Spatially Informed Cell Type Deconvolution for Spatial Transcriptomics

Ying Ma[1], **Xiang Zhou**[1,2]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

[2]Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA

Abstract

Many spatially resolved transcriptomic technologies do not have single-cell resolution but measure the average gene expression for each spot from a mixture of cells of potentially heterogeneous cell types. Here, we introduce a deconvolution method, conditional autoregressive deconvolution (CARD), that combines cell type–specific expression information from single-cell RNA sequencing (scRNA-seq) with correlation in cell type composition across tissue locations. Modeling spatial correlation allows us to borrow the cell-type composition information across locations, improving accuracy of deconvolution even with a mismatched scRNA-seq reference. CARD can also impute cell type compositions and gene expression levels at unmeasured tissue locations, enable the construction of a refined spatial tissue map with a resolution arbitrarily higher than that measured in the original study, and perform deconvolution without a scRNA-seq reference. Applications to four datasets including a pancreatic cancer dataset identified multiple cell types and molecular markers with distinct spatial localization that define the progression, heterogeneity, and compartmentalization of pancreatic cancer.

Introduction

Spatially resolved transcriptomic technologies perform gene expression profiling on many tissue locations with spatial localization information[1], enabling the characterization of transcriptomic landscape on tissues[2–10]. Despite fast technological development, however, most technologies are of limited spatial resolution. In particular, almost all sequencing-based technologies collect expression measurements on tissue locations that consist of a few to a few dozen single cells belonging to potentially distinct cell types[11–14]. Because each measured location contains a mixture of cells, these sequencing-based technologies effectively quantify the average expression level across many cells on the location. Consequently, performing cell type deconvolution on tissue locations becomes an essential

correspondence to XZ (xzhousph@umich.edu).

Author Contributions
XZ conceived the idea and provided funding support. YM and XZ designed the experiments. YM developed the method, implemented the software, performed simulations, and analyzed real data. YM and XZ wrote the manuscript.

Competing interests
The authors declare that they have no competing interests.

analytic task for disentangling the spatial localization of cell types and characterizing the complex tissue architecture[15,16].

Deconvolution of spatial transcriptomics data requires cell type specific gene expression information and tailored spatial methods. Cell type specific gene expression information are nowadays readily available from single-cell RNA sequencing (scRNA-seq) studies[17], which have been previously used for deconvoluting bulk RNA-seq data[18] by recently developed deconvolution methods including MuSiC[19], SCDC[20], and Bisque[21]. These methods can in principle be directly applied to spatial transcriptomics and are being adapted so by several recently developed methods[22–31] such as RCTD[23], stereoscope[29], SPOTlight[22], cell2location[30], and spatialDWLS[31] (details in Supplementary Notes). All these methods, however, do not make use of the rich spatial localization information available in spatial transcriptomics.

Spatial localization information in spatial transcriptomics measures the relative distance between tissue locations and contains potentially invaluable information for deconvolution. Specifically, a tissue is composed of multiple cell types that are segregated in a spatially correlated fashion into tissue domains[32–35], which are characterized by a domain-specific composition of cell types, with similar cell types colocalized spatially[36,37]. Histological characterization of various tissues[12,38–40], including the Hematoxylin and Eosin (H&E) staining images accompanying spatial transcriptomics datasets[12,14], highlight the spatial segregation of cell types and neighboring cell type composition similarity. In single cell resolution spatial transcriptomics[41,42], we also observed that similar cell types tend to colocalize, with colocalization pattern decaying with distance (Supplementary Figures 1-2). Consequently, neighboring locations on the tissue likely contain more similar cell type compositions as compared to locations that are far away. Therefore, modeling the neighborhood similarity in cell type compositions and accommodating their spatial correlation would allow us to borrow composition information across locations on the entire tissue section to enable accurate deconvolution of spatial transcriptomics on each individual location.

Here, we develop a method, named Conditional AutoRegressive based Deconvolution (CARD), to perform such spatially informed deconvolution of cell types for spatial transcriptomic. CARD builds upon a non-negative matrix factorization model to use the cell type specific gene expression information from scRNA-seq data for deconvoluting spatial transcriptomics. A unique feature of CARD is its ability to accommodate the spatial correlation structure in cell type composition across tissue locations by a conditional autoregressive modeling assumption[43,44]. As a result, CARD can take advantage of the spatial correlation structure to enable accurate and robust deconvolution of spatial transcriptomics across technologies with different spatial resolutions and in the presence of mismatched scRNA-seq references. In addition, modeling spatial correlation allows CARD to impute cell type compositions as well as gene expression levels on new locations of the tissue, facilitating the construction of a refined spatial map with an arbitrarily high resolution for any spatial transcriptomics technologies -- both these features are in direct contrast to a recent method BayesSpace[45] that can only enhance Spatial Transcriptomics (ST) or 10x Visium data with a fixed resolution of either six or nine times higher than that of the original. Importantly, an extension of CARD is also capable of performing reference-free deconvolution without a scRNA-seq reference. We develop a computationally efficient algorithm for constrained maximum likelihood inference, making CARD scalable to data with tens of thousands of spatial locations and tens of thousands of genes. We illustrate the benefits of CARD through extensive simulations and applications to four published spatial transcriptomics studies with distinct technologies, spatial resolutions, tissue structures, and scRNA-seq references.

Results

Simulations

CARD is described in Materials and Methods, with its technical details provided in Supplementary Notes and its method schematic shown in Figure 1. We performed simulations to evaluate the performance of CARD and compared it with six existing deconvolution methods: MuSiC, SPOTlight, RCTD, cell2location, spatialDWLS, and stereoscope (details in Materials and Methods). Briefly, we used a scRNA-seq data[46] to construct spatial transcriptomics and we varied a noise level parameter p_n to modify cell type compositions and spatial correlation patterns across locations (Supplementary Figures 3-4). The simulated data are realistic, preserving data features observed in the published spatial transcriptomics data (Supplementary Figure 5). We examined four simulation settings, each of which consists of five simulation replicates. In each replicate, we applied various deconvolution methods to deconvolute the spatial transcriptomics data, using either the same set of scRNA-seq data or its modified version or another set as reference. We then followed[19] and quantified the deconvolution performance by computing the root mean square error (RMSE) between the estimated cell type composition and the underlying truth on each location. We primarily displayed RMSE difference plots where we contrasted the RMSE of other methods with respect to CARD following[47,48]. We kept the original RMSE and rank plots in the supplements, which show consistent results.

We first explored a baseline analysis scenario (scenario I), where we used the same scRNA-seq data used in the simulations for deconvolution. Here, CARD outperforms all other deconvolution methods across all simulation settings (median RMSE = 0.079), with 9%, 8%, 33%, 7%, 23%, and 18% improvement in terms of RMSE as compared to MuSiC (0.087), RCTD (0.086), SPOTlight (0.118), cell2location (0.085), spatialDWLS (0.103), and stereoscope (0.096), respectively (Figure 2 scenario I, Supplementary Figures 6-8). In addition, CARD identifies the dominant cell type on each spatial location accurately as measured by AUC and ARI (Supplementary Figure 9).

To examine the robustness of different deconvolution methods, we explored four additional scenarios (Supplementary Notes) where we either removed one cell type in the scRNA-seq reference (scenario II); added one cell type (scenario III); used miss-classified cell types (scenario IV); or used another scRNA-seq data sequenced on a different platform for deconvolution (scenario V). Compared to scenario I, the performance of all methods remains similar in scenarios III (except SPOTlight) and generally reduces in other scenarios, though their relative rank remains largely consistent across scenarios. In addition, CARD outperforms the other methods in all settings, with its performance gain more apparent than scenario I (Figure 2). Specifically, in scenario II, CARD loses a median of 3% accuracy across settings as compared to using the original scRNA-seq data (Supplementary Notes). However, CARD is more accurate than the other methods across settings with 13% ~ 32% accuracy improvement (Figure 2, Supplementary Figures 10-14). In scenario III, CARD only loses a median of 0.4% accuracy across settings as compared to using the original scRNA-seq data. It remains the most accurate method across settings with 7% ~ 40% accuracy improvement over the other methods (Figure 2, Supplementary Figure 15). In scenario IV, CARD loses a median of 4% accuracy as compared to using the original scRNA-seq data (Figure 2, Supplementary Figure 16). However, CARD is again more accurate than the other methods across settings (Figure 2, Supplementary Figure 17), with 6% ~ 32% accuracy improvement across misclassified cell types (Supplementary Figure 18). In scenario V, CARD loses a median of 10% accuracy across settings as compared to using the original scRNA-seq data. But it remains the most accurate method across settings with 5% ~ 35% accuracy improvement over the other methods (Figure 2, Supplementary Figure 19).

We examined the deconvolution accuracy of different methods at distinct cell type resolution levels (Supplementary Notes) and found that the deconvolution accuracy of most methods improved initially with increasing number of sub-cell types (Supplementary Figure 20) and reached a saturation point with sufficiently large number of sub-cell types, where many sub-cell types are no longer distinguishable from each other (Supplementary Figure 21). Regardless of the cell type resolution, the relative performance of most deconvolution methods remains consistent (Supplementary Figure 22). We also carried out additional model-based simulations where we can more effectively control for spatial correlation (Supplementary Notes) and found as expected that the advantage of CARD over the other methods shows a clear dependency on spatial correlation (Supplementary Figure 23).

Mouse Olfactory Bulb Data

We applied CARD and the other methods to analyze four published spatial transcriptomics data that include two obtained from Spatial Transcriptomics (ST), one from Slide-seq, and one from 10x Visium (details in Supplementary Notes). In each data, the majority of marker genes (92% by Moran’s I test and 54% by Geary’s C test) display statistically significant spatial autocorrelation (adjusted p-value < 0.05; Supplementary Table 1), with the semivariance generally increasing with distance (Supplementary Figure 24) and the expression correlation between locations decreasing with distance (Supplementary Figure 25), supporting cell type composition similarity between neighboring locations. We used scRNA-seq data from sequencing platforms different from the spatial transcriptomics for deconvolution.

We first examined the mouse olfactory bulb (MOB) data[14], where we used a scRNA-seq data[49] from 10x Chromium on the same tissue for deconvolution (Supplementary Tables 2-3). The MOB data consists of four main anatomic layers organized in an inside out fashion annotated based on H&E staining: the granule cell layer (GCL), the mitral cell layer (MCL), the glomerular layer (GL), and the nerve layer (ONL) (Figure 3A, details in Materials and

Methods). The cell type compositions inferred by CARD accurately depict such expected layered structure[50], as is evident by visualizing either the first principal component (PC1) of the estimated cell type composition matrix (Supplementary Figure 26) or the inferred dominant cell types (Figure 3B, Supplementary Table 4, Supplementary Figure 27). In contrast, MuSiC, SPOTlight, spatialDWLS, and stereoscope were unable to distinguish the three outer layers from each other, while RCTD was unable to clearly distinguish the nerve layer from the glomerular layer. RCTD, cell2location, and spatialDWLS showed a blurry boundary between GCL and MCL/GL on top of the tissue section, while cell2location could not clearly identify the boundaries between MCL and GL.

Careful examination of the cell type composition and corresponding cell type marker genes in different layers further confirm the accuracy of CARD deconvolution (Figure 3C-3D, Supplementary Notes). For example, CARD distinguished correctly the adjacent MCL and GL, with distinct enrichment of mitral/tufted cells and periglomerular cells in the two layers, respectively, despite the similarity between these two cell types; while others cannot (Supplementary Figures 28-29). We also observed that multiple cell types inferred by CARD show spatially co-localization patterns (Figure 3E, Supplementary Notes).

A key benefit of CARD is its ability to model the spatial correlation structure across spatial locations, which facilitates the imputation of cell type composition and gene expression on locations not measured in the original study. We performed location masking analysis for CARD and validated that the imputed expression levels are highly consistent with the truth regardless of the percentage of masked locations (Pearson's correlation=0.44–0.56; Figure 3F, Supplementary Figure 30). Imputation on new locations allows us to construct a refined spatial map of cell type composition or gene expression with arbitrarily high spatial resolution (details in Materials and Methods), which captures fine grained details of the layered structure in the olfactory bulb (Figure 3G, Supplementary Figures 31-32) and facilitates the identification of marker genes with spatial expression patterns (Supplementary Figure 33, Supplementary Notes). In contrast, the fixed resolution enhancement by BayesSpace failed to capture the expected spatial expression pattern for a few marker genes at high resolution (Supplementary Figures 34-35). We quantitatively compared the performance of CARD and BayesSpace for resolution enhancement by performing clustering analysis on the imputed expression data (details in Materials and Methods). We found that the clustering results based on CARD displayed a clear inside-out layered structure that resembles the anatomic organization of the olfactory bulb, more so than that obtained with the original scale data or by BayesSpace (Supplementary Figure 36). CARD is also computationally efficient: CARD takes only 0.4 seconds to construct the refined expression map for all genes, is 5,816 times faster than BayesSpace, and represents a scalable solution for fine map reconstruction in much larger datasets.

Human Pancreatic Ductal Adenocarcinomas Data

The second data we examined is a human pancreatic ductal adenocarcinomas (PDAC) data from spatial transcriptomics[51]. For deconvolution, we first used a matched scRNA-seq data for the same patient obtained through inDrop[51] (denoted as PDAC-A). The PDAC data contains multiple tissue regions (cancer, pancreatic, ductal, and stroma regions) annotated by histologists based on H&E staining[51] (Figure 4A). Through deconvolution, CARD located various pancreatic and tumoral cell types into different tissue regions (Figure 4B). The PC1 of the estimated cell type composition matrix from CARD can clearly capture a gross regional segregation between cancer and non-cancer regions, between the ductal and stroma regions, and between the pancreatic and ductal regions. In contrast, none of the other methods were as effective in differentiating these regions (Supplementary Figures 37 - 40, Supplementary Notes). The dominant cell types on each location from CARD also capture the segregation between cancer and non-cancer regions (Supplementary Figure 41), with the neoplastic cells such as cancer clone A and clone B cells highly enriched in the former (Wilcoxon test p-value = 1.9e-48, 1.1e-43 respectively, Figure 4D). CARD also reveals distinct distribution of two macrophage subpopulations between the cancer and non-cancer regions (Figure 4D), representing a key functional signature of the regional compartmentalization of the cancer tissue that were missed by the other methods (Supplementary Figure 42).

CARD further divides the cancer region into two sub-regions, a pattern missed by the other methods (Figure 4B-4C, Supplementary Figures 41, 43): an upper subregion dominated by cancer clone A cells with an enrichment of marker gene *Tm4sf1*, and a bottom subregion dominated by cancer clone B cells with an enrichment of marker gene *S100a4* (Figure 4B-4C, Supplementary Figure 43). *S100A4* is a prognostic marker for early-stage pancreatic cancer and its spatial enrichment suggests that the bottom cancer subregion is likely an early cancer region. In contrast, *Tm4sf1* is essential for PDAC migration and invasion[52–54] and its spatial enrichment suggests that the upper cancer subregion is likely a late-stage cancer region with metastasis capability. Indeed, the upper cancer subregion is also detected by CARD to be enriched with fibroblast cells, along with fibroblast cell marker gene Cd248 (Figure 4C), a cell type known to be associated with advanced TNM stage[55].

CARD also localizes many other cell types into specific tissue regions, consistent with the expression pattern of the corresponding marker genees (Figure 4B-4C, Supplementary Figure 43, Supplementary Notes). In contrast, none of the other methods capture the expected spatial localization of both ductal centroacinar and terminal ductal cells. In addition, acinar cells inferred by CARD are mainly enriched in the normal pancreatic tissue region; but they are inferred by the other methods to be either absent in the pancreatic region or diffused outward from the pancreatic region to the stroma region and cancer region. Several cell types inferred by CARD are also co-localized spatially in PDAC (Figure 4F), such as those between ductal high hypoxic cells and cancer cells and those between endothelial cells and fibroblast cells, supporting the role of the former in forming the hypoxic and nutrient-poor tumor microenvironment (TME) and the role of the later in pancreatic-cancer stroma interaction of the tumor microenvironment[56,57]. The mean cell type proportions inferred by CARD in the ST data are also highly correlated with that measured in the scRNA-seq dataset obtained on the same patient, more so than that obtained by the other methods (Figure 4E).

Next, we examined the robustness of deconvolution by using unmatched scRNA-seq datasets (Supplementary Table 2, Supplementary Notes). Despite the platform and sample differences in the scRNA-seq references, we found that the estimated cell type compositions for the major cell types are consistent across different scRNA-seq references, with the highest consistency achieved by CARD (Supplementary Figure 44). Regardless of which unmatched scRNA-seq data was used, CARD shows superior performance than the other methods in capturing the gross segregation of cancer and non-cancer regions, identifying two distinct cancer subregions, accurately localizing cell types, and revealing a possible TME supporting tumor progression[58–61] (Supplementary Figures 45 – 46, Supplementary Notes).

Finally, we found that the imputed gene expression by CARD are highly consistent with the truth across a range of masking percentages (Pearson's correlation=0.29–0.52; Figure 4G, Supplementary Figure 47). Such consistency is higher when the matched scRNA-seq data from the same patient is used as the reference, as compared to using an unmatched scRNA-seq data (Supplementary Figure 48). The high-resolution spatial map of cell type composition or gene expression obtained by CARD also reveals refined boundaries between different tissue subregions (Supplementary Figure 49) and the spatial expression pattern of marker genes (Figure 4H, Supplementary Figure 50). Besides marker genes, CARD also discovered multiple genes that display clear spatial expression pattern in the refined spatial map but not in the original map (Supplementary Figure 51, Supplementary Notes). In contrast, the high-resolution map of BayesSpace does not show a clear pattern of multiple known marker genes (Supplementary Figure 52) and additional genes (Supplementary Figure 53). Clustering analysis on CARD imputed high resolution data also revealed clear segregation of the two cancer sub-regions, the normal pancreatic region, and the ductal region, more so than the original data or the refined data by BayesSpace (Supplementary Figure 54).

Mouse Hippocampus Data from Multiple Sources

We analyzed two mouse hippocampus datasets: one directly on hippocampus measured using Slide-seq V2[62] and the other on a coronal brain section containing hippocampus measured using 10x Visium[12]. We used the hippocampus scRNA-seq dataset by Drop-seq[23,63] for deconvoluting both datasets (Supplementary table 2). We only applied cell2location to the 10x Visium data but not the Slide-seq V2 data due to its heavy computational burden.

The hippocampus primarily consists of three regions -- the CA1/CA2 region, the CA3 region, and the dentate gyrus -- all visualizable by total UMI counts per location displayed on the tissue (Figure 5A). The cell type compositions inferred by CARD accurately depict the three anatomic structures of hippocampus, with the compositional PC1 capturing the curved shape of hippocampus accurately, more so than the other three methods (Figure 5A, Supplementary Figure 55). The dominant cell type on each location inferred by CARD also matches the expectation (Figure 5B): CA1 cells are highly enriched in CA1; CA3 cells mainly localize in CA3; dentate cells reside in a C-shaped ring region of dentate gyrus; ependymal cells form an irregular and columnar shape and line the ventricles of the brain[64]; while choroid cells reside right below the ependymal cells and locate in the choroid plexus[65] along with Cajal-Retzius cells[66] (Supplementary Figure 56). In contrast, MuSiC is unable to localize the main cell types such as CA1 and CA3 cells correctly and thus unable to reveal the main structures of the hippocampus (Figure 5B, Supplementary Figure 57). SPOTlight detects an incorrectly diffused pattern of ependymal cells and incorrectly locates many CA3 cells to the CA1 region or outside hippocampus (Figure 5B, Supplementary Figure 58). RCTD, spatialDWLS, and stereoscope perform similarly, all locating CA3 cells incorrectly in CA1 (Figure 5B; Supplementary Figures 59-61), with the CA1 cell marker gene enriched in locations dominated by CA3 cells inferred by these methods (Supplementary Figure 62). Additionally, they all allocate different cell types to hippocampus structures that appear to be much wider than expected[13,67,68](Figures 5A-5B). Careful examination of marker genes further confirms the accuracy of CARD deconvolution (Figure 5C). We quantified the deconvolution performance of different methods by examining the expression levels of the marker genes on each of the three hippocampal structures inferred based on the estimated cell type composition by different methods. Quantifications again support more accurate deconvolution by CARD than the other methods (Figure 5D; Supplementary Figure 63, Supplementary Notes).

We observed that multiple cell types inferred by CARD are co-localized together (Supplementary Figure 64). The highest co-localization occurs between *Slc17a6/Vglut2* neurons and entorhinal cells, highlighting the cell compositional architecture underlying the hippocampus-entorhinal cortex network[69]. The imputed gene expression by CARD are consistent with the truth across a range of masking percentages (Supplementary Figure 65). Although the resolution of this dataset is already high, the refined spatial map of cell type composition by CARD again reveals refined boundaries between different subregions of hippocampus (Supplementary Figure 66), with the refined gene expression recovers strong spatial pattern for various marker genes (Figure 5E, Supplementary Figure 67) and additional genes (Supplementary Figure 68, Supplementary Notes). We examined the reliability of the refined spatial map by creating a low-resolution version of the Slide-seq V2 data and then applied CARD to construct a refined spatial expression map at the original Slide-seq V2 resolution (Supplementary Notes). We found that the refined spatial map recovers a consistent and sometime stronger spatial pattern than the original Slide-seq V2 data (Supplementary Figures 69-72), supporting the accuracy and effectiveness of refined spatial map construction. Here, we were unable to apply BayesSpace due to both its heavy computational burden and its required input of pixel coordinates that are not available from Slide-seq technologies.

Finally, we examined the hippocampus region from the 10x Visium[12] data. Again, CARD captures the key structures of the hippocampus (Figures 5F-5G). The estimated cell type compositions on CA1, CA3 and dentate gyrus from both CARD and MuSiC matched the corresponding structures on the H&E image, while those from the other methods appear to also occupy regions outside the expected structure boundaries (Figure 5F, Supplementary Figure 73), a pattern confirmed with quantifications (Supplementary Figures 74-75).

Extension of CARD for Reference-free Deconvolution

We further developed CARDfree, an extension of CARD for reference-free cell type deconvolution that does not require a scRNA-seq reference data (Supplementary Notes). CARDfree only requires users to input a list of gene names for previously known cell type markers, which determines the dimensionality of the input gene expression matrix. Compared to CARD, CARDfree yields generally similar cell type composition estimates in the real data, but likely with lower accuracy. For example, CARDfree captures the general tissue domain segregation pattern as CARD in both MOB and PDAC data, though it was unable to differentiate the two cancer sub-regions as CARD did (Supplementary Figure 76). CARDfree does not perform as well as CARD in the high-resolution Slide-seq V2 data and did not identify the CA3 structure based on its estimated cell type proportions, as the Slide-seq V2 data is highly sparse and thus could be benefited from reference-based deconvolution. However, in the hippocampus region of the Slide-seq V2 data, we did notice that CARDfree identified a region with a unique cell type composition (Supplementary Figure 77, CT15 colored in blue) that was not found by other deconvolution methods. This region appears to part of the entorhinal cortex, which consists of endothelial tip cells that are highly related to angiogenesis in mouse brain[70]. The results suggest that reference-free deconvolution may sometimes have added benefits.

Discussion

We have presented CARD for accurate and spatially informed deconvolution of spatial transcriptomics. CARD is computationally efficient: it is 0.8–7,761.8 times faster while using 0.2%−109% of the physical memory as compared to the other deconvolution methods (Supplementary Figure 78, Supplementary Table 5); it is 5,875–7,028 times faster and uses only 14%−17% of the physical memory as compared to BayesSpace in creating refined spatial maps (Supplementary Figures 79-80, Supplementary Table 6). We have demonstrated the benefits of CARD in both simulations and applications to four spatial transcriptomics datasets.

We have primarily focused on examining the sequencing-based technologies that measure the average gene expression from a mixture of cells on each tissue location. Non-sequencing-based technologies, such as seqFISH[71] and MERFISH[72], mostly rely on single molecular fluorescent *in situ* hybridization (smFISH) and are directly of single cell resolution. However, it remains computationally challenging to detect the accurate boundaries between cells on the smFISH image data, especially when the cell density is high[73–75]. Consequently, the expression data measured on each "single cell" in smFISH may consist of transcripts from a mixture of neighboring cells. Therefore, CARD can also be applied to analyze these datasets. In a mouse cortex data from seqFISH+[42], we found that the cell type compositions inferred by CARD clearly displayed a layered structure that resembled the laminar organization of the cortex, with each layer harboring a distinct composition of neuronal populations (Supplementary Figures 81-82).

We have presented an extension of CARD, CARDfree, for reference-free deconvolution. CARDfree requires a post-processing step to correctly label the inferred cell types. Such post-processing often requires cell type specific gene expression profiles and can be challenging to carry out accurately. For example, in PDAC, CARDfree infers cell type composition on each location for 20 inferred cell types. But it is not trivial to find the name for each inferred cell type: for instance, it is not easy to tell whether the inferred cell type #14 (CT14) corresponds to the ductal centroacinar cell or the endothelial cell, as markers for both cell types are enriched in locations with a high proportion of CT14 cells (Supplementary Figure 83). Therefore, new computational algorithms are likely needed for labeling cell types inferred from reference-free deconvolution methods. We also present another extension of CARD (Supplementary Notes) to facilitate the construction of single-cell resolution spatial transcriptomics from non-single-cell resolution spatial transcriptomics (Supplementary Figures 84 - 90). Such extension requires knowing the spatial localization information for all single cells on the tissue, which remains challenging to obtain from non-single-cell resolution spatial transcriptomics. Because the spatial transcriptomics data itself does not contain information for inferring the single cell positions, H&E image segmentation is often required to identify single cells on the tissue and extract their locations. However, common software is not always accurate in inferring the location for single cells (e.g., Supplementary Figure 91). In addition, aligning H&E image with spatial transcriptomics can be computationally challenging[76]. Future efforts are needed to address these challenges.

Additional extensions of CARD are possible. First, CARD models normalized spatial transcriptomes data and could be benefited from extensions for direct modeling of raw count data using an over-dispersed Poisson model[77,78]. Second, we only explored the use of the Gaussian kernel[79] for modeling spatial correlation. Exploring the use of other kernels such as the periodic kernels[79] or incorporating histological image information such as image intensity level as additional coordinates[6,80], which can be readily done in CARD, may capture diverse and rich spatial correlation patterns in the future. Third, the spatial imputation feature of CARD facilitates not only the construction of a refined spatial map but also the selection of scRNA-seq references when multiple scRNA-seq resources are available. Specifically, we can evaluate through data masking the imputation accuracy resulted from pairing with different scRNA-seq references and select the scRNA-seq data with the best imputation accuracy for deconvolution. In PDAC, the matched scRNA-seq indeed produced the best imputation performance and would be selected as the optimal reference data for deconvolution.

Materials and Methods

CARD Method Overview

We present an overview of CARD here, with its technical details provided in Supplementary Notes. CARD is a deconvolution method for spatial transcriptomics studies with regional resolution. These studies perform transcriptomic profiling on multiple tissue locations, each of which contains multiple single cells. CARD aims to estimate the cell type composition on each tissue location while properly accounting for the spatial correlation among them. CARD requires both spatial transcriptomics data and a single cell RNA-seq (scRNA-seq) data as input. The scRNA-seq data serves as a reference and consists of K cell types with a set of G cell type informative genes. Cell types and informative genes in scRNA-seq can be obtained through standard analysis pipelines for clustering and informative gene identification[81,82]. In scRNA-seq, we denote $\boldsymbol{B}$ as the G by K cell type specific expression matrix for the informative genes, where each element represents the mean expression level of an informative gene in a specific cell type. The expression matrix $\boldsymbol{B}$ is commonly referred to as the reference basis matrix. In the spatial transcriptomics data, we denote $\boldsymbol{X}$ as the G by N gene expression matrix for the same set of informative genes measured on N spatial locations. We denote $\boldsymbol{V}$ as the N by K cell type composition matrix, where each row of $\boldsymbol{V}$ represents the proportions of the K cell types on each spatial location. Our objective is to estimate $\boldsymbol{V}$ given both $\boldsymbol{X}$ from the spatial transcriptomics data and $\boldsymbol{B}$ constructed from the scRNA-seq data. To do so, we consider a non-negative matrix factorization model to link the three matrices:

$$\boldsymbol{X} = \boldsymbol{B}\boldsymbol{V}^{\boldsymbol{T}} + \boldsymbol{E}, \tag{1}$$

where each element in $\boldsymbol{V}$ is constrained to be non-negative; and $\boldsymbol{E}$ is an G by N residual error matrix with each element independently and identically following a normal distribution $\boldsymbol{E}_{gi} \sim N\left(0, \sigma_e^2\right)$. A detailed biological interpretation of equation (1) in the context of deconvolution is provided in Supplementary Notes.

The non-negative matrix factorization model in equation (1) has been applied for cell type deconvolution in bulk RNA-seq studies. However, this model is not directly applicable for deconvoluting spatial transcriptomics as it does not account for the spatial correlation structure in the cell type compositions across locations. Intuitively, cell type compositions on two neighboring locations of a tissue are likely to be similar to each other, more so than those on locations that are far away. Consequently, the cell type compositions on neighboring locations contain valuable information for inferring the cell type composition on the location of interest. The similarity in cell type compositions on neighboring locations effectively induces spatial correlation among rows of $\boldsymbol{V}$ in the above factorization model. Thus, modeling spatial correlation in $\boldsymbol{V}$ is relevant for spatial transcriptomics as it would allow us to borrow cell type composition information across spatial locations to enable accurate estimation of $\boldsymbol{V}$. To accommodate the spatial correlation in $\boldsymbol{V}$, we specify a conditional autoregressive (CAR)[43,83,84] modeling assumption on each column of $\boldsymbol{V}$. Specifically, for the column/cell type k, we assume:

$$\boldsymbol{V}_{ik} = \boldsymbol{b}_k + \phi \sum_{j\,=\,1,\, j\, \neq\, i}^{n} \boldsymbol{W}_{ij}\left(\boldsymbol{V}_{jk} - \boldsymbol{b}_k\right) + \boldsymbol{\epsilon}_{ik} \tag{2}$$

where $\boldsymbol{V}_{ik}$ represents the proportion of cell type k on the i-th location; $\boldsymbol{b}_k$ is the k-th cell type specific intercept that represents the average cell type composition across locations; W is a N-by-N non-negative weight matrix with each element $\boldsymbol{W}_{ij}$ specifying the weight used for inferring the cell type composition on the i-th location based on the cell type composition information on the j-th location; ϕ is a spatial autocorrelation parameter that determines the strength of the spatial correlation in cell type composition; and $\boldsymbol{\epsilon}_{ik}$ is the residual error that follows a normal distribution $\boldsymbol{\epsilon}_{ik} \sim N\left(0, \sigma_{ik}^2\right)$. The CAR modeling assumption on V effectively expresses the composition of the k-th cell type on the i-th location, $\boldsymbol{V}_{ik}$, as a weighted summation of the k-th cell type compositions on all other locations, $\boldsymbol{V}_{jk}\ \left(j \neq i\right)$. Consequently, the CAR modeling assumption on V allows us to borrow information across locations to infer the cell type composition on the location of interest.

We follow[79] to express the weight matrix $\boldsymbol{W}$ in the form of a Gaussian kernel function constructed based on the Euclidean distance between pairs of spatial locations (details in Supplementary Notes). The Gaussian kernel function has been widely used to model a range of correlation patterns that decay over distance across tissue locations in many other analytic tasks in spatial transcriptomics [85,86]. While we primarily focus on using a Gaussian kernel for $\boldsymbol{W}$, our method and software can easily incorporate other types of kernels to capture diverse spatial correlation patterns encountered in different data sets. With the Gaussian kernel matrix $\boldsymbol{W}$, we further obtain a row-standardized weight matrix $\widetilde{\boldsymbol{W}}$ through transformation $\widetilde{\boldsymbol{W}}_{ij} = \boldsymbol{W}_{ij}/\boldsymbol{W}_{i+}$, with $\boldsymbol{W}_{i+} = \sum_{j=1}^{n} \boldsymbol{W}_{ij}$. Because the weight matrix and the residual error variance need to satisfy the symmetric condition[87,88], we set $\sigma_{ik}^2 = \lambda_k/\boldsymbol{W}_{i+}$ to ensure $\widetilde{\boldsymbol{W}}_{ij}\sigma_{jk}^2 = \widetilde{\boldsymbol{W}}_{ji}\sigma_{ik}^2$, where λ_k is a scalar. With the above parameterization, we can follow the Brook's Lemma[84,89] to obtain the joint distribution for the N-size column vector $\boldsymbol{V}_k$ as

$$\boldsymbol{V}_k \sim MVN(\boldsymbol{b}_k \mathbf{1}_N, \boldsymbol{\Sigma}_{\boldsymbol{k}}), \tag{3}$$

where $\mathbf{1}_N$ is a N-vector of 1's; $\boldsymbol{\Sigma}_{\boldsymbol{k}} = \left(\boldsymbol{I}_N - \phi\widetilde{\boldsymbol{W}}\right)^{-1}\boldsymbol{M}_{\boldsymbol{k}}$ is a positive definite covariance matrix with $\boldsymbol{M}_{\boldsymbol{k}} = diag\left(\sigma_{1\mathrm{k}}^2, \ldots \sigma_{Nk}^2\right)$; and MVN denotes a multivariate normal distribution (details in Supplementary Notes).

Equations (1) and (3) together define a factor model with a CAR modeling assumption on the latent factors to induce spatial correlation across rows of $\boldsymbol{V}$. By modeling the spatial correlation in $\boldsymbol{V}$, our model allows us to borrow cell type composition information across spatial locations for spatially informed cell type deconvolution. We developed a constrained optimization algorithm in the maximum likelihood framework to estimate the cell type composition matrix $\boldsymbol{V}$, with non-negativity constraints on each of its elements (details in Supplementary Notes). Our algorithm treats the hyper-parameters ($\boldsymbol{b}_k$, $\boldsymbol{\lambda}_\mathrm{k}$, ϕ and σ_e^2) as unknown and infers these parameters based on the data at hand to ensure optimal deconvolution performance. Our algorithm has several computational advantages that makes it highly computationally efficient. First, the modeling framework of CARD is in essence a linear factor model, expressing the mean gene expression profile in the spatial transcriptomics as a linear function of that from scRNA-seq. The linear factor modeling framework streamlines the inference procedure and facilitates scalable computation. Second, CARD makes use of the fast multiplicative updating rules[90,91] for updating the nonnegative cell type composition matrix in each optimization iteration. The multiplicative updating rules allow for algorithmic optimization without explicit inverse of the spatial covariance matrix, which is otherwise required for spatial deconvolution, and which incurs heavy computation burden (Supplementary Notes). Third, CARD takes advantage of the modern computing architecture and explicitly expresses the most computationally intensive part of the algorithm in the form of large matrix operations instead of multiple scalar operations. For example, it updates the entire cell type composition matrix jointly at each optimization iteration instead of updating each element in the cell type composition matrix on each spatial location separately. Finally, while CARD is implemented in R, its core deconvolution algorithm is implemented with an efficient C++ code that is linked back to the main functions of CARD through Rcpp, ensuring scalable computation.

Imputation and Construction of High-Resolution Spatial Maps for Cell Type Composition and Gene Expression

A key feature of CARD is its ability to model the spatial correlation structure in $\boldsymbol{V}$. By modeling the spatial correlation in $\boldsymbol{V}$, CARD can predict and impute the cell type compositions on new, unmeasured, spatial locations on the tissue. Imputing cell type compositions on new locations would allow us to obtain a refined cell type composition map of the tissue with a spatial resolution much higher than that measured in the original study. To enable imputation and construction of a refined cell type composition map, we first outlined the shape of the tissue by applying a two-dimensional concave hull algorithm[92] on the existing locations. We then created an equally spaced grid within the tissue outline and set the number of grid points to exceed the number of spatial locations measured in the original study. We denote the cell type composition matrix on the original N spatial locations as $\boldsymbol{V}$ and denote the corresponding matrix on the $N*$ new locations as $V*$. Based on equation (3), the $(N + N*)$-sized cell type composition vector for the k-th cell type, $(\boldsymbol{V}_k,\ \boldsymbol{V}_k^*)^T$, follows a multivariate normal distribution $MVN(\boldsymbol{b}_k\mathbf{1}_{N + N*}, \boldsymbol{\Sigma})$. We partition the covariance matrix $\boldsymbol{\Sigma}$ into $\begin{bmatrix}\Sigma_{oo} & \Sigma_{on}\\ \Sigma_{no} & \Sigma_{nn}\end{bmatrix}$, where o are the indices that correspond to the original locations while n are the indices that correspond to the new locations. We can then estimate $\boldsymbol{V}_k^*$ via its conditional mean

$$\widehat{\boldsymbol{V}}_k^* \ = \ \boldsymbol{b}_k\mathbf{1}_{N*} + \boldsymbol{\Sigma}_{no}\boldsymbol{\Sigma}_{oo}^{-1}(\boldsymbol{V}_k - \boldsymbol{b}_k\mathbf{1}_N), \tag{4}$$

where the parameters on the right-hand side of the equation are replaced by the corresponding estimates. The estimates $\widehat{\boldsymbol{V}}_k^*$ on the new locations are almost always non-negative as they are effectively represented as a weighted summation of the non-negative cell type proportions on the original locations. To ensure scalable imputation, we used a sparse approximation of the covariance matrix $\boldsymbol{\Sigma}$ by using only the nearest 10 neighbors for each location. With the imputed cell type compositions, we can further impute the gene expression levels on the new locations by multiplying the above conditional mean in equation (4) with the basis matrix to obtain $\boldsymbol{B}\,\widehat{\boldsymbol{V}}_k^*$.

Basis Matrix Construction

We constructed the reference basis matrix $\boldsymbol{B}$ following the main ideas of MuSiC using three detailed steps (details in Supplementary Notes). (1) We selected genes that are expressed in both the scRNA-seq reference data and the spatial transcriptomic data. (2) We selected among them the candidate cell type informative genes with a mean expression level in a given cell type at least 1.25 log fold higher than its mean expression level across all remaining cell types. (3) We removed among them the outlier genes that show high expression heterogeneity within a cell type by calculating gene-specific expression dispersion (Supplementary Figure 92). In particular, we calculated the expression dispersion as the variance to mean ratio for each gene in each cell type. We then obtained the gene-specific dispersion by averaging the estimated expression dispersion across cell types. We finally removed the top 1% genes with the largest gene-specific dispersion values.

Simulations and Deconvolution Analysis Evaluation

All simulations are described in the Supplementary Notes. In each simulation replicate, we calculated the true cell type proportions on each spatial location as the number of cells in each cell type divided by the total number of cells on the location. We denote the true cell type composition matrix as $\boldsymbol{V}$. After we obtained the estimated cell type composition matrix $\widehat{\boldsymbol{V}}$, we evaluated deconvolution performance by computing the root mean square error (RMSE) between $\widehat{\boldsymbol{V}}$ and $\boldsymbol{V}$ through

$$RMSE = \sqrt{\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(\boldsymbol{V}_{ik} - \widehat{\boldsymbol{V}}_{ik}\right)^2},$$

where N=260 is the total number of spatial locations and K is the total number of cell types. Note that the above formula for RMSE calculation is based on all cell types (Supplementary Notes).

Compared Methods

We compared CARD with four deconvolution methods: (1) MuSiC[19] (version 0.1.1), (2) SPOTlight[22] (version 0.1.0), and (3) RCTD[23] (version 1.1.0), (4) cell2location[30] (version 0.07a), (5) spatialDWLS (implemented in the R package Giotto, version 1.0.4), (6) stereoscope (version 0.2.0). For all methods, we followed the tutorial on the corresponding GitHub pages and used the recommended default parameter settings for deconvolution analysis. cell2location requires users to input additional parameters. For these parameters, we set them to be close to what we used in the simulations and to be close to what we best know of in the real data applications. Specifically, in the simulations, we set "cells_per_spot" to be a random number from a uniform distribution U(8, 12) with an expected value of 10. We set "factors_per_spot" and "combs_per_spot" to be exactly the number of cell types available in the corresponding scRNA-seq reference. In the real data applications, we set "cells_per_spot" to be 30 for the mouse olfactory spatial transcriptomics data and human pancreatic ductal adenocarcinoma data and set it to be 10 for the 10x Visium data. We set both the "factors_per_spot" and "combs_per_spot" to be 7 following the software tutorial.

We also compared the high-resolution spatial map constructed by CARD with a recently developed method BayesSpace (version 1.1.4). Because BayesSpace only implements that neighborhood structure suitable for Spatial Transcriptomics (ST) and 10x Visium, we only evaluated its performance on the mouse olfactory spatial transcriptomics data and human pancreatic ductal adenocarcinoma (PDAC) data. We followed the tutorial on GitHub and used the recommended default parameter settings for resolution enhancement. Specifically, we set the required number of clusters qs based on their recommended pseudo-log-likelihood as the following: qs = 5 for mouse olfactory spatial transcriptomics data and qs = 8 for PDAC data. Note that BayesSpace is restricted in creating a neighborhood structure that has a fixed number of sub-spots at each location in the original data (5 for Visium technology and 9 for ST technology). In order to compare the high-resolution spatial gene expression constructed by CARD and BayesSpace on the same set of sub-spots, we applied CARD to directly impute the gene expression on the sub-spots generated by BayesSpace. Afterwards, we performed PCA dimension reduction on the high-resolution data and applied the K-means algorithm analysis on the top 20 PCs to cluster spatial locations into six clusters for the mouse olfactory data and eighteen clusters for the PDAC following the original studies.

Real Data Analyses

All real datasets used in the present study are described in the Supplementary Notes. We first examined cell type composition similarity in these real datasets. Because we do not know the true cell type composition in these data, we used cell type marker genes as surrogates to examine the spatial distribution of cell types on the tissue[93]. We reasoned that, if the cell type composition is similar among neighboring locations, then we would also expect the cell type maker genes to show spatial correlation in their expression pattern on the tissue. Therefore, for each of the three spatial transcriptomics datasets examined in the present study, we looked at one marker at a time (from the same set of markers in real data applications) and examined its spatial autocorrelation pattern by carrying out spatial autocorrelation tests using Moran I[94] and Geary's C[95]. Note that we were unable to carry out Moran's I test[96] and Geary's C test[96] on the large SlideseqV2 dataset due to heavy computational cost. Besides examining cell type marker genes, we also calculated correlation in the expression profile of the marker genes between neighboring locations (Supplementary Notes). Intuitively, if the cell type composition is similar between neighboring locations, then the expression profile of marker genes will also be correlated between neighboring locations, more so than that between locations that are far away.

Next, we applied different methods to deconvolute the above datasets. In each analysis, we supplied the same spatial transcriptomics data and the same scRNA-seq data as input for all methods (preprocessing details in Supplementary Notes). After deconvolution, we followed[97] to assign the dominant cell type on each spatial location and examined the distribution of each cell type on the tissue. For the two datasets that contain a matched H&E image (MOB and PDAC), we compared the distribution of the dominant cell types inferred from spatial transcriptomics with the tissue structures annotated based on the H&E image. Specifically, we obtained tissue structure annotations based on the H&E image, overlayed spatial transcriptomics locations on top the H&E image, and manually annotated each measured location in spatial transcriptomics with the tissue structure annotations extracted from the H&E image. For the MOB dataset, we annotated four main structural layers in the olfactory bulb: the granule cell layer (GCL, which contains n = 67 spatial locations), the mitral cell layer (MCL, n = 75), the glomerular layer (GL, n =80), and the nerve layer (ONL, n = 55). For the PDAC dataset, we annotated four main structural regions on the cancer tissue: cancer region (n = 137), ductal region (n = 72), pancreatic region (n = 70), and stroma region (n = 147). In the MOB dataset, because each olfactory layer is dominated by one cell type, we directly compared the dominant cell type inferred from CARD with the layer annotations based on H&E image via adjusted rand index (ARI) and Purity, using the *compare* function in the *igraph* R package (v1.0.0) and *purity* function in the *funtimes* R packages (v8.1), respectively (details in Supplementary Notes). In the PDAC dataset, because each tissue region is substantially more heterogenous than that in the MOB data and contains potentially multiple cell types, using ARI would penalize methods that detected fine tissue regions that were not detected in the original study. Therefore, we carefully examined the distribution of inferred cell types on each annotated tissue region based on the transcriptomic profile and existing biological literature.

Because CARD directly models spatial correlation, CARD can be used to impute gene expression on unmeasured locations. To evaluate the accuracy of such imputation, we performed location masking analysis. Specifically, in each real data application, we randomly masked a fixed percentage of the spatial locations to be missing, used the unmasked spatial locations to perform CARD deconvolution, relied on the cell type composition estimates obtained on the unmasked locations to predict and impute the cell type composition on the masked locations, and further imputed the gene expression levels on the masked locations. We then compared the imputed gene expression level with the measured expression level on the masked locations using RMSE. RMSE serves as an indicator on how accurate CARD imputation works, which also reflects its deconvolution performance. The magnitude of RMSE can vary substantially across datasets depending on factors such as the sequencing read depth per location. In the analysis, we set the mask percentage to be either 1%, 2%, 5%, 10% or 20% for all datasets.

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article

Data Availability

This study made use of publicly available datasets. These include the mouse olfactory bulb dataset (http://www.spatialtranscriptomicsresearch.org), human pancreatic ductal adenocarcinoma (PDAC) dataset (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111672),

Mouse hippocampus Slide-seqV2 dataset ((https://singlecell.broadinstitute.org/single_cell/study/SCP948/robust-decomposition-of-cell-type-mixtures-in-spatial-transcriptomics), and mouse brain (coronal section) 10x Visium (https://www.10xgenomics.com/resources/datasets/). For the scRNAseq references used in this study, they are all publicly available with details provided in supplementary tables 2-3.

Code Availability

The CARD software package and source code have been deposited at [www.xzlab.org/software.html]. All scripts used to reproduce all the analysis are also available at the same website.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This study was supported by the National Institutes of Health (NIH) Grants R01GM126553, R01HG011883, and R01GM144960, all to XZ.

References for main text

1. Burgess DJ Spatial transcriptomics coming of age. Nature Reviews Genetics (2019) doi:10.1038/s41576-019-0129-z.
2. Soldatov R et al. Spatiotemporal structure of cell fate decisions in murine neural crest. Science (80-.). (2019) doi:10.1126/science.aas9536.
3. Prinz M, Priller J, Sisodia SS & Ransohoff RM Heterogeneity of CNS myeloid cells and their roles in neurodegeneration. Nature Neuroscience (2011) doi:10.1038/nn.2923.
4. Svensson V, Teichmann SA & Stegle O SpatialDE: Identification of spatially variable genes. Nat. Methods (2018) doi:10.1038/nmeth.4636.
5. Dries R et al. Giotto, a pipeline for integrative analysis and visualization of single-cell spatial transcriptomic data. bioRxiv (2019) doi:10.1101/701680.
6. Pham D et al. stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. bioRxiv (2020).
7. Biancalani T et al. Deep learning and alignment of spatially-resolved whole transcriptomes of single cells in the mouse brain with Tangram. bioRxiv (2020).
8. Chen J et al. Unsupervised Spatially Embedded Deep Representation of Spatial Transcriptomics. (2021).
9. Fischl AM, Heron PM, Stromberg AJ & McClintock TS Activity-dependent genes in mouse olfactory sensory neurons. Chem. Senses (2014) doi:10.1093/chemse/bju015.
10. Moses L & Pachter L Museum of Spatial Transcriptomics. (2021).
11. Asp M, Bergenstråhle J & Lundeberg J Spatially Resolved Transcriptomes—Next Generation Tools for Tissue Exploration. BioEssays (2020) doi:10.1002/bies.201900221.
12. Genomics, 10x. 10x Genomics Visium. https://www.10xgenomics.com/spatial-transcriptomics/.
13. Rodriques SG et al. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science (80-.). (2019) doi:10.1126/science.aaw1219.
14. Ståhl PL et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science (2016) doi:10.1126/science.aaf2403.
15. Liao J, Lu X, Shao X, Zhu L & Fan X Uncovering an Organ's Molecular Architecture at Single-Cell Resolution by Spatially Resolved Transcriptomics. Trends in Biotechnology (2020) doi:10.1016/j.tibtech.2020.05.006.
16. Rao A, Barkley D, França GS & Yanai I Exploring tissue architecture using spatial transcriptomics. Nature 596, 211–220 (2021). [PubMed: 34381231]
17. Hwang B, Lee JH & Bang D Single-cell RNA sequencing technologies and bioinformatics pipelines. Experimental and Molecular Medicine (2018) doi:10.1038/s12276-018-0071-8.
18. Cobos FA, Alquicira-Hernandez J, Powell JE, Mestdagh P & De Preter K Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nat. Commun. 11, 1–14 (2020). [PubMed: 31911652]
19. Wang X, Park J, Susztak K, Zhang NR & Li M Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat. Commun. (2019) doi:10.1038/s41467-018-08023-x.
20. Dong M et al. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. Brief. Bioinform. (2020) doi:10.1093/bib/bbz166.
21. Jew B et al. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Nat. Commun. (2020) doi:10.1038/s41467-020-15816-6.

22. Elosua-Bayes M, Nieto P, Mereu E, Gut I & Heyn H SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. Nucleic Acids Res. 49, e50–e50 (2021). [PubMed: 33544846]

23. Cable DM et al. Robust decomposition of cell type mixtures in spatial transcriptomics. Nat. Biotechnol. 1–10 (2021). [PubMed: 33376248]

24. Song Q & Su J DSTG: deconvoluting spatial transcriptomics data through graph-based artificial intelligence. Brief. Bioinform. (2021).

25. Lopez R et al. Multi-resolution deconvolution of spatial transcriptomics data reveals continuous patterns of inflammation. bioRxiv (2021).

26. Biancalani T et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. Nat. Methods 1–11 (2021). [PubMed: 33408396]

27. Danaher P et al. Advances in mixed cell deconvolution enable quantification of cell types in spatially-resolved gene expression data. bioRxiv (2020).

28. Gayoso A et al. scvi-tools: a library for deep probabilistic analysis of single-cell omics data. bioRxiv (2021).

29. Andersson A et al. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. Commun. Biol. 3, 1–8 (2020). [PubMed: 31925316]

30. Kleshchevnikov V et al. Cell2location maps fine-grained cell types in spatial transcriptomics. Nat. Biotechnol. 1–11 (2022) doi:10.1038/s41587-021-01139-4. [PubMed: 34980916]

31. Dong R & Yuan G-C SpatialDWLS: accurate deconvolution of spatial transcriptomic data. Genome Biol. 22, 1–10 (2021). [PubMed: 33397451]

32. Stoltzfus CR et al. CytoMAP: a spatial analysis toolbox reveals features of myeloid cell organization in lymphoid tissues. Cell Rep. 31, 107523 (2020).

33. Dudas M, Wysocki A, Gelpi B & Tuan T-L Memory encoded throughout our bodies: molecular and cellular basis of tissue regeneration. Pediatr. Res. 63, 502–512 (2008). [PubMed: 18427295]

34. Bove A et al. Local cellular neighborhood controls proliferation in cell competition. Mol. Biol. Cell 28, 3215–3228 (2017). [PubMed: 28931601]

35. Van Vliet S et al. Spatially correlated gene expression in bacterial groups: the role of lineage history, spatial gradients, and cell-cell interactions. Cell Syst. 6, 496–507 (2018). [PubMed: 29655705]

36. Phillips D et al. Immune cell topography predicts response to PD-1 blockade in cutaneous T cell lymphoma. medRxiv (2020).

37. Schürch CM et al. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. Cell 182, 1341–1359 (2020). [PubMed: 32763154]

38. Allen Reference Atlas – Mouse Brain [brain atlas]. Available from atlas.brain-map.org.

39. Spatial Research, available from https://www.spatialresearch.org.

40. Public Health Image Library, available from https://phil.cdc.gov.

41. Xia C, Fan J, Emanuel G, Hao J & Zhuang X Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. Proc. Natl. Acad. Sci. 116, 19490–19499 (2019).

42. Eng CHL et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. Nature (2019) doi:10.1038/s41586-019-1049-y.

43. Banerjee S, Carlin BP & Gelfand AE Hierarchical Modeling and Analysis for Spatial Data. Hierarchical Modeling and Analysis for Spatial Data (2014). doi:10.1201/b17115.

44. Lee D A comparison of conditional autoregressive models used in Bayesian disease mapping. Spat. Spatiotemporal. Epidemiol. (2011) doi:10.1016/j.sste.2011.03.001.

45. Zhao E et al. Spatial transcriptomics at subspot resolution with BayesSpace. Nat. Biotechnol. 1–10 (2021). [PubMed: 33376248]

46. Zeisel A et al. Molecular Architecture of the Mouse Nervous System. Cell (2018) doi:10.1016/j.cell.2018.06.021.

47. Yang S & Zhou X Accurate and Scalable Construction of Polygenic Scores in Large Biobank Data Sets. Am. J. Hum. Genet. (2020) doi:10.1016/j.ajhg.2020.03.013.

48. Zhou X, Carbonetto P & Stephens M Polygenic Modeling with Bayesian Sparse Linear Mixed Models. PLoS Genet. (2013) doi:10.1371/other.pgen.1003264.

49. 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) doi:10.1016/j.celrep.2018.11.034.

50. Nagayama S, Homma R & Imamura F Neuronal organization of olfactory bulb circuits. Frontiers in Neural Circuits (2014) doi:10.3389/fncir.2014.00098.

51. Moncada R et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. Nat. Biotechnol. (2020) doi:10.1038/s41587-019-0392-8.

52. Zheng B et al. TM4SF1 as a prognostic marker of pancreatic ductal adenocarcinoma is involved in migration and invasion of cancer cells. Int. J. Oncol. (2015) doi:10.3892/ijo.2015.3022.

53. Fu F et al. Role of transmembrane 4 L six family 1 in the development and progression of cancer. Front. Mol. Biosci. 7, (2020).

54. Xu D et al. Lost miR-141 and upregulated TM4SF1 expressions associate with poor prognosis of pancreatic cancer: regulation of EMT and angiogenesis by miR-141 and TM4SF1 via AKT. Cancer Biol. Ther. 21, 354–363 (2020). [PubMed: 31906774]

55. Zhang X et al. Expression pattern of cancer-associated fibroblast and its clinical relevance in intrahepatic cholangiocarcinoma. Hum. Pathol. 65, 92–100 (2017). [PubMed: 28457731]

56. Morvaridi S, Dhall D, Greene MI, Pandol SJ & Wang Q Role of YAP and TAZ in pancreatic ductal adenocarcinoma and in stellate cells associated with cancer and chronic pancreatitis. Sci. Rep. (2015) doi:10.1038/srep16759.

57. Nielsen MFB, Mortensen MB & Detlefsen S Key players in pancreatic cancer-stroma interaction: Cancer-associated fibroblasts, endothelial and inflammatory cells. World J. Gastroenterol. 22, 2678 (2016). [PubMed: 26973408]

58. Zheng C et al. Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. Cell (2017) doi:10.1016/j.cell.2017.05.035.

59. Comito G, Ippolito L, Chiarugi P & Cirri P Nutritional Exchanges Within Tumor Microenvironment: Impact for Cancer Aggressiveness. Frontiers in Oncology (2020) doi:10.3389/fonc.2020.00396.

60. Lambrechts D et al. Phenotype molding of stromal cells in the lung tumor microenvironment. Nat. Med. (2018) doi:10.1038/s41591-018-0096-5.

61. Peng J et al. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Res. (2019) doi:10.1038/s41422-019-0195-y.

62. Stickels RR et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. Nat. Biotechnol. 39, 313–319 (2021). [PubMed: 33288904]

63. Saunders A et al. Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. Cell (2018) doi:10.1016/j.cell.2018.07.028.

64. Del Bigio MR Ependymal cells: biology and pathology. Acta Neuropathol. 119, 55–73 (2010). [PubMed: 20024659]

65. Encyclopedia of the human brain. Choice Rev. Online (2003) doi:10.5860/choice.40-2552.

66. Meyer G Building a human cortex: The evolutionary differentiation of Cajal-Retzius cells and the cortical hem. J. Anat. (2010) doi:10.1111/j.1469-7580.2010.01266.x.

67. Stickels RR et al. Sensitive spatial genome wide expression profiling at cellular resolution. bioRxiv (2020).

68. Hawrylycz M et al. The allen brain atlas. in Springer Handbook of Bio-/Neuroinformatics (2014). doi:10.1007/978-3-642-30574-0_62.

69. Wozny C et al. VGLUT2 functions as a differential marker for hippocampal output neurons. Front. Cell. Neurosci. (2018) doi:10.3389/fncel.2018.00337.

70. Wälchli T et al. Quantitative assessment of angiogenesis, perfused blood vessels and endothelial tip cells in the postnatal mouse brain. Nat. Protoc. 10, 53–74 (2015). [PubMed: 25502884]

71. Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M & Cai L Single-cell in situ RNA profiling by sequential hybridization. Nature Methods (2014) doi:10.1038/nmeth.2892.

72. Chen KH, Boettiger AN, Moffitt JR, Wang S & Zhuang X Spatially resolved, highly multiplexed RNA profiling in single cells. Science (80-.). (2015) doi:10.1126/science.aaa6090.

73. Moffitt JR et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. Science (80-.). 362, (2018).

74. He Y et al. ClusterMap: multi-scale clustering analysis of spatial gene expression. bioRxiv (2021).

75. Moen E et al. Deep learning for cellular image analysis. Nat. Methods 16, 1233–1246 (2019). [PubMed: 31133758]

76. Bergenstråhle J, Larsson L & Lundeberg J Seamless integration of image and molecular analysis for spatial transcriptomics workflows. BMC Genomics 21, 1–7 (2020).

77. Sun S et al. Differential expression analysis for RNAseq using Poisson mixed models. Nucleic Acids Res. 45, e106–e106 (2017). [PubMed: 28369632]

78. Sun S et al. Heritability estimation and differential analysis of count data with generalized linear mixed models in genomic sequencing studies. Bioinformatics 35, 487–496 (2019). [PubMed: 30020412]

79. Sun S, Zhu J & Zhou X Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nat. Methods (2020) doi:10.1038/s41592-019-0701-7.

80. Hu J et al. Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. bioRxiv (2020).

81. Soneson C & Robinson MD Bias, robustness and scalability in single-cell differential expression analysis. Nat. Methods (2018) doi:10.1038/nmeth.4612.

82. Duò A, Robinson MD & Soneson C A systematic performance evaluation of clustering methods for single-cell RNA-seq data. F1000Research (2018) doi:10.12688/f1000research.15666.2.

83. De Oliveira V Bayesian analysis of conditional autoregressive models. Ann. Inst. Stat. Math. (2012) doi:10.1007/s10463-010-0298-1.

84. Besag J Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc. Ser. B 36, 192–225 (1974).

85. Vanhatalo J, Pietiläinen V & Vehtari A Approximate inference for disease mapping with sparse Gaussian processes. Stat. Med. (2010) doi:10.1002/sim.3895.

86. Rousset F & Ferdy JB Testing environmental and genetic effects in the presence of spatial autocorrelation. Ecography (Cop.). (2014) doi:10.1111/ecog.00566.

87. Cressie N STATISTICS FOR SPATIAL DATA. Terra Nov. (1992) doi:10.1111/j.1365-3121.1992.tb00605.x.

88. Rue H & Held L Gaussian markov random fields: Theory and applications. Gaussian Markov Random Fields: Theory and Applications (2005). doi:10.1198/tech.2006.s352.

89. Brook D On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. Biometrika 51, 481–483 (1964).

90. Lee DD & Seung HS Algorithms for non-negative matrix factorization. in Advances in Neural Information Processing Systems (2001).

91. Janecek A & Tan Y Iterative improvement of the multiplicative update nmf algorithm using nature-inspired optimization. in 2011 Seventh International Conference on Natural Computation vol. 3 1668–1672 (IEEE, 2011).

92. Park J-S & Oh S-J A new concave hull algorithm and concaveness measure for n-dimensional datasets. J. Inf. Sci. Eng. 28, 587–600 (2012).

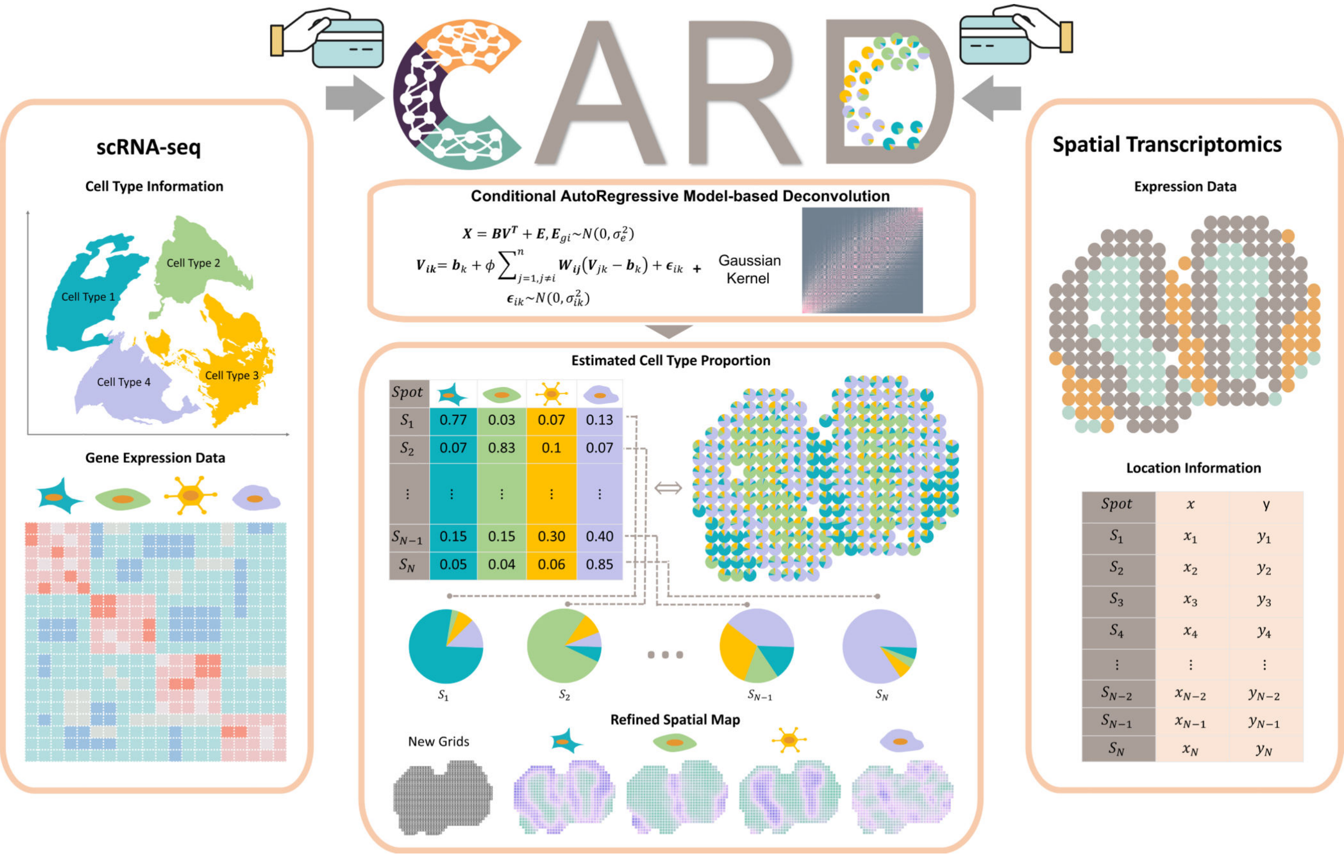
93. Ralston A & Shaw K Gene expression regulates cell differentiation. Nat Educ 1, 127–131 (2008).

94. Li H, Calder CA & Cressie N Beyond Moran's I: testing for spatial dependence based on the spatial autoregressive model. Geogr. Anal. 39, 357–375 (2007).

95. Radeloff VC, Miller TF, He HS & Mladenoff DJ Periodicity in spatial data and geostatistical models: autocorrelation between patches. Ecography (Cop.). 23, 81–91 (2000).

96. Bivand R et al. spdep: Spatial dependence: weighting schemes, statistics and models. (2011).

97. Teschendorff AE, Zhu T, Breeze CE & Beck S EPISCORE: Cell type deconvolution of bulk tissue DNA methylomes from single-cell RNA-Seq data. Genome Biol. (2020) doi:10.1186/s13059-020-02126-9.



$$X = BV^T + E, E_{gi} \sim N(0, \sigma_e^2)$$

$$V_{ik} = b_k + \phi \sum_{j=1, j \neq i}^{n} W_{ij}(V_{jk} - b_k) + \epsilon_{ik} + \text{Gaussian Kernel}$$

$$\epsilon_{ik} \sim N(0, \sigma_{ik}^2)$$

Spot				
S_1	0.77	0.03	0.07	0.13
S_2	0.07	0.83	0.1	0.07
⋮	⋮	⋮	⋮	⋮
S_{N-1}	0.15	0.15	0.30	0.40
S_N	0.05	0.04	0.06	0.85

Spot	x	y
S_1	x_1	y_1
S_2	x_2	y_2
S_3	x_3	y_3
S_4	x_4	y_4
⋮	⋮	⋮
S_{N-2}	x_{N-2}	y_{N-2}
S_{N-1}	x_{N-1}	y_{N-1}
S_N	x_N	y_N

Figure 1. Schematic overview of CARD.

CARD is designed to deconvolute spatial transcriptomics data and infer cell type composition on each spatial location based on the reference scRNA-seq data. CARD requires a scRNA-seq data with cell type specific gene expression information (left box) along with the spatial transcriptomics data with localization information (right box). With these two inputs, CARD performs deconvolution through a non-negative matrix factorization framework and outputs the estimated cell type composition across spatial locations (bottom box). A unique feature of CARD is its ability to account for the spatial correlation of cell type compositions across spatial locations through a conditional autoregressive (CAR) model (top box). By accounting for the spatial correlation of cell type compositions across spatial locations, CARD is also capable of imputing cell type compositions and gene expression levels on locations not measured in the original study, facilitating the construction of a refined high-resolution spatial map on the tissue (bottom box).

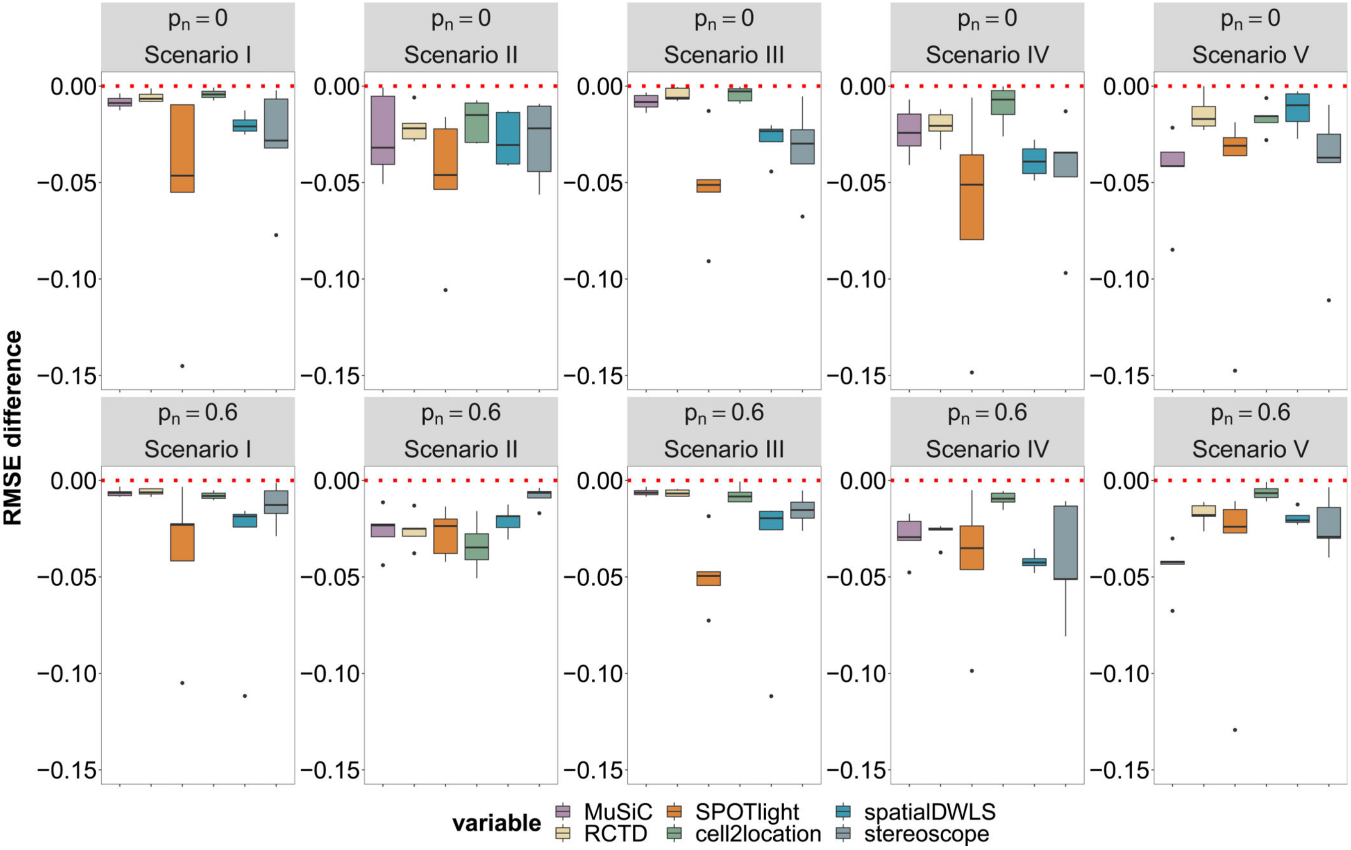


Figure 2. Comparison of deconvolution accuracy of different methods in simulations under the analysis scenarios I-V.

In the analysis scenario I, the same scRNA-seq dataset used in simulations is used as the reference for deconvolution. In the analysis scenario II, the same scRNA-seq data but with one missing cell type (e.g., Neuron cells) is used as the reference for deconvolution. In the analysis scenario III, the same scRNA-seq data but with one additional cell type (e.g., Blood cells) is used as the reference for deconvolution. In the analysis scenario IV, the same scRNA-seq reference data but with miss-classified cell type in the reference for deconvolution. In the analysis scenario V, the different scRNA-seq reference sequenced from a different platform but with similar cell types is used as the reference for deconvolution. Compared deconvolution methods (x-axis) include MuSiC (purple), RCTD (yellow), SPOTlight (orange), cell2location (green), spatialDWLS (blue), and stereoscope (blue-gray). Simulations were performed under different spatial correlation strength as represented by the proportion of noisy locations (p_n). High p_n corresponds to low spatial correlation. We calculated the root mean square errors (RMSE) between the estimated cell type compositions and the true cell type compositions for each method to measure its deconvolution performance. We further contrasted RMSE of the other methods with respect to that of CARD by computing an RMSE difference to remove the unnecessarily difficulty level variation across replicates. An RMSE difference (y-axis) below zero suggests that CARD performs better than other methods. Differences of RMSE across five simulation replicates (n = 5) were displayed in the form of box plots. Each boxplot ranges from the third and first quartiles with the median as the horizontal line while whiskers represent 1.5 times the interquartile range from the lower and upper bounds of the box.

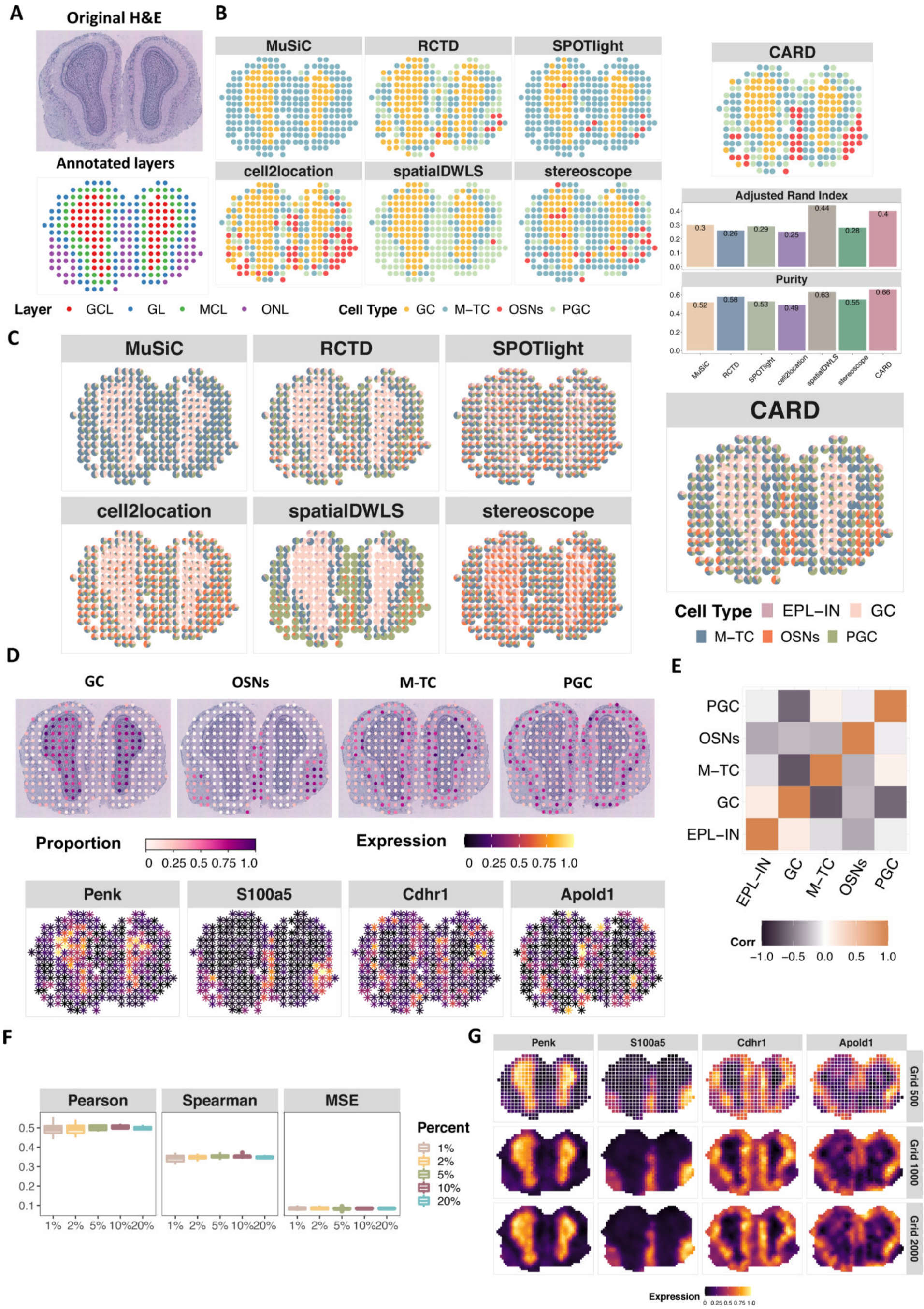


Figure 3. Analyzing the mouse olfactory bulb data.

(A) Hematoxylin and eosin (H&E) staining of the olfactory bulb (top panel) displays four anatomic layers that are organized in an inside out fashion (bottom panel): the granule cell layer (GCL), the mitral cell layer (MCL), the glomerular layer (GL), and the nerve layer (ONL). **(B)** Left panel shows on each spatial location the dominant cell type inferred from four different deconvolution methods. The examined cell types include granule cells (GC), olfactory sensory neurons (OSNs), periglomerular cells (PGC) and mitral/tufted cells (M-TC). Compared deconvolution methods include MuSiC, RCTD, SPOTlight, cell2location, spatialDWLS, stereoscope and CARD. Right bottom panel displays the adjusted rand index (ARI; y-axis) and the Purity (y-axis), which quantify the similarity between the inferred dominant cell types from different methods (x-axis) and the anatomic layers annotated based on the H&E image. **(C)** Spatial scatter pie plot displays inferred cell type composition on each spatial location from different deconvolution methods. **(D)** Top panels display on each spatial location the proportion of each of the four cell types inferred by CARD. Bottom panels display the expression levels of four corresponding cell type specific marker genes. **(E)** Correlations in cell type proportion across spatial locations between pairs of cell types inferred by CARD. Color is scaled by the correlation value. **(F)** Accuracy of CARD imputation in the masking analysis across 10 replicates (n = 10). A fixed percentage of locations are masked as missing (x-axis) and CARD is used to impute the gene expression on the masked locations. Three different metrics (y-axis) are used to evaluate imputation accuracy in terms of the similarity between the imputed expression and true expression on masked locations: Pearson's correlation, Spearman's correlation and mean square error (MSE). Each boxplot ranges from the first and third quartiles with the median as the horizontal line while whiskers represent 1.5 times the interquartile range from the lower and upper bounds of the box. **(G)** CARD imputes gene expression for four marker genes on a fine grid set of spatial locations (number of grid points = 500, 1,000, or 2,000), resulting in a refined spatial map of gene expression.

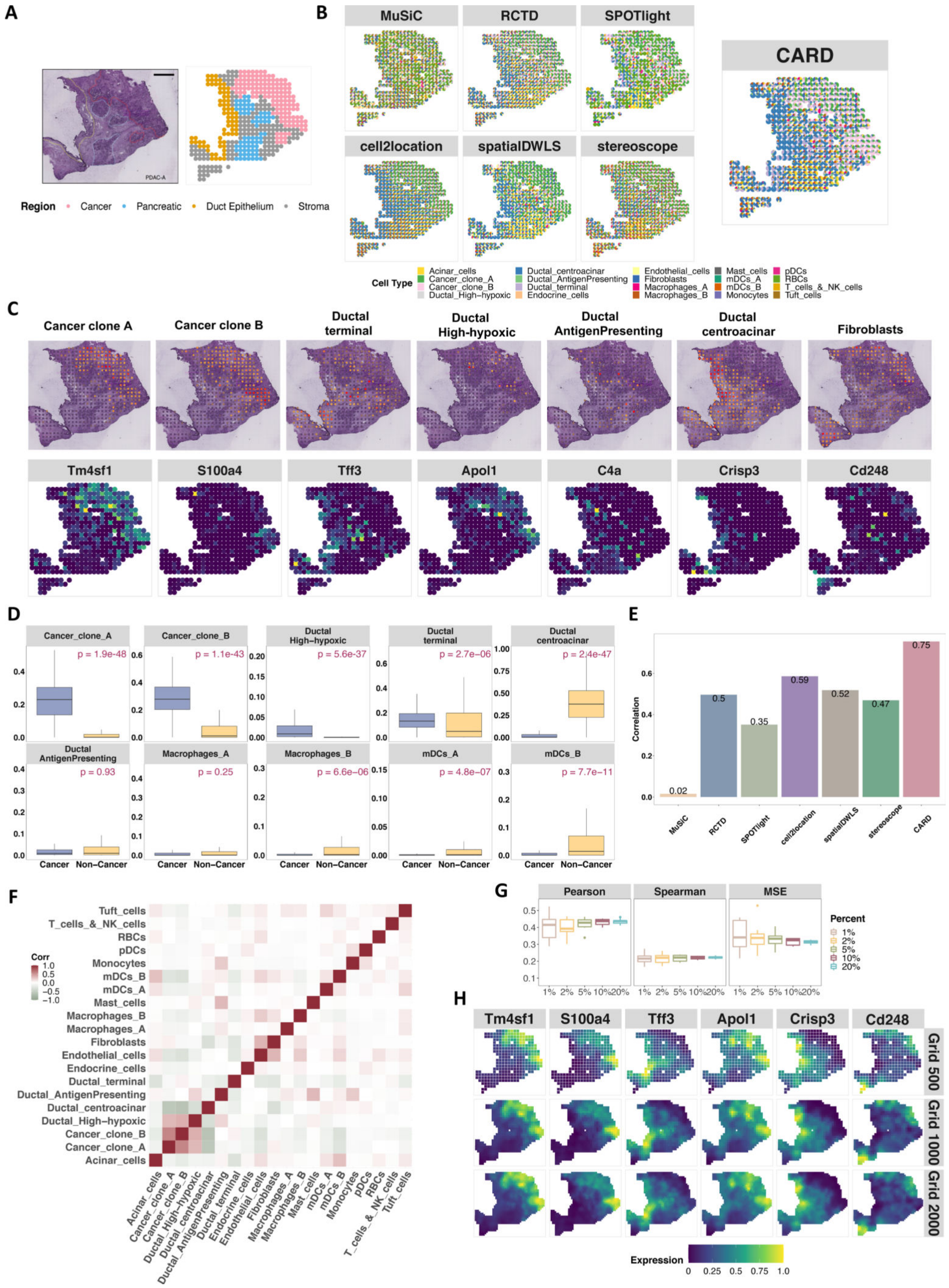


Figure 4. Analyzing the pancreatic ductal adenocarcinoma (PDAC) data.

(A) Hematoxylin and eosin (H&E) staining of the PDAC (left panel) displays four regions (right panel) annotated from the original publication[51]: Cancer, Pancreatic, Duct and stroma regions. **(B)** Spatial scatter pie plot displays inferred cell type composition on each spatial location from different deconvolution methods. Compared deconvolution methods include MuSiC, RCTD, SPOTlight, cell2location, spatialDWLS, stereoscope and CARD. **(C)** Top panels display on each spatial location the proportion of each of the cell types inferred by CARD. Bottom panels display the expression levels of corresponding cell type specific marker genes. **(D)** Comparisons of cell type proportions inferred by CARD in cancer region (n = 137) vs non-cancer region (n = 289) with p-value tested by two-sided Wilcoxon Rank Sum test. **(E)** Correlation between mean cell type proportions inferred by CARD and that in the matched scRNA-seq reference data. **(F)** Correlations in cell type proportion across spatial locations between pairs of cell types inferred by CARD. Color is scaled by the correlation value. **(G)** Accuracy of CARD imputation in the masking analysis across 10 replicates (n = 10). A fixed percentage of locations are masked as missing (x-axis), and CARD is used to impute the gene expression on the masked locations. Three different metrics (y-axis) are used to evaluate imputation accuracy in terms of the similarity between the imputed expression and true expression on masked locations: Pearson's correlation, Spearman's correlation and mean square error (MSE). **(H)** CARD imputes gene expression for four marker genes on a fine grid set of spatial locations (number of grid points = 500, 1,000, or 2,000), resulting in a refined spatial map of gene expression. Each boxplot in (D) and (G) ranges from the first and third quartiles with the median as the horizontal line while whiskers represent 1.5 times the interquartile range from the lower and upper bounds of the box.

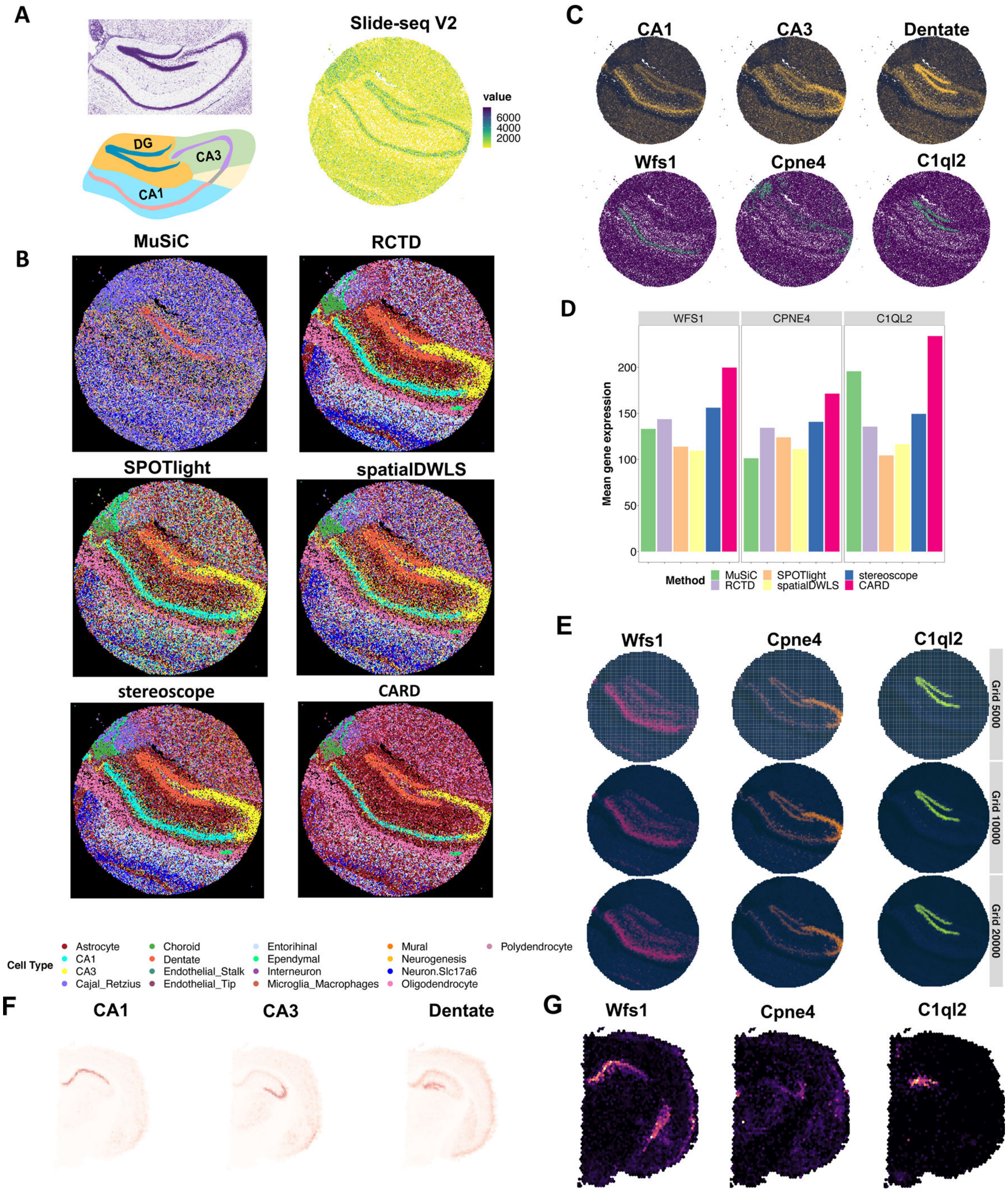


Figure 5. Analyzing the hippocampus region in the Slide-seq V2 and 10x Visium Mouse Brain (Coronal) data.

(A) The UMI counts of Slide-seq V2 data (right panel) displays the structure and the shape of hippocampus tissue, highly consistent with the image from Allen Reference Atlas[38] (left panel) **(B)** The dominant cell type on each location inferred from four different deconvolution methods. Compared deconvolution methods include MuSiC, RCTD, SPOTlight, spatialDWLS, stereoscope and CARD. **(C)** Top panels display on each spatial location the proportion of each of the cell types inferred by CARD. Bottom panels display the expression levels of corresponding cell type specific marker genes. The examined cell types are CA1 cells, CA3 cells and dentate cells. **(D)** Bar plots display the comparisons of the mean gene expression level of marker genes in the major regions inferred by different deconvolution methods; **(E)** CARD imputes gene expression for four marker genes on a fine grid set of spatial locations, resulting in a refined spatial map of gene expression. **(F)** The proportion of each of the cell types on each location inferred by CARD in the 10x Visium dataset. **(G)** The expression levels of corresponding cell type specific marker genes in the 10x Visium dataset.