulations for high cell density

To simulate ST data with high cell density, we kept the physical spot size the same as in the sequencing-based simulation but increased the total number of cells in each spot by 3 or 6 times, which corresponds to approximately 3 to 36 cells and 6 to 72 cells per spot with an average of 10.8 and 21.6 cells per spot, respectively. In each spot, the cell-type proportions remained the same but for each existing cell type, three or six times more cells were randomly selected from the inDrops without replacement to calculate the simulated spot data.

Simulations for small number of cell types

To evaluate the necessity of adaptive Lasso when the number of cell types is small, we removed all cells that do not belong to five chosen cell types (eL2/3, eL4, eL5, eL6, and Oligo) from the STARmap data. The same simulation procedure was conducted to simulate STARmap-based ST data. When conducting deconvolution on the simulated ST data using external reference, cells that do not belong to the five chosen cell types were also removed from the reference data.

Ablation tests

To understand the contribution of different components in SDePER, we conducted ablation tests by disabling the CVAE for platform effects removal, pseudo-spots inclusion in the CVAE training, adaptive LASSO penalty for sparsity, or graph Laplacian penalty for spatial correlation in SDePER. Three datasets were used including those from the STARmap-based simulation, sequencing-based simulation, and the simulation for high cell density. For each dataset, we consider the performance of SDePER as the baseline. The performance of SDePER with each component disabled was compared to the baseline performance to assess the contribution of the component. For adaptive LASSO penalty, the dataset from simulations for a small number of cell types was also used for the ablation test.

Performance evaluation criteria

To evaluate the method performance, we compare the cell-type compositions estimated by each method, $\hat{\theta}_i$, to the ground truth θ_i for each spot i using the root mean square error (RMSE) that quantifies the overall estimation accuracy, Jensen–Shannon Divergence (JSD) that assesses similarity between the estimated cell-type distribution and ground truth per spot, Pearson’s correlation coefficient that measures the similarity of estimation to ground truth, and false discovery rate (FDR) that measures how many cell types were falsely predicted to be present. Formulas of these criteria are as follows:

$$\text{RMSE}(\hat{\theta}_i) = \sqrt{\frac{1}{K} \sum_{k=1}^K (\hat{\theta}_{ik} - \theta_{ik})^2}.$$

$\text{JSD}(\hat{\theta}_i \parallel \theta_i) = \frac{1}{2} (\text{KL}(\hat{\theta}_i \parallel \frac{\hat{\theta}_i + \theta_i}{2}) + \text{KL}(\theta_i \parallel \frac{\hat{\theta}_i + \theta_i}{2}))$, where $\text{KL}(\cdot \parallel \cdot)$ represents Kullback–Leibler divergence.

$$\text{cor}(\hat{\theta}_i, \theta_i) = \frac{\sum_{k=1}^K (\hat{\theta}_{ik} - \bar{\hat{\theta}}_i)(\theta_{ik} - \bar{\theta}_i)}{\sqrt{\sum_{k=1}^K (\hat{\theta}_{ik} - \bar{\hat{\theta}}_i)^2} \sqrt{\sum_{k=1}^K (\theta_{ik} - \bar{\theta}_i)^2}}.$$

$$\text{FDR}_i = \frac{\sum_{k=1}^K I(\hat{\theta}_{ik} \neq 0, \theta_{ik} = 0)}{\sum_{k=1}^K I(\hat{\theta}_{ik} \neq 0)}.$$

Real dataset analysis

Mouse olfactory bulb dataset

We obtain the mouse olfactory bulb (MOB) ST data from the Spatial Research lab [4]. We focus on the “MOB replicate 12” file which contains 16,034 genes and 282 spots. An

independent scRNA-seq data is also downloaded as the reference scRNA-seq data (GEO accession number: GSE121891) [36], which consists of 18,560 genes and 12,801 cells from 5 cell types: granule cells, olfactory sensory neurons, periglomerular cells, mitral and tufted cells, and external plexiform layer interneurons.

To apply SDePER on the MOB dataset, we select 250 highly variable genes and 244 cell-type marker genes from the reference scRNA-seq data, which form a set of 434 unique genes used in the CVAE component for platform effects removal. We randomly select 10–40 single cells from the scRNA-seq data to generate a pseudo spot to mimic the number of cells per spot suggested in the MOB data; 168 cell-type marker genes are used in GLRM component. The hyperparameter λ_r is chosen to be 1.931 and λ_l to be 5.179 based on cross-validation. The running time of SDePER is 0.56 h in total using a 20-core, 100 GB RAM, Intel Xeon 2.6 GHz CPU machine.

Melanoma dataset

We download the melanoma dataset from the Spatial Research lab [7]. We focus on the second replicate from biopsy 1 because it contains regions annotated as lymphoid tissue and is extensively examined in the original paper. Biopsy 1 contains 16,148 genes and 293 spots. The reference scRNA-seq dataset is downloaded from GEO database (accession number: GSE115978), which contains 23,686 genes and 2495 cells from 8 selected samples. In total, seven cell types are present in the reference data, including malignant cells, T cells, B cells, natural killer (NK) cells, macrophages, cancer-associated fibroblasts (CAFs), and endothelial cells [38].

We choose the top 300 highly variable genes and 280 marker genes, corresponding to 534 unique genes for the CVAE component in SDePER. The number of cells per pseudo-spot is set to be 5–40 cells as provided in the original paper; 145 cell-type marker genes are used in GLRM component. The hyperparameter λ_r is chosen to be 1.931 and λ_l to be 37.276 based on cross-validation. The running time of SDePER is 0.56 h in total using a 32-core, 100 GB RAM, Intel Xeon 2.6 GHz CPU machine.

Breast cancer dataset

We obtain the HER2-positive breast cancer spatial transcriptomics dataset from a previous study [8]. The first section of patient H with 15,029 genes and 613 spots is selected for the analysis. We obtain the scRNA-seq data of five HER2-positive tumors from GEO database (accession number: GSE176078) [39] as the reference scRNA-seq data for deconvolution. The reference data consists of 29,733 genes and 19,311 cells from 9 cell types.

For the CVAE component of SDePER, we select the top 1500 highly variable genes and 824 cell-type marker genes, corresponding to 1942 unique genes. Each pseudo-spot is assumed to contain 20–70 cells based on estimation by other cancer ST studies using the same ST platform [68]; 290 cell-type marker genes are used in GLRM component. The hyperparameter λ_r is chosen to be 1.931 and λ_l to be 37.276 based on cross-validation. The running time of SDePER is 1.8 h in total using a 32-core, 100 GB RAM, Intel Xeon 2.6 GHz CPU machine.

Idiopathic pulmonary fibrotic lung dataset

We measured the ST data of a human IPF lung sample from an explanted lung of a patient with end-stage IPF explanted lung (Yale IRB:1601017047) using the $10 \times s$ Genomics Visium platform, which is a complex and challenging sample with vague structure and different lung compartments including the bronchi, vascular, mesenchyme, and immune compartment. We selected one block of frozen lung tissue obtained from a patient with Idiopathic Pulmonary Fibrosis (IPF). Sections of $10 \mu\text{m}$ fresh frozen samples were cut from the blocks onto Visium slides ($10 \times$ Genomics) and processed according to the manufacturer's protocol tissue sections were hematoxylin and eosin stained and finally imaged ($20 \times$) using a scanning microscope (EvosM700, ThermoFischer Scientific). Tissue was permeabilized and mRNAs were hybridized to the barcoded capture probes directly underneath. cDNA synthesis connects the spatial barcode and the captured mRNA. After RT and amplification by PCR, dual-indexed libraries were prepared as in the $10 \times$ Genomics protocol and sequenced (two samples/HiSeq 6000 flow cell) with read lengths 28 bp R1, 10 bp i7 index, 10 bp i5 index, and 90 bp R2. Base calls were converted to reads with the software SpaceRanger's implementation mkfastq (SpaceRanger v1.2.2). Multiple fastq files from the same library and strand were catenated to single files. Read2 files were subjected to two passes of contaminant trimming with cutadapt (v1.17): for the template switch oligo sequence (AAGCAGTGGTATCAACGCAGAGTACATGGG) anchored on the 5' end and for poly(A) sequences on the 3' end. Following trimming, read pairs were removed if the read2 was trimmed below 30 bp. Visium libraries were mapped on the human genome ($10 \times$ -provided GRCh38 reference), using STARsolo (STARsolo v2.7.6a).

This data measured the expression of 60,651 genes at 4992 spatial locations. We filtered out the spatial location not covered by tissue and the genes not expressed on all spatial locations. Finally, we performed cell-type deconvolution on 32,078 genes and 3532 spatial spots.

We used the scRNA-seq data from an IPF lung in a previous study as the reference [44]. This dataset consists of 60,651 genes and 12,070 cells. These cells have already been annotated into 39 cell types. Some of the cell types had insufficient cells to provide sufficient information to perform deconvolution. We reannotated the scRNA-seq dataset with 44 cell types in total and selected major cell types with a sufficient number of cells. Therefore, we only considered 11,227 cells from 26 major cell types, and we further filtered out the genes not expressed on these cells. Finally, a final set of 35,483 genes and 11,227 cells serves as the reference scRNA-seq data for deconvolution.

We choose the top 2000 highly variable genes and a set of manually selected 2534 marker genes, corresponding to 3101 unique genes for the CVAE component in SDePER. The number of cells per pseudo-spot is set to be 2–10 cells as provided by expert advice; 1788 cell-type marker genes are used in GLRM component. The hyperparameter λ_r is chosen to be 0.72 and λ_l to be 13.895. The running time of SDePER is 8.58 h in total using a 64-core, 100 GB RAM, Intel Xeon 2.6 GHz CPU machine.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-024-03416-2>.

Supplementary Material 1.

Supplementary Material 2.

Peer review information

Zhana Duren, Kevin Pang, and Veronique van den Berghe were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional File 2.

Authors' contributions

Z.W. and X.Y. conceived the idea and provided funding support. N.L., Y.L., A.A., Z.W., and X.Y. designed the study. N.L., Y.L., J.Q., and G.X. developed the method, implemented the software, and performed the simulations. Y.L., N.L., J.Q., J.Z., N.W., X.H., and W.J. conducted the real data analysis. A.J., T.S.A., I.R., R.H., and N.K. provided the IPF dataset. A.J., T.S.A., and N.K. aided in the result interpretation. Y.L., N.L., J.Q., Z.W., and X.Y. wrote the manuscript. Z.W. and X.Y. supervised the research. All authors read and approved the final manuscript.

Funding

This study was supported by the National Institutes of Health (NIH) grants R01LM014087 (to X.Y. and Z.W.), R21LM012884 (to X.Y.), National Science Foundation (NSF) grant DMS1916246 (to Z.W.). A.J. is supported by funding from Fond de dotation du Souffle, Philippe Foundation, Bourse de Mobilite CHU de Caen, Bourse de mobilite internation interregion Nord Ouest G4.

Availability of data and materials

The SDePER implementation is freely available at Github (<https://github.com/az7jh2/SDePER>) [69] and Zenodo (<https://zenodo.org/doi/https://doi.org/10.5281/zenodo.8328020>) [70]. The source code is released under MIT license. A Docker image of SDePER is also freely available at Docker Hub (<https://hub.docker.com/r/az7jh2/sdeper>) [71]. The scripts used to conduct all the simulation and real data analyses are freely available at Github (https://github.com/az7jh2/SDePER_Analysis) [72] and Zenodo (<https://zenodo.org/doi/https://doi.org/10.5281/zenodo.13702536>) [73] together with all the simulated data and real data. This study assembles five publicly available datasets and one private dataset generated in the laboratory of Dr. Naftali Kaminski. The public datasets used in the simulation studies include the STARmap data (<https://kangaroo-goby.squarespace.com/data>) and inDrops data (GSE102827) [74] with the external reference data (GSE115746) [75]. The three public datasets used in the real data analyses include the MOB data (<https://www.spatialresearch.org/resources-published-datasets/doi-10-1126science-aaf2403/>) [76] with its reference data (GSE121891) [77], melanoma data (<https://www.spatialresearch.org/resources-published-datasets/doi-10-1158-0008-5472-can-18-0747/>) with its reference data (GSE115978) [78], and breast cancer data [79] with its reference data (GSE176078) [80]. The private dataset includes the IPF data (GSE231385) [81] and its reference scRNA-seq data (GSE136831) [82].

Declarations

Ethics approval and consent to participate

No ethical approval was required for this study. All public datasets used in the paper were generated by other organizations that have obtained ethical approval.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA. ²SJTU-Yale Join Center for Biostatistics and Data Science, Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China. ³The Second Affiliated Hospital of The Chinese University of Hong Kong, Shenzhen, Shenzhen, Guangdong, China. ⁴Department of Mathematical Sciences, University of Nevada, Las Vegas, NV, USA. ⁵Section of Pulmonary, Critical Care and Sleep Medicine, Yale School of Medicine, New Haven, CT, USA. ⁶Service de Pneumologie, Centre de Competences de Maladies Pulmonaires Rares, CHU de Caen UNICAEN, CEA, CNRS, ISTCT/CERVOxy Group, GIP CYCERON, Normandie University, Caen, France. ⁷Department of Pathology, Yale School of Medicine, New Haven, CT, USA. ⁸Department of Medicine, Baylor College of Medicine, Houston, TX, USA. ⁹Department of Biomedical Informatics & Data Science, Yale School of Medicine, New Haven, CT, USA.

Received: 26 November 2023 Accepted: 1 October 2024

Published online: 14 October 2024

References

- Asp M, Bergenstrahle J, Lundeberg J. Spatially resolved transcriptomes-next generation tools for tissue exploration. *BioEssays*. 2020;42(10):e1900221.
- Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-cell genomics. *Nat Biotechnol*. 2016;34(11):1145–60.
- Li YH, et al. Visualization and analysis of gene expression in stanford type A aortic dissection tissue section by spatial transcriptomics. *Front Genet*. 2021;12:698124.
- Stahl PL, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*. 2016;353(6294):78–82.
- Stickels RR, et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat Biotechnol*. 2021;39(3):313–9.
- Vickovic S, et al. High-definition spatial transcriptomics for in situ tissue profiling. *Nat Methods*. 2019;16(10):987–90.
- Thrane K, et al. Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. *Cancer Res*. 2018;78(20):5970–9.
- Andersson A, et al. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nat Commun*. 2021;12(1):6012.
- Asp M, et al. A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. *Cell*. 2019;179(7):1647–1660 e19.
- Janosevic D, et al. The orchestrated cellular and molecular responses of the kidney to endotoxin define a precise sepsis timeline. *Elife*. 2021;10.
- Li B, et al. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nat Methods*. 2022;19(6):662–70.
- Cable DM, et al. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat Biotechnol*. 2022;40(4):517–26.
- Andersson A, et al. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun Biol*. 2020;3(1):565.
- Lopez R, et al. DestVI identifies continuums of cell types in spatial transcriptomics data. *Nat Biotechnol*. 2022;40(9):1360–9.
- Kleshchevnikov V, et al. Cell 2location maps fine-grained cell types in spatial transcriptomics. *Nat Biotechnol*. 2022;40(5):661–71.
- Biancalani T, et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat Methods*. 2021;18(11):1352–62.
- Long Y, et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nat Commun*. 2023;14(1):1155.
- Miller BF, et al. Reference-free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data. *Nat Commun*. 2022;13(1):2339.
- Tu, JJ, et al., EnDecon: cell type deconvolution of spatially resolved transcriptomics data via ensemble learning. *Bioinformatics*. 2023;39(1).
- Li H, et al. SD2: spatially resolved transcriptomics deconvolution through integration of dropout and spatial information. *Bioinformatics*. 2022;38(21):4878–84.
- Sun D, et al. STRIDE: accurately decomposing and integrating spatial transcriptomics using single-cell RNA sequencing. *Nucleic Acids Res*. 2022;50(7):e42.
- Song, Q. and J. Su, DSTG: deconvoluting spatial transcriptomics data through graph-based artificial intelligence. *Brief Bioinform*. 2021;22(5).
- Rodrigues SG, et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science*. 2019;363(6434):1463–7.
- Ma Y, Zhou X. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nat Biotechnol*. 2022;40(9):1349–59.
- Elosua-Bayes M, et al. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res*. 2021;49(9):e50.
- Dong R, Yuan GC. SpatialDWLS: accurate deconvolution of spatial transcriptomic data. *Genome Biol*. 2021;22(1):145.
- Danaher P, et al. Advances in mixed cell deconvolution enable quantification of cell types in spatial transcriptomic data. *Nat Commun*. 2022;13(1):385.
- Liu Z, et al. SONAR enables cell type deconvolution with spatially weighted Poisson-Gamma model for spatial transcriptomics. *Nat Commun*. 2023;14(1):4727.
- Chen J, et al., A comprehensive comparison on cell-type composition inference for spatial transcriptomics data. *Brief Bioinform*. 2022;23(4).
- Yan, L. and X. Sun. Benchmarking and integration of methods for deconvoluting spatial transcriptomic data. *Bioinformatics*. 2023;39(1).
- Zhang Y, et al. Deconvolution algorithms for inference of the cell-type composition of the spatial transcriptome. *Comput Struct Biotechnol J*. 2023;21:176–84.
- Shang L, Zhou X. Spatially aware dimension reduction for spatial transcriptomics. *Nat Commun*. 2022;13(1):7203.
- Zhao E, et al. Spatial transcriptomics at subspot resolution with BayesSpace. *Nat Biotechnol*. 2021;39(11):1375–84.
- Sohn, K., H. Lee, and X. Yan, Learning structured output representation using deep conditional generative models. *Advances in Neural Information Processing Systems*. 2015;28.
- Wang, X., et al., Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*. 2018;361(6400).
- 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 Rep*. 2018;25(10):2689–2703 e3.
- Tirosh I, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*. 2016;352(6282):189–96.
- Jerby-Arnon, L., et al., *A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade*. *Cell*, 2018. **175**(4): p. 984–997 e24.

39. Wu SZ, et al. A single-cell and spatially resolved atlas of human breast cancers. *Nat Genet.* 2021;53(9):1334–47.
40. Ueda Y, et al. Overexpression of HER2 (erbB2) in human breast epithelial cells unmasks transforming growth factor beta-induced cell motility. *J Biol Chem.* 2004;279(23):24505–13.
41. Rossi A, et al. Stromal and immune cell dynamics in tumor associated tertiary lymphoid structures and anti-tumor immune responses. *Front Cell Dev Biol.* 2022;10:933113.
42. Bergomas F, et al. Tertiary intratumor lymphoid tissue in colo-rectal cancer. *Cancers (Basel).* 2011;4(1):1–10.
43. Sautes-Fridman C, et al. Tertiary lymphoid structures in the era of cancer immunotherapy. *Nat Rev Cancer.* 2019;19(6):307–25.
44. Adams TS, et al. Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci Adv.* 2020;6(28):eaba1983.
45. Vannan, A., et al., Image-based spatial transcriptomics identifies molecular niche dysregulation associated with distal lung remodeling in pulmonary fibrosis. *bioRxiv*, 2023.
46. Burda, Y., R. Grosse, and R. Salakhutdinov, *Importance weighted autoencoders*. eprint [arXiv:1509.00519v4](https://arxiv.org/abs/1509.00519v4) [cs.LG], 2015.
47. Cremer, C., Q. Morris, and D. Duvenaud, Reinterpreting importance-weighted autoencoders. eprint [arXiv:1704.02916v2](https://arxiv.org/abs/1704.02916v2) [stat.ML], 2017.
48. Cybenko G. Approximation by superpositions of a sigmoidal function. *Math Control Signals Systems.* 1989;2(4):303–14.
49. Csáji, B.C., Approximation with artificial neural networks, in Faculty of Sciences. 2001, Eötvös Loránd University: Hungary. p. 48.
50. Lopez R, et al. Deep generative modeling for single-cell transcriptomics. *Nat Methods.* 2018;15(12):1053–8.
51. Kingma, D. and J. Ba, *Adam: a method for stochastic optimization*. eprint [arXiv:1412.6980v9](https://arxiv.org/abs/1412.6980v9) [cs.LG], 2014.
52. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* 2018;19(1):15.
53. Haghverdi L, et al. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol.* 2018;36(5):421–7.
54. Korsunsky I, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods.* 2019;16(12):1289–96.
55. Stuart T, et al. Comprehensive Integration of Single-Cell Data. *Cell.* 2019;177(7):1888–1902 e21.
56. Swami, A. Non-Gaussian mixture models for detection and estimation in heavy-tailed noise. in 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100). 2000. IEEE.
57. Quail DF, Joyce JA. Microenvironmental regulation of tumor progression and metastasis. *Nat Med.* 2013;19(11):1423–37.
58. Riquelme PA, Drapeau E, Doetsch F. Brain micro-ecologies: neural stem cell niches in the adult mammalian brain. *Philos Trans R Soc Lond B Biol Sci.* 2008;363(1489):123–37.
59. Swain PS, Elowitz MB, Siggia ED. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Natl Acad Sci U S A.* 2002;99(20):12795–800.
60. Zhang J, Li L. Stem cell niche: microenvironment and beyond. *J Biol Chem.* 2008;283(15):9499–503.
61. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc.* 2006;101(476):1418–29.
62. Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. *Math Program.* 1989;45(1):503–28.
63. Virtanen P, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods.* 2020;17(3):261–72.
64. Tuck J, Barratt S, Boyd S. A distributed method for fitting Laplacian regularized stratified models. *J Mach Learn Res.* 2021;22(60):1–37.
65. Bradski G. The openCV library. *Dr Dobb's Journal of Software Tools.* 2000;25(11):120–5.
66. Hrvatin S, et al. Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nat Neurosci.* 2018;21(1):120–9.
67. Tasic B, et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature.* 2018;563(7729):72–8.
68. Moncada R, et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat Biotechnol.* 2020;38(3):333–42.
69. Liu, Y., et al., SDePER. Github. <https://github.com/az7jh2/SDePER> (2024).
70. Liu, Y., et al., SDePER. Zenodo. <https://zenodo.org/doi/10.5281/zenodo.8328020> (2024).
71. Liu, Y., et al., SDePER Docker Image. Docker Hub. <https://hub.docker.com/r/az7jh2/sdeper> (2024).
72. Liu, Y., SDePER Analysis Scripts. Github. https://github.com/az7jh2/SDePER_Analysis (2024).
73. Liu, Y., SDePER Analysis Scripts. Zenodo. <https://zenodo.org/doi/10.5281/zenodo.13702536> (2024).
74. Hrvatin, S., et al., inDrop Data. Datasets. Gene Expression Omnibus. <http://identifiers.org/geo/GSE102827> (2017).
75. Tasic, B., et al., STARmap Data External Reference. Datasets. Gene Expression Omnibus. <http://identifiers.org/geo/GSE115746> (2018).
76. Stahl, P.L., et al., MOB Data. Datasets. Bioproject. <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA316587> (2016).
77. Tepe, B., et al., MOB Reference Data. Datasets. Gene Expression Omnibus. <http://identifiers.org/geo/GSE121891> (2018).
78. Jerby-Arnon, L., et al., *Melanoma Reference Data*. Datasets. Gene Expression Omnibus. <http://identifiers.org/geo/GSE115978> (2018).
79. Andersson A, et al. Breast Cancer Data Datasets. 2020. Zenodo. <https://doi.org/10.5281/zenodo.4751624>.
80. Wu, S.Z., et al., Breast Cancer Reference Data. Datasets. Gene Expression Omnibus. <http://identifiers.org/geo/GSE176078> (2021).
81. Liu, Y., et al., IPF Visium Data. Datasets. Gene Expression Omnibus. <http://identifiers.org/geo/GSE231385> (2024).
82. Adams, T.S., et al., IPF Reference Data. Datasets. Gene Expression Omnibus. <http://identifiers.org/geo/GSE136831> (2020).

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.