nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-022-01273-7

Spatially informed cell-type deconvolution for spatial transcriptomics

Ying Ma[1] and Xiang Zhou[1,2] ✉

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-based 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 to enable the construction of a refined spatial tissue map with a resolution arbitrarily higher than that measured in the original study and can perform deconvolution without an 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.

Spatially resolved transcriptomic technologies perform gene expression profiling on many tissue locations with spatial localization information[1], enabling the characterization of transcriptomic landscapes 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 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 is currently 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[22–30], such as RCTD[23], stereoscope[28], SPOTlight[22], cell2location[29] and spatialDWLS[30] (details in Supplementary Notes). All of 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[31–34], which are characterized by a domain-specific composition of cell types, with similar cell types colocalized spatially[35,36]. Histological characterization of various tissues (https://atlas.brain-map.org/, https://www.spatialresearch.org and https://phil.cdc.gov/), including hematoxylin and eosin (H&E) staining images accompanying spatial transcriptomics datasets[12,14], highlights the spatial segregation of cell types and neighboring cell-type composition similarity. In single-cell resolution spatial transcriptomics[37,38], we also observed that similar cell types tend to colocalize, with colocalization patterns decaying with distance (Supplementary Figs. 1 and 2). Consequently, neighboring locations on the tissue likely contain more similar cell-type compositions than 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 transcriptomics. 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 data. 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 (CAR) modeling assumption[39,40]. As a result, CARD can take advantage of the spatial correlation structure to enable accurate and robust deconvolution of spatial transcriptomics data 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 of these features are in direct contrast to a recent method BayesSpace[41] 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 an scRNA-seq reference. We develop a computationally efficient

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA. [2]Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA. ✉e-mail: xzhousph@umich.edu

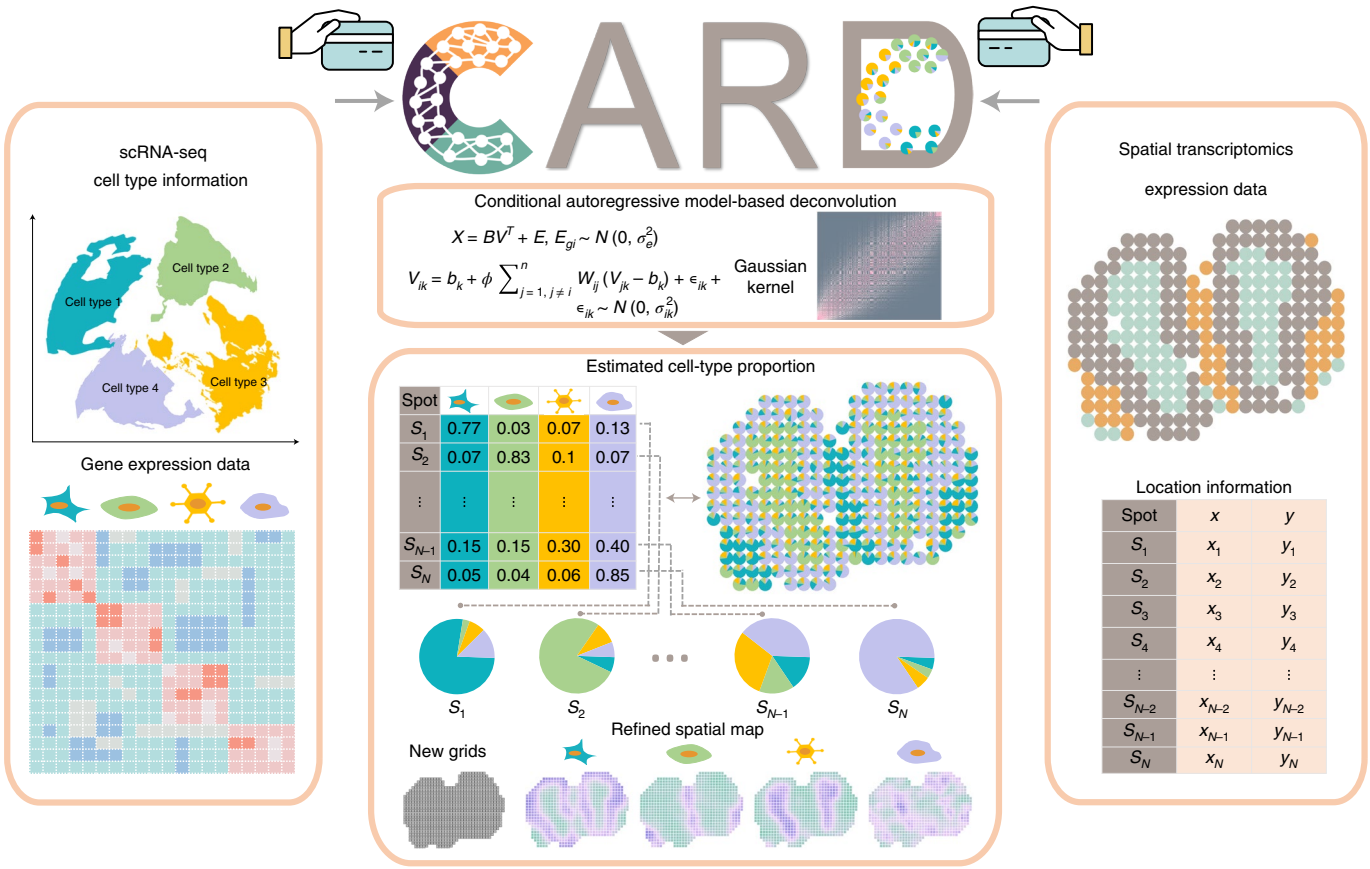


Fig. 1 | Schematic overview of CARD. CARD is designed to deconvolute spatial transcriptomics data and infer cell-type composition on each spatial location based on the reference scRNA-seq data. CARD requires scRNA-seq data with cell-type-specific gene expression information (left) along with spatial transcriptomics data with localization information (right). 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). A unique feature of CARD is its ability to account for the spatial correlation of cell-type compositions across spatial locations through a CAR model (top). 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).

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 the Methods, with its technical details provided in the Supplementary Notes and its method schematic shown in Fig. 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 (Methods). Briefly, we used scRNA-seq data[42] 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 Figs. 3 and 4). The simulated data are realistic, preserving data features observed in the published spatial transcriptomics data (Supplementary Fig. 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 a reference. We then followed ref. [19] and quantified the deconvolution performance by computing the root mean square error (r.m.s.e.) between the estimated cell-type composition and the underlying truth on each location. We primarily displayed r.m.s.e. difference plots where we contrasted the r.m.s.e. of other methods with respect to CARD following refs. [43,44]. We kept the original r.m.s.e. and rank plots in the supplements, which show consistent results.

We first explored a baseline analysis scenario (scenario 1), 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 r.m.s.e. = 0.079), with 9%, 8%, 33%, 7%, 23% and 18% improvement in terms of r.m.s.e. compared to MuSiC (0.087), RCTD (0.086), SPOTlight (0.118), cell2location (0.085), spatialDWLS (0.103) and stereoscope (0.096), respectively (Fig. 2, scenario 1, and Supplementary Figs. 6 and 8). In addition, CARD identifies the dominant cell type on each spatial location accurately, as measured by area under the curve (AUC) and adjusted rand index (ARI; Supplementary Fig. 9).

To examine the robustness of different deconvolution methods, we explored four additional scenarios (Supplementary Notes) where we removed one cell type in the scRNA-seq reference (scenario 2), added one cell type (scenario 3), used misclassified cell types (scenario 4) or used other scRNA-seq data sequenced on a different platform for deconvolution (scenario 5). Compared to scenario 1, the performances of all methods remain similar in scenario 3 (except SPOTlight) and generally are reduced in other scenarios,

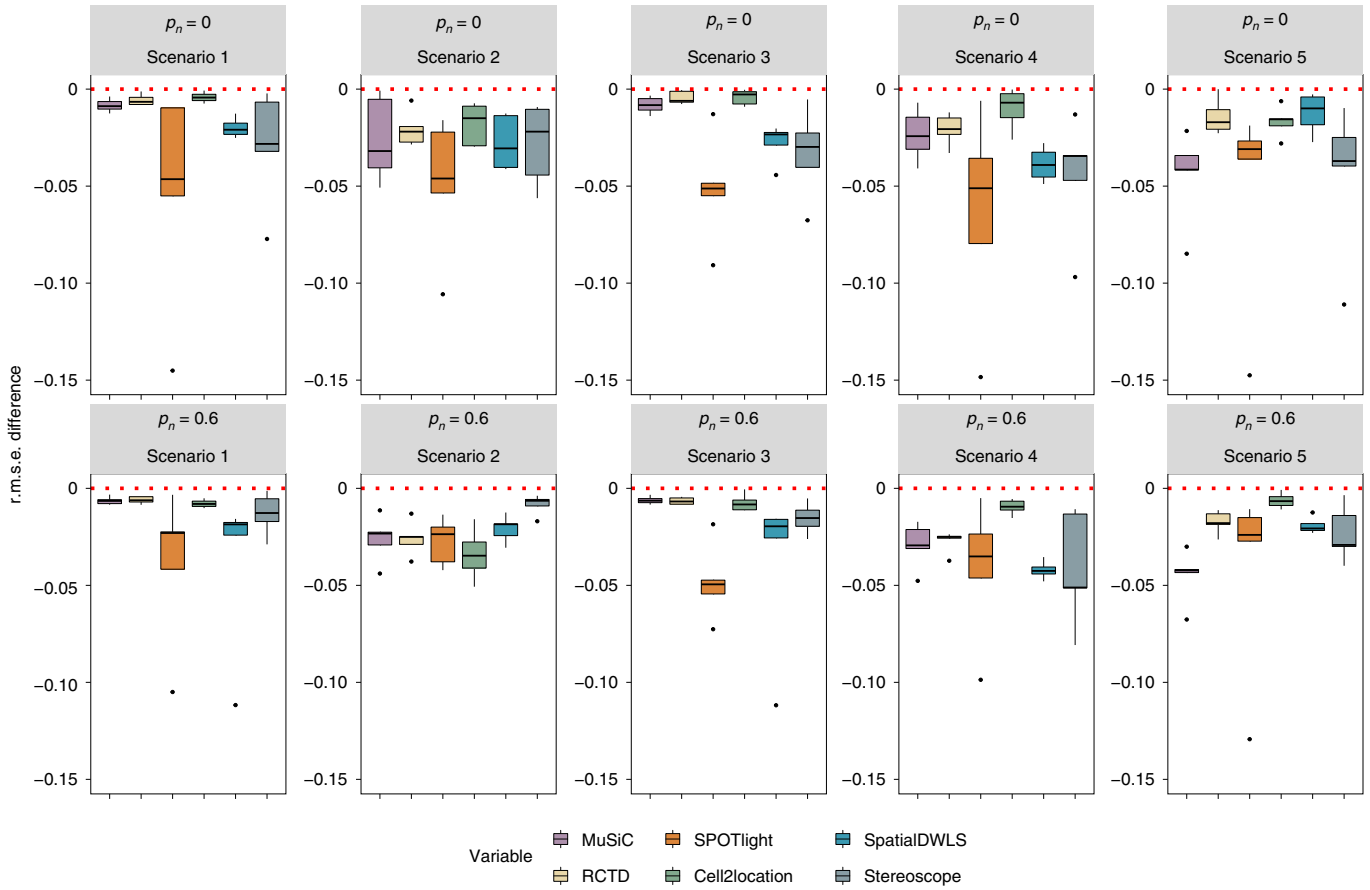


Fig. 2 | Comparison of deconvolution accuracy of different methods in simulations under analysis scenarios 1–5. In analysis scenario 1, the same scRNA-seq dataset used in simulations is used as the reference for deconvolution. In analysis scenario 2, the same scRNA-seq dataset but with one missing cell type (for example, neurons) is used as the reference for deconvolution. In analysis scenario 3, the same scRNA-seq dataset but with one additional cell type (for example, blood cells) is used as the reference for deconvolution. In analysis scenario 4, the same scRNA-seq reference dataset but with a misclassified cell type in the reference is used for deconvolution. In analysis scenario 5, a 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 r.m.s.e. between the estimated cell-type compositions and the true cell-type compositions for each method to measure its deconvolution performance. We further contrasted r.m.s.e. of the other methods with respect to that of CARD by computing an r.m.s.e. difference to remove the unnecessary difficulty level variation across replicates. An r.m.s.e. difference (*y* axis) below zero suggests that CARD performs better than other methods. Differences of r.m.s.e. across five simulation replicates ($n = 5$) were displayed in the form of box plots. Each box plot 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.

although their relative ranks remain largely consistent across scenarios. In addition, CARD outperforms the other methods in all settings, with its performance gains more apparent than scenario 1 (Fig. 2). Specifically, in scenario 2, CARD loses a median of 3% accuracy across settings compared to using the original scRNA-seq data (Supplementary Notes). However, CARD is more accurate than the other methods across settings, with a 13–32% accuracy improvement (Fig. 2 and Supplementary Figs. 10–14). In scenario 3, CARD only loses a median of 0.4% accuracy across settings 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 (Fig. 2 and Supplementary Fig. 15). In scenario 4, CARD loses a median of 4% accuracy compared to using the original scRNA-seq data (Fig. 2 and Supplementary Fig. 16). However, CARD is again more accurate than the other methods across settings (Fig. 2 and Supplementary Fig. 17), with 6–32% accuracy improvement across misclassified cell types (Supplementary Fig. 18). In scenario 5, CARD loses a median of 10% accuracy across settings 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 (Fig. 2 and Supplementary Fig. 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 Fig. 20) and reached a saturation point with a sufficiently large number of sub-cell types, where many sub-cell types are no longer distinguishable from each other (Supplementary Fig. 21). Regardless of the cell-type resolution, the relative performances of most deconvolution methods remain consistent (Supplementary Fig. 22). We also performed 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 Fig. 23).

Mouse olfactory bulb (MOB) data. We applied CARD and the other methods to analyze four published spatial transcriptomics datasets that include two obtained from ST, one from Slide-seqV2 and one from 10x Visium (details in Supplementary Notes). In each dataset, 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 of <0.05; Supplementary Table 1), with the semivariance generally increasing with distance (Supplementary Fig. 24) and the expression correlation between locations decreasing with distance (Supplementary Fig. 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 MOB data[14], where we used scRNA-seq data[45] from 10x Chromium on the same tissue for deconvolution (Supplementary Tables 2 and 3). The MOB data consist 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; Fig. 3a and Methods). The cell-type compositions inferred by CARD accurately depict such expected layered structure[46], as is evident by visualizing either the first principal component (PC1) of the estimated cell-type composition matrix (Supplementary Fig. 26) or the inferred dominant cell types (Fig. 3b, Supplementary Table 4 and Supplementary Fig. 27). By 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 ONL from the GL. 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.

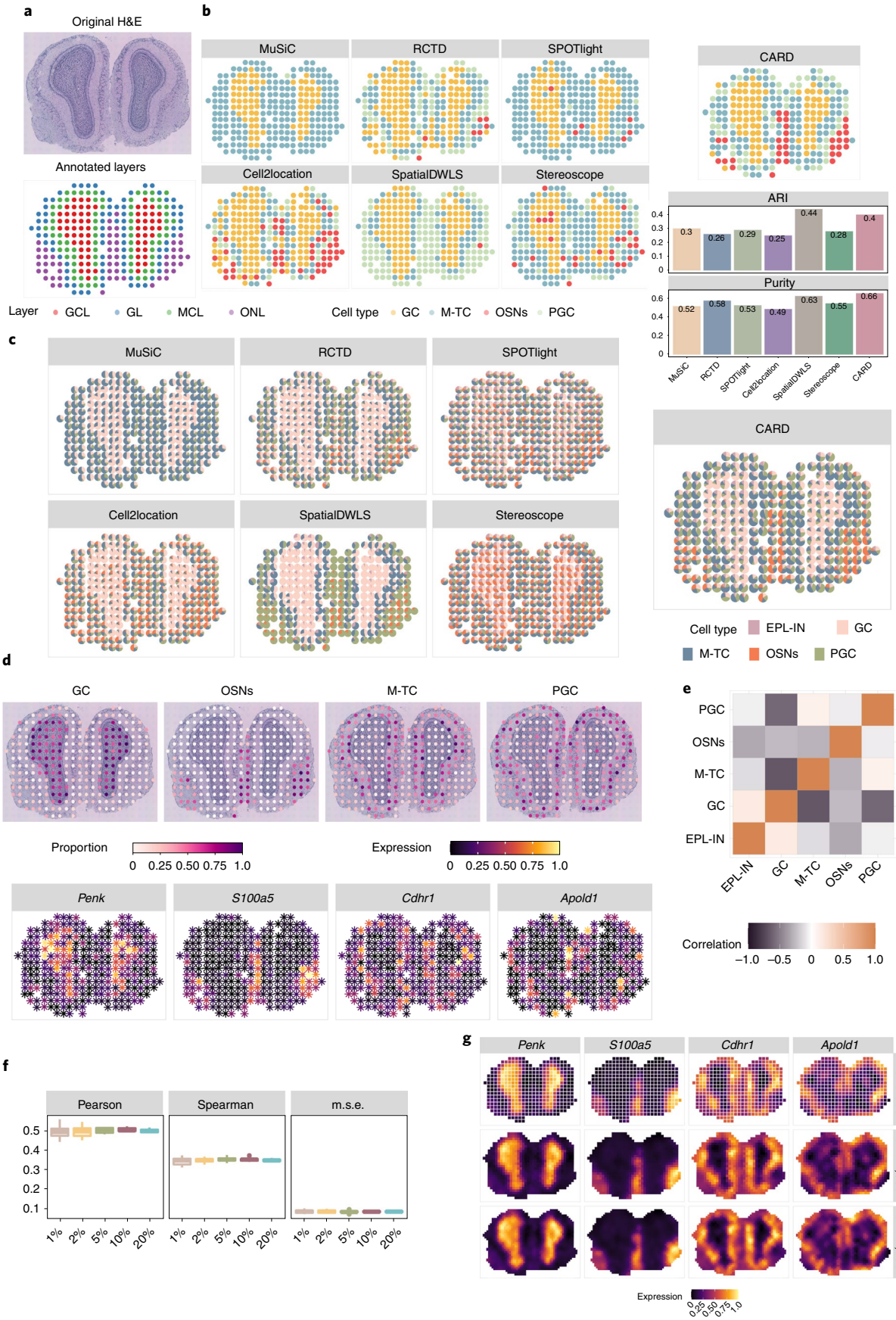
Careful examination of the cell-type composition and corresponding cell-type marker genes in different layers further confirmed the accuracy of CARD deconvolution (Fig. 3c,d and Supplementary Notes). For example, CARD distinguished correctly the adjacent MCL and GL, with distinct enrichment of mitral/tufted cells and periglomerular cells in the two layers, respectively, despite the similarity between these two cell types; however, others cannot (Supplementary Figs. 28 and 29). We also observed that multiple cell types inferred by CARD show spatial colocalization patterns (Fig. 3e and 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; Fig. 3f and Supplementary Fig. 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 (Methods), which captures fine-grained details of the layered structure in the olfactory bulb (Fig. 3g and Supplementary Figs. 31 and 32) and facilitates the identification of marker genes with spatial expression patterns (Supplementary Fig. 33 and Supplementary Notes). By contrast, the fixed resolution enhancement by BayesSpace failed to capture the expected spatial expression pattern for a few marker genes at high resolution (Supplementary Figs. 34 and 35). We quantitatively compared the performance of CARD and BayesSpace for resolution enhancement by performing clustering analysis on the imputed expression data (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 Fig. 36). CARD is also computationally efficient; CARD takes only 0.4 s 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 adenocarcinoma (PDAC) data. The second dataset we examined was a human PDAC dataset from ST[47]. For deconvolution, we first used a matched scRNA-seq dataset for the same individual obtained through inDrop[47] (denoted as PDAC-A). The PDAC data contain multiple tissue regions (cancer, pancreatic, ductal and stroma regions) annotated by histologists based on H&E staining[47] (Fig. 4a). Through deconvolution, CARD located various pancreatic and tumoral cell types into different tissue regions (Fig. 4b). The PC1 of the estimated cell-type composition matrix from CARD clearly captured a gross regional segregation between cancer and non-cancer regions, between the ductal and stroma regions and between the pancreatic and ductal regions. By contrast, none of the other methods were as effective in differentiating these regions (Supplementary Figs. 37– 40 and Supplementary Notes). The dominant cell types on each location from CARD also captured the segregation between cancer and non-cancer regions (Supplementary Fig. 41), with neoplastic cells, such as cancer clone A and clone B cells, highly enriched in the former (Wilcoxon test $P = 1.9 \times 10^{-48}$ and 1.1×10^{-43}, respectively; Fig. 4d). CARD also reveals the distinct distribution of two macrophage subpopulations between the cancer and non-cancer regions (Fig. 4d), representing a key functional signature of the regional compartmentalization of the cancer tissue that was missed by the other methods (Supplementary Fig. 42).

CARD further divides the cancer region into two subregions, a pattern missed by the other methods (Fig. 4b,c and Supplementary Figs. 41 and 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* (Fig. 4b,c and Supplementary Fig. 43). *S100A4*

Fig. 3 | Analyzing MOB data. a, H&E staining of the olfactory bulb (top) displays four anatomic layers that are organized in an inside–out fashion (bottom): the GCL, MCL, GL and ONL. **b**, Left, on each spatial location, the dominant cell type inferred from four different deconvolution methods is shown. The examined cell types include granule cells (GC), olfactory sensory neurons (OSNs), periglomerular cells (PGC), mitral/tufted cells (M-TC) and external plexiform layer interneurons (EPL-IN). EPL-IN is not the dominant cell type on any spatial location and is thus not shown in **b**. Compared deconvolution methods include MuSiC, RCTD, SPOTlight, cell2location, spatialDWLS, stereoscope and CARD. Right, ARI (top *y* axis) and purity (bottom *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 images, are shown. **c**, A spatial scatter pie plot displays inferred cell-type composition on each spatial location from different deconvolution methods. **d**, Top, the proportion of each of the four cell types inferred by CARD on each spatial location. Bottom, expression levels of the 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 ten replicates ($n = 10$). A fixed percentage of locations are masked as missing (*x* axis), and CARD is used to impute 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 m.s.e. Each box plot 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.


a
Original H&E
Annotated layers
Layer GCL GL MCL ONL
b
MuSiC
RCTD
SPOTlight
Cell2location
SpatialDWLS
Stereoscope
CARD
Cell type GC M-TC OSNs PGC
ARI
0.3 0.26 0.29 0.25 0.44 0.28 0.4
Purity
0.52 0.58 0.53 0.49 0.63 0.55 0.66
MuSiC RCTD SPOTlight Cell2location SpatialDWLS Stereoscope CARD
c
MuSiC
RCTD
SPOTlight
Cell2location
SpatialDWLS
Stereoscope
CARD
Cell type EPL-IN GC M-TC OSNs PGC
d
GC
OSNs
M-TC
PGC
Proportion 0 0.25 0.50 0.75 1.0
Expression 0 0.25 0.50 0.75 1.0
Penk
S100a5
Cdhr1
Apold1
e
PGC
OSNs
M-TC
GC
EPL-IN
Correlation −1.0 −0.5 0 0.5 1.0
f
Pearson
Spearman
m.s.e.
0.5
0.4
0.3
0.2
0.1
1% 2% 5% 10% 20%
g
Penk
S100a5
Cdhr1
Apold1
Grid 500
Grid 1,000
Grid 2,000
Expression 0 0.25 0.50 0.75 1.0

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. By contrast, *TM4SF1* is essential for PDAC migration and invasion[48–50], 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* (Fig. 4c), a cell type known to be associated with advanced tumor-node-metastasis stage[51].

CARD also localizes many other cell types into specific tissue regions, consistent with the expression pattern of the corresponding marker genes (Fig. 4b,c, Supplementary Fig. 43 and Supplementary Notes). By 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 colocalized spatially in PDAC (Fig. 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 and the role of the latter in pancreatic cancer stroma interaction of the tumor microenvironment[52,53]. The mean cell-type proportions inferred by CARD in the ST data are also highly correlated with those measured in the scRNA-seq dataset obtained from the same individual, more so than those obtained by the other methods (Fig. 4e).

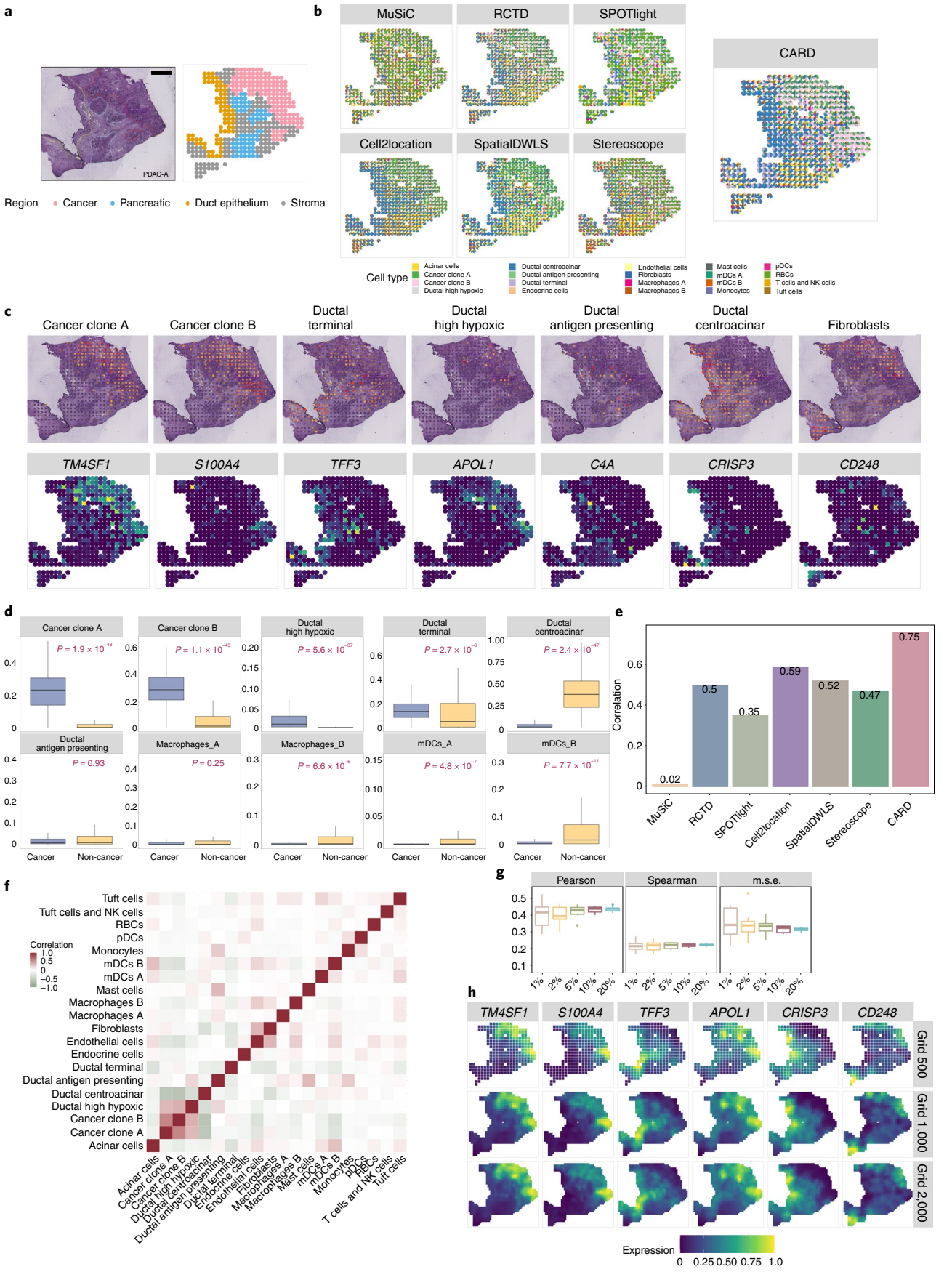
Next, we examined the robustness of deconvolution by using unmatched scRNA-seq datasets (Supplementary Table 2 and Supplementary Notes). Despite the platform and sample differences in the scRNA-seq references, we found that the estimated cell-type compositions for the major cell types are consistent across different scRNA-seq references, with the highest consistency achieved by CARD (Supplementary Fig. 44). Regardless of which unmatched scRNA-seq data were used, CARD showed superior performance compared to 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 tumor microenvironment supporting tumor progression[54–57] (Supplementary Figs. 45 and 46 and Supplementary Notes).

Finally, we found that the imputed gene expression by CARD is highly consistent with the truth across a range of masking percentages (Pearson's correlation = 0.29–0.52; Fig. 4g and Supplementary Fig. 47). Such consistency is higher when the matched scRNA-seq data from the same individual is used as the reference than when an unmatched scRNA-seq dataset is used (Supplementary Fig. 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 Fig. 49) and the spatial expression pattern of marker genes (Fig. 4h and Supplementary Fig. 50). Besides marker genes, CARD also discovered multiple genes that display clear spatial expression patterns in the refined spatial map but not in the original map (Supplementary Fig. 51 and Supplementary Notes). By contrast, the high-resolution map of BayesSpace does not show a clear pattern of multiple known marker genes (Supplementary Fig. 52) and additional genes (Supplementary Fig. 53). Clustering analysis on CARD-imputed high-resolution data also revealed clear segregation of the two cancer subregions, the normal pancreatic region, and the ductal region, more so than the original data or the refined data by BayesSpace (Supplementary Fig. 54).

Mouse hippocampus data from multiple sources. We analyzed two mouse hippocampus datasets: one directly on hippocampus measured using Slide-seqV2 (ref. [58]) 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[23,59] for deconvoluting both datasets (Supplementary Table 2). We only applied cell2location to the 10x Visium data but not the Slide-seqV2 data due to its heavy computational burden.

The hippocampus primarily consists of three regions, the cornu ammonis 1 (CA1)/CA2 region, the CA3 region and the dentate gyrus, all visualizable by total unique molecular identifier (UMI) counts per location displayed on the tissue (Fig. 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 the hippocampus accurately, more so than the other three methods (Fig. 5a and Supplementary Fig. 55). The dominant cell type on each location inferred by CARD also matches the expectation (Fig. 5b): CA1 cells are highly enriched in CA1, CA3 cells mainly localize in CA3, dentate cells reside in a C-shaped ring region of the dentate gyrus and ependymal cells form an irregular and columnar shape and line the ventricles of the brain[60], while choroid cells reside right below the ependymal cells and locate in the choroid plexus[61] along with Cajal–Retzius cells[62] (Supplementary Fig. 56). By contrast, MuSiC is unable to localize the main cell types, such as CA1 and CA3 cells, correctly and thus is unable to reveal the main structures of the hippocampus (Fig. 5b and Supplementary Fig. 57). SPOTlight detects an incorrectly diffused pattern of ependymal cells and incorrectly locates many CA3 cells to the CA1 region or outside the hippocampus (Fig. 5b and Supplementary Fig. 58). RCTD, spatialDWLS and stereoscope perform similarly, all locating CA3 cells incorrectly in CA1 (Fig. 5b and Supplementary Figs. 59–61), with the CA1 cell marker gene enriched in locations dominated by CA3 cells inferred by these methods (Supplementary Fig. 62). Additionally, they all allocate different cell types to hippocampal structures that appear to be much wider than expected[13,63](Fig. 5a,b). Careful examination of marker genes further confirmed the accuracy of CARD deconvolution (Fig. 5c).

Fig. 4 | Analyzing the PDAC data. a, H&E staining of the PDAC (left) displays four regions (right) annotated from the original publication[47]: cancer, pancreatic, ductal and stroma regions. **b**, A 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; mDCs, myeloid dendritic cells; pDCs, plasmacytoid dendritic cells; RBCs, red blood cells; NK cells, natural killer cells. **c**, Top, the proportion of each of the cell types inferred by CARD is displayed on each spatial location. Bottom, the expression levels of corresponding cell-type-specific marker genes are displayed. **d**, Comparisons of cell-type proportions inferred by CARD in cancer regions ($n = 137$) versus non-cancer regions ($n = 289$) with a P value tested by a 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. The color is scaled by the correlation value. **g**, Accuracy of CARD imputation in the masking analysis across ten 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 m.s.e. **h**, CARD imputes gene expression for six 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 box plot 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.


a
PDAC-A
Region Cancer Pancreatic Duct epithelium Stroma
b
MuSiC
RCTD
SPOTlight
Cell2location
SpatialDWLS
Stereoscope
CARD
Cell type
Acinar cells
Cancer clone A
Cancer clone B
Ductal high hypoxic
Ductal centroacinar
Ductal antigen presenting
Ductal terminal
Endocrine cells
Endothelial cells
Fibroblasts
Macrophages A
Macrophages B
Mast cells
mDCs A
mDCs B
Monocytes
pDCs
RBCs
T cells and NK cells
Tuft cells
c
Cancer clone A
Cancer clone B
Ductal terminal
Ductal high hypoxic
Ductal antigen presenting
Ductal centroacinar
Fibroblasts
TM4SF1
S100A4
TFF3
APOL1
C4A
CRISP3
CD248
d
Cancer clone A
P = 1.9 × 10^-48
Cancer clone B
P = 1.1 × 10^-43
Ductal high hypoxic
P = 5.6 × 10^-37
Ductal terminal
P = 2.7 × 10^-6
Ductal centroacinar
P = 2.4 × 10^-47
Ductal antigen presenting
P = 0.93
Macrophages_A
P = 0.25
Macrophages_B
P = 6.6 × 10^-6
mDCs_A
P = 4.8 × 10^-7
mDCs_B
P = 7.7 × 10^-11
Cancer
Non-cancer
e
Correlation
0.02
0.5
0.35
0.59
0.52
0.47
0.75
MuSiC
RCTD
SPOTlight
Cell2location
SpatialDWLS
Stereoscope
CARD
f
Correlation
Tuft cells
Tuft cells and NK cells
RBCs
pDCs
Monocytes
mDCs B
mDCs A
Mast cells
Macrophages B
Macrophages A
Fibroblasts
Endothelial cells
Endocrine cells
Ductal terminal
Ductal antigen presenting
Ductal centroacinar
Ductal high hypoxic
Cancer clone B
Cancer clone A
Acinar cells
g
Pearson
Spearman
m.s.e.
1%
2%
5%
10%
20%
h
TM4SF1
S100A4
TFF3
APOL1
CRISP3
CD248
Grid 500
Grid 1,000
Grid 2,000
Expression
0 0.25 0.50 0.75 1.0

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 supported more accurate deconvolution by CARD than the other methods (Fig. 5d, Supplementary Fig. 63 and Supplementary Notes).

We observed that multiple cell types inferred by CARD are colocalized together (Supplementary Fig. 64). The highest colocalization occurs between *Slc17a6*/*Vglut2* neurons and entorhinal cells, highlighting the cell compositional architecture underlying the hippocampus–entorhinal cortex network[64]. The imputed gene expression data by CARD are consistent with the truth across a range of masking percentages (Supplementary Fig. 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 the hippocampus (Supplementary Fig. 66), with the refined gene expression recovering strong spatial patterns for various marker genes (Fig. 5e and Supplementary Fig. 67) and additional genes (Supplementary Fig. 68 and Supplementary Notes). We examined the reliability of the refined spatial map by creating a low-resolution version of the Slide-seqV2 data and then applied CARD to construct a refined spatial expression map at the original Slide-seqV2 resolution (Supplementary Notes). We found that the refined spatial map recovers a consistent and sometimes stronger spatial pattern than the original Slide-seqV2 data (Supplementary Figs. 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 data. Again, CARD captures the key structures of the hippocampus (Fig. 5f,g). The estimated cell-type compositions on CA1, CA3 and the 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 (Fig. 5f and Supplementary Fig. 73), a pattern confirmed with quantifications (Supplementary Figs. 74 and 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 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 same general tissue domain segregation pattern as CARD in both MOB and PDAC data, although it was unable to differentiate the two cancer subregions as CARD did (Supplementary Fig. 76). CARDfree does not perform as well as CARD with the high-resolution Slide-seqV2 data and did not identify the CA3 structure based on its estimated cell-type proportions, as the Slide-seqV2 data are highly sparse and thus could benefit from reference-based deconvolution. However, in the hippocampus region of the Slide-seqV2 data, we did notice that CARDfree identified a region with a unique cell-type composition (Supplementary Fig. 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[65]. The results suggest that reference-free deconvolution may sometimes have added benefits.

Discussion

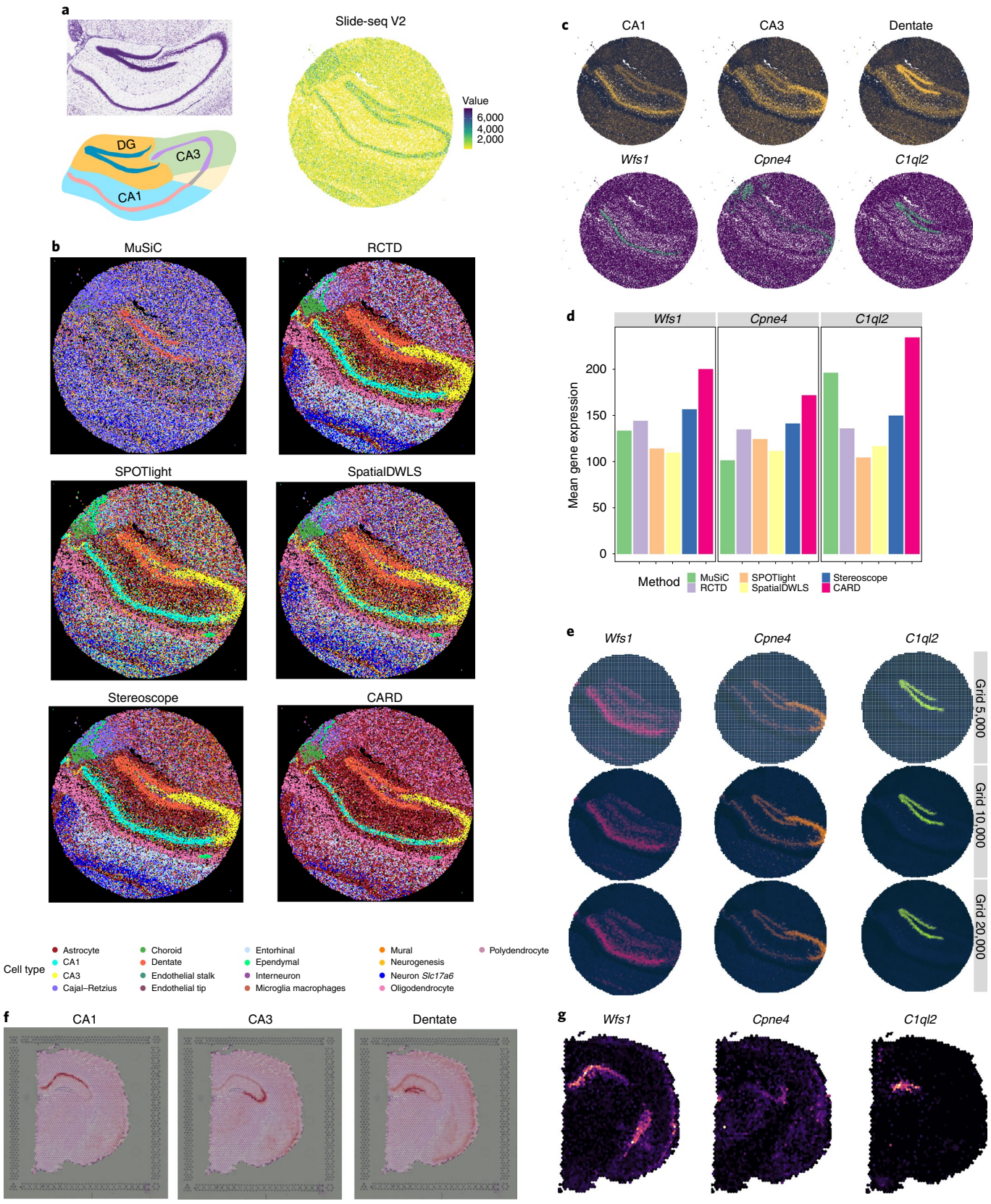
We have presented CARD for accurate and spatially informed deconvolution of spatial transcriptomics data. CARD is computationally efficient: it is 0.8–7,761.8 times faster and uses 0.2–109% of the physical memory compared to the other deconvolution methods (Supplementary Fig. 78 and Supplementary Table 5), and it is 5,875–7,028 times faster and uses only 14–17% of the physical memory compared to BayesSpace in creating refined spatial maps (Supplementary Figs. 79 and 80 and 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[66] and MERFISH[67], mostly rely on single-molecule fluorescence 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[68–70]. 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 mouse cortex data from seqFISH+[38], 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 Figs. 81 and 82).

We have presented an extension of CARD, CARDfree, for reference-free deconvolution. CARDfree requires a postprocessing step to correctly label the inferred cell types. Such postprocessing often requires cell-type-specific gene expression profiles and can be challenging to perform accurately. For example, in PDAC, CARDfree infers cell-type composition on each location for 20 inferred cell types. However, 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 corresponds to ductal centroacinar cells or endothelial cells, as markers for both cell types are enriched in locations with a high proportion of cell type 14 cells (Supplementary Fig. 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 Figs. 84–90). Such extension requires knowing the spatial localization information for all

Fig. 5 | Analyzing the hippocampus region in Slide-seqV2 and 10x Visium mouse brain (coronal) data. a, The UMI counts of Slide-seqV2 data (right) display the structure and shape of hippocampal tissue, highly consistent with the image from Allen Reference Atlas (left); DG, dentate gyrus. **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, the proportion of each of the cell types inferred by CARD is displayed on each spatial location. Bottom, 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 three 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.

single cells on the tissue, which remains challenging to obtain from non-single-cell resolution spatial transcriptomics data. Because the spatial transcriptomics data itself do 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 (for example, see Supplementary Fig. 91). In addition, aligning H&E images with



spatial transcriptomics data can be computationally challenging[71]. Future efforts are needed to address these challenges.

Additional extensions of CARD are possible. First, CARD models normalized spatial transcriptomics data and could benefit from extensions for direct modeling of raw count data using an overdispersed Poisson model[72,73]. Second, we only explored the use of the Gaussian kernel[74] for modeling spatial correlation. Exploring the use of other kernels, such as the periodic kernels[74], or incorporating histological image information, such as image intensity level, as additional coordinates[6,75], 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 resulting 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 data indeed produced the best imputation performance and would be selected as the optimal reference data for deconvolution.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-022-01273-7.

Received: 9 June 2021; Accepted: 7 March 2022;
Published online: 02 May 2022

References

1. Burgess, D. J. Spatial transcriptomics coming of age. *Nat. Rev. Genet.* **20**, 317 (2019).
2. Soldatov, R. et al. Spatiotemporal structure of cell fate decisions in murine neural crest. *Science* **364**, eaas9536 (2019).
3. Prinz, M., Priller, J., Sisodia, S. S. & Ransohoff, R. M. Heterogeneity of CNS myeloid cells and their roles in neurodegeneration. *Nat. Neurosci.* **14**, 1227–1235 (2011).
4. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: Identification of spatially variable genes. *Nat. Methods* **15**, 343–346 (2018).
5. Dries, R. et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol.* **22**, 78 (2021).
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. Preprint at *bioRxiv* https://doi.org/10.1101/2020.05.31.125658 (2020).
7. Biancalani, T. et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat. Methods* **18**, 1352–1362 (2021).
8. Fu, H. et al. Unsupervised spatially embedded deep representation of spatial transcriptomics. Preprint at *bioRxiv* https://doi.org/10.1101/2021.06.15.448542 (2021).
9. Fischl, A. M., Heron, P. M., Stromberg, A. J. & McClintock, T. S. Activity-dependent genes in mouse olfactory sensory neurons. *Chem. Senses* **39**, 439–449 (2014).
10. Moses, L. & Pachter, L. Museum of spatial transcriptomics. *Nat. Methods* https://doi.org/10.1038/s41592-022-01409-2 (2022).
11. Asp, M., Bergenstråhle, J. & Lundeberg, J. Spatially resolved transcriptomes—next generation tools for tissue exploration. *Bioessays* **42**, e1900221 (2020).
12. Genomics, 10x. 10x Genomics Visium. https://www.10xgenomics.com/spatial-transcriptomics/
13. Rodriques, S. G. et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).
14. Ståhl, P. L. et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
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 Biotechnol.* **39**, 43–58 (2020).
16. Rao, A., Barkley, D., França, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).
17. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50**, 1–14 (2018).
18. Cobos, F. A., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.* **11**, 5650 (2020).
19. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**, 380 (2019).
20. Dong, M. et al. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Brief. Bioinform.* **22**, 416–427 (2020).
21. Jew, B. et al. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nat. Commun.* **11**, 1971 (2020).
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 (2021).
23. Cable, D. M. et al. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-021-00830-w (2021).
24. Song, Q. & Su, J. DSTG: deconvoluting spatial transcriptomics data through graph-based artificial intelligence. *Brief. Bioinform.* **22**, bbaa414 (2021).
25. Lopez, R. et al. Multi-resolution deconvolution of spatial transcriptomics data reveals continuous patterns of inflammation. Preprint at *bioRxiv* https://doi.org/10.1101/2021.05.10.443517 (2021).
26. Danaher, P. et al. Advances in mixed cell deconvolution enable quantification of cell types in spatial transcriptomic data. *Nat. Commun.* **13**, 385 (2022).
27. Gayoso, A. et al. scvi-tools: a library for deep probabilistic analysis of single-cell omics data. Preprint at *bioRxiv* https://doi.org/10.1101/2021.04.28.441833 (2021).
28. Andersson, A. et al. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun. Biol.* **3**, 565 (2020).
29. Kleshchevnikov, V. et al. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-021-01139-4 (2022).
30. Dong, R. & Yuan, G.-C. SpatialDWLS: accurate deconvolution of spatial transcriptomic data. *Genome Biol.* **22**, 145 (2021).
31. Stoltzfus, C. R. et al. CytoMAP: a spatial analysis toolbox reveals features of myeloid cell organization in lymphoid tissues. *Cell Rep.* **31**, 107523 (2020).
32. 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).
33. Bove, A. et al. Local cellular neighborhood controls proliferation in cell competition. *Mol. Biol. Cell* **28**, 3215–3228 (2017).
34. 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).
35. Phillips, D. et al. Immune cell topography predicts response to PD-1 blockade in cutaneous T cell lymphoma. *Nat. Commun.* **12**, 6726 (2021).
36. Schürch, C. M. et al. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell* **182**, 1341–1359 (2020).
37. 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. USA* **116**, 19490–19499 (2019).
38. Eng, C. H. L. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* **568**, 235–239 (2019).
39. Banerjee, S., Carlin, B. P. & Gelfand, A. E. *Hierarchical Modeling and Analysis for Spatial Data* 2nd edn (Chapman and Hall/CRC, 2014).
40. Lee, D. A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spat. Spatiotemporal Epidemiol.* **2**, 79–89 (2011).
41. Zhao, E. et al. Spatial transcriptomics at subspot resolution with BayesSpace. *Nat. Biotechnol.* **39**, 1375–1384 (2021).
42. Zeisel, A. et al. Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014 (2018).
43. Yang, S. & Zhou, X. Accurate and scalable construction of polygenic scores in large biobank data sets. *Am. J. Hum. Genet.* **106**, 679–693 (2020).
44. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet.* **9**, e1003264 (2013).
45. 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.* **25**, 2689–2703 (2018).
46. Nagayama, S., Homma, R. & Imamura, F. Neuronal organization of olfactory bulb circuits. *Front. Neural Circuits* **8**, 98 (2014).
47. Moncada, R. et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat. Biotechnol.* **38**, 333–342 (2020).

48. 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.* **47**, 490–498 (2015).
49. Fu, F. et al. Role of transmembrane 4 L six family 1 in the development and progression of cancer. *Front. Mol. Biosci.* **7**, 202 (2020).
50. 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).
51. Zhang, X. et al. Expression pattern of cancer-associated fibroblast and its clinical relevance in intrahepatic cholangiocarcinoma. *Hum. Pathol.* **65**, 92–100 (2017).
52. Morvaridi, S., Dhall, D., Greene, M. I., Pandol, S. J. & Wang, Q. Role of YAP and TAZ in pancreatic ductal adenocarcinoma and in stellate cells associated with cancer and chronic pancreatitis. *Sci. Rep.* **5**, 16759 (2015).
53. Nielsen, M. F. B., Mortensen, M. B. & Detlefsen, S. Key players in pancreatic cancer-stroma interaction: cancer-associated fibroblasts, endothelial and inflammatory cells. *World J. Gastroenterol.* **22**, 2678–2700 (2016).
54. Zheng, C. et al. Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. *Cell* **169**, 1342–1356 (2017).
55. Comito, G., Ippolito, L., Chiarugi, P. & Cirri, P. Nutritional exchanges within tumor microenvironment: impact for cancer aggressiveness. *Front. Oncol.* **10**, 396 (2020).
56. Lambrechts, D. et al. Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat. Med.* **24**, 1277–1289 (2018).
57. Peng, J. et al. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Res.* **29**, 725–738 (2019).
58. Stickels, R. R. et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39**, 313–319 (2021).
59. Saunders, A. et al. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* **174**, 1015–1030 (2018).
60. Del Bigio, M. R. Ependymal cells: biology and pathology. *Acta Neuropathol.* **119**, 55–73 (2010).
61. Ramachandran, V. S. (ed.) *Encyclopedia of the Human Brain*, 1st edn (Elsevier, 2003).
62. Meyer, G. Building a human cortex: the evolutionary differentiation of Cajal–Retzius cells and the cortical hem. *J. Anat.* **217**, 334–343 (2010).
63. Hawrylycz, M. et al. in *Springer Handbook of Bio-/Neuroinformatics* (ed. Kasabov N.) 1111–1126 (Springer, 2014).
64. Wozny, C. et al. VGLUT2 functions as a differential marker for hippocampal output neurons. *Front. Cell. Neurosci.* **12**, 337 (2018).
65. 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).
66. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* **11**, 360–361 (2014).
67. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
68. Moffitt, J. R. et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, eaau5324 (2018).
69. He, Y. et al. ClusterMap for multi-scale clustering analysis of spatial gene expression. *Nat. Commun.* **12**, 5909 (2021).
70. Moen, E. et al. Deep learning for cellular image analysis. *Nat. Methods* **16**, 1233–1246 (2019).
71. Bergenstråhle, J., Larsson, L. & Lundeberg, J. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. *BMC Genomics* **21**, 482 (2020).
72. Sun, S. et al. Differential expression analysis for RNAseq using Poisson mixed models. *Nucleic Acids Res.* **45**, e106 (2017).
73. 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).
74. Sun, S., Zhu, J. & Zhou, X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat. Methods* **17**, 193–200 (2020).
75. Hu, J. et al. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods.* **18**, 1342–1351 (2021).

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 scRNA-seq data as input. The scRNA-seq data serve as a reference and consist of K cell types with a set of G cell-type-informative genes. Cell types and informative genes in scRNA-seq data can be obtained through standard analysis pipelines for clustering and informative gene identification[76,77]. In scRNA-seq, we denote 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 B is commonly referred to as the reference basis matrix. In the spatial transcriptomics data, we denote X as the G-by-N gene expression matrix for the same set of informative genes measured on N spatial locations. We denote V as the N-by-K cell-type composition matrix, where each row of V represents the proportions of the K cell types on each spatial location. Our objective is to estimate V given both X from the spatial transcriptomics data and B constructed from the scRNA-seq data. To do so, we consider a non-negative matrix factorization model to link the three matrices:

$$X = BV^T + E, \tag{1}$$

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

The non-negative matrix factorization model in Eq. (1) has been applied for cell-type deconvolution in bulk RNA-seq studies. However, this model is not directly applicable for deconvoluting spatial transcriptomics data, 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 V in the above factorization model. Thus, modeling spatial correlation in 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 V. To accommodate the spatial correlation in V, we specify a CAR[39,78,79] modeling assumption on each column of V. Specifically, for the column/cell type k, we assume

$$V_{ik} = b_k + \phi \sum_{j=1, j\neq i}^{n} W_{ij} \left(V_{jk} - b_k\right) + \epsilon_{ik}, \tag{2}$$

where V_{ik} represents the proportion of cell type k on the ith location, b_k is the kth cell-type-specific intercept that represents the average cell-type composition across locations, W is an N-by-N non-negative weight matrix with each element W_{ij} specifying the weight used for inferring the cell-type composition on the ith location based on the cell-type composition information on the jth location, ϕ is a spatial autocorrelation parameter that determines the strength of the spatial correlation in cell-type composition, and ϵ_{ik} is the residual error that follows a normal distribution $\epsilon_{ik} \sim N(0, \sigma_{ik}^2)$. The CAR modeling assumption on V effectively expresses the composition of the kth cell type on the ith location, V_{ik}, as a weighted summation of the kth cell-type compositions on all other locations, $V_{jk} (j \neq i)$. 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 ref. [74] to express the weight matrix W in the form of a Gaussian kernel function constructed based on the Euclidean distance between pairs of spatial locations (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[80,81]. While we primarily focus on using a Gaussian kernel for W, our method and software can easily incorporate other types of kernels to capture diverse spatial correlation patterns encountered in different datasets. With the Gaussian kernel matrix W, we further obtain a row-standardized weight matrix $\widetilde{W}$ through transformation $\widetilde{W}_{ij} = W_{ij}/W_{i+}$, with $W_{i+} = \sum_{j=1}^{n} W_{ij}$. Because the weight matrix and the residual error variance need to satisfy the symmetric condition[82,83], we set $\sigma_{ik}^2 = \lambda_k/W_{i+}$ to ensure $\widetilde{W}_{ij}\sigma_{jk}^2 = \widetilde{W}_{ji}\sigma_{ik}^2$, where λ_k is a scalar. With the above parameterization, we can follow the Brook's Lemma[79,84] to obtain the joint distribution for the N-size column vector V_k as

$$V_k \sim MVN(b_k 1_N, \Sigma_k), \tag{3}$$

where 1_N is an N vector of 1s, $\Sigma_k = \left(I_N - \phi\widetilde{W}\right)^{-1} M_k$ is a positive definite covariance matrix with $M_k = \text{diag}\left(\sigma_{1k}^2, \ldots \sigma_{Nk}^2\right)$, and MVN denotes a multivariate normal distribution (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 V. By modeling the spatial correlation in 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 V, with non-negativity constraints on each of its elements (Supplementary Notes). Our algorithm treats the hyperparameters (b_k, λ_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 make 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 data 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[85,86] for updating the non-negative 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 each column in the cell-type composition matrix 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 V. By modeling the spatial correlation in 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[87] 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 V and denote the corresponding matrix on the N^* new locations as V^*. Based on Eq. (3), the $(N + N^*)$-sized cell-type composition vector for the kth cell type, $(V_k, V_k^*)^T$, follows a multivariate normal distribution $MVN(b_k 1_{N+N^*}, \Sigma)$. We partition the covariance matrix Σ into $\begin{bmatrix} \Sigma_{oo} & \Sigma_{on} \\ \Sigma_{no} & \Sigma_{nn} \end{bmatrix}$, where o are the indices that correspond to the original locations, and n are the indices that correspond to the new locations. We can then estimate V_k^* via its conditional mean

$$\widetilde{V}_k^* = b_k 1_{N^*} + \Sigma_{no} \sum\nolimits_{oo}^{-1} (V_k - b_k 1_N), \tag{4}$$

where the parameters on the right of the equation are replaced by the corresponding estimates. The estimates $\widetilde{V}_k^*$ 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 Σ by using only the nearest ten 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 Eq. (4) with the basis matrix to obtain $B\ \widetilde{V}_k^*$.

Basis matrix construction. We constructed the reference basis matrix B following the main ideas of MuSiC using three detailed steps (Supplementary Notes). We (1) selected genes that are expressed in both the scRNA-seq reference data and the spatial transcriptomic data, (2) selected among them the candidate cell-type-informative genes with a mean expression level in a given cell type of at least 1.25-log-fold higher than its mean expression level across all remaining cell types and (3) removed among them the outlier genes that show high expression heterogeneity within a cell type by calculating gene-specific expression dispersion (Supplementary Fig. 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% of 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 V. After we obtained the estimated cell-type composition matrix $\hat{V}$, we evaluated deconvolution performance by computing the r.m.s.e. between $\hat{V}$ and V through

$$\text{r.m.s.e.} = \sqrt{\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(V_{ik} - \hat{V}_{ik}\right)^2}$$, 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 r.m.s.e. calculation is based on all cell types (Supplementary Notes).

Compared methods. We compared CARD with six 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[29] (version 0.07a), (5) spatialDWLS (implemented in the R package Giotto, version 1.0.4) and (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 that users 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 PDAC data and set it to be 10 for the 10x Visium data. We set both the factors_per_spot and combs_per_spot values 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 a neighborhood structure suitable for ST and 10x Visium data, we only evaluated its performance on the mouse olfactory spatial transcriptomics data and human 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 subspots at each location in the original data (five for Visium technology and nine for ST technology). To compare the high-resolution spatial gene expression data constructed by CARD and BayesSpace on the same set of subspots, we applied CARD to directly impute gene expression on the subspots generated by BayesSpace. Afterward, we performed principal-component analysis dimension reduction on the high-resolution data and applied the K-means algorithm analysis on the top 20 PCs to cluster spatial locations into 6 clusters for the mouse olfactory data and 18 clusters for the PDAC data 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 did 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[88]. We reasoned that, if the cell-type composition is similar among neighboring locations, then we would also expect the cell-type marker 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 performing spatial autocorrelation tests using Moran I (ref. [89]) and Geary's C (ref. [90]). Note that we were unable to perform Moran's I test[91] and Geary's C test[91] on the large Slide-seqV2 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 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 ref. [92] 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 GCL (which contains $n = 67$ spatial locations), the MCL ($n = 75$), the GL ($n = 80$) and the 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 images via 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 heterogeneous than 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 Pearson's correlation, Spearman's correlation and m.s.e. These serve as indicators on how accurate CARD imputation works, which also reflect its deconvolution performance. Importantly, the magnitude of m.s.e. 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 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 MOB dataset (http://www.spatialtranscriptomicsresearch.org), human 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 dataset (https://www.10xgenomics.com/resources/datasets/). For the scRNA-seq references used in this study, all are publicly available, with details provided in Supplementary Tables 2 and 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 analyses are also available at the same website.

References

76. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261 (2018).
77. Duò, A., Robinson, M. D. & Soneson, C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res.* **7**, 1141 (2018).
78. De Oliveira, V. Bayesian analysis of conditional autoregressive models. *Ann. Inst. Stat. Math.* **64**, 107–133 (2012).
79. Besag, J. Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Ser. B* **36**, 192–225 (1974).
80. Vanhatalo, J., Pietiläinen, V. & Vehtari, A. Approximate inference for disease mapping with sparse Gaussian processes. *Stat. Med.* **29**, 1580–1607 (2010).
81. Rousset, F. & Ferdy, J. B. Testing environmental and genetic effects in the presence of spatial autocorrelation. *Ecography* **37**, 781–790 (2014).
82. Cressie, N. Statistics for spatial data. *Terra Nova* **4**, 613–617 (1992).
83. Rue, H. & Held, L. *Gaussian Markov Random Fields: Theory and Applications* 1st edn (Chapman and Hall/CRC, 2005).
84. 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).
85. Lee, D. D. & Seung, H. S. in *Advances in Neural Information Processing Systems* (eds Leen T., Dietterich T. & Tresp V.) 556–562 (MIT Press, 2001).
86. 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).
87. 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).

88. Ralston, A. & Shaw, K. Gene expression regulates cell differentiation. *Nat. Educ.* **1**, 127–131 (2008).
89. Li, H., Calder, C. A. & Cressie, N. Beyond Moran's *I*: testing for spatial dependence based on the spatial autoregressive model. *Geogr. Anal.* **39**, 357–375 (2007).
90. Radeloff, V. C., Miller, T. F., He, H. S. & Mladenoff, D. J. Periodicity in spatial data and geostatistical models: autocorrelation between patches. *Ecography* **23**, 81–91 (2000).
91. Bivand, R. et al. *spdep: Spatial Dependence: Weighting Schemes, Statistics and Models.* R package version 0.5-37 (2011).
92. Teschendorff, A. E., Zhu, T., Breeze, C. E. & Beck, S. EPISCORE: cell type deconvolution of bulk tissue DNA methylomes from single-cell RNA-seq data. *Genome Biol.* **21**, 221 (2020).

Acknowledgements

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

Author contributions

X.Z. conceived the idea and provided funding support. Y.M. and X.Z. designed the experiments. Y.M. developed the method, implemented the software, performed simulations and analyzed real data. Y.M. and X.Z. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-022-01273-7.

Correspondence and requests for materials should be addressed to Xiang Zhou.

Peer review information *Nature Biotechnology* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

nature portfolio

Corresponding author(s): Xiang Zhou

Last updated by author(s): Feb 22, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection: No software is used in data collecting

Data analysis

We used the newly developed R package CARD for data analysis. CARD is described in the Methods section and deposited at github [https://github.com/YingMa0107/CARD] (R package, version 1.0). In addition, we also used the following software for comparative analysis.
MuSiC[https://github.com/xuranw/MuSiC](R package, version 0.1.1): a statistical method for estimating cell type proportions in bulk RNA-seq data using cross-subject scRNA-seq as reference.
SPOTlight[https://github.com/MarcElosua/SPOTlight](R package, version 0.1.0): a statistical method for decomposition of cell type mixtures in spatial transcriptomics data based on topic profile signatures and NMF model.
RCTD[https://github.com/dmcable/RCTD/](R package, version 1.1.0): a statistical method for robust decomposition of cell type mixtures in spatial transcriptomics data based on Poisson factor analysis model.
cell2location[https://github.com/BayraktarLab/cell2location/] (Python package, version 0.07a): a statistical method for comprehensive mapping of tissue cell architecture via integrated single cell and spatial transcriptomics data.
spatialDWLS[https://github.com/RubD/Giotto] (R package, version 1.1.0): a statistical method that extended the previous bulk RNAseq deconvolution methods DWLS into deconvoluting cell type compositions in spatial transcriptomics data. Specifically, the method was implemented in the package Giotto
stereoscope[https://github.com/almaan/stereoscope] (Python package, version 0.2.0): a model-based probabilistic method that uses single cell data to deconvolve the cell mixtures in spatial transcriptomics data.
BayesSpace[https://github.com/edward130603/BayesSpace] (R package, version 1.5.1): a statistical method for clustering and enhancing the resolution of spatial gene expression experiments.
In addition, we also used the following packages for data analysis: Seurat [https://github.com/satijalab/seurat] (R package, version 4.1.0), igraph [https://github.com/igraph/igraph] (R package, 1.0.0), funtimes (R package, version 8.1)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

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/). The following scRNAseq references used in this study are publicly available with details provided in supplementary tables 2-3:

• Simulation Scenarios I – IV scRNAseq reference dataset (Zeisel et al. 2018). Count matrices were downloaded from http://mousebrain.org/adolescent/loomfiles_level_L1.html
• Simulation Scenario V scRNAseq reference dataset (Mizrak et al. 2019). Count matrices were downloaded from the Gene Expression Omnibus website (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE109447) with the GEO accession number GSE109447
• Mouse olfactory bulb scRNAseq reference dataset (Tepe et al. 2018). Count matrices were downloaded from the Gene Expression Omnibus website (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE121891) with the GEO accession number GSE121891
• Pancreatic ductal adenocarcinomas scRNAseq dataset (Moncada et al. 2020). Count matrices were downloaded from the Gene Expression Omnibus website (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111672) with the GEO accession number GSE111672. In this website, we downloaded both sample matched scRNASeq dataset and sample unmatched scRNAseq dataset
• Peng scRNAseq dataset (Peng et al. 2019). Count matrices were downloaded from the Genome Sequence Archive under project PRJCA001063
• Mouse Hippocampus scRNAseq dataset (Saunders et al. 2018). Count matrices were downloaded from the Broad Institute Single Cell Portal at https://singlecell.broadinstitute.org/single_cell/study/SCP948
• Mouse brain cortex scRNAseq dataset (Hrvatin et al. 2018). Count matrices were downloaded from the Gene Expression Omnibus website (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE102827) with the GEO accession number GSE102827

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	CARD was evaluated across four spatially resolved transcriptomics datasets in real data applications and discussed one potential application of CARD using 1 single cell resolution seqFISH+ dataset, coupled with 8 single cell RNAseq data references. Besides real data applications, this paper also performed analysis using 2 single cell RNAseq datasets as references.
Data exclusions	For each spatial transcriptomics and single cell RNASeq reference dataset, we used all spots or or cells that were detected to have non-zero expression as determined by their original publications. No data were excluded from this study.
Replication	To evaluate the deconvolution accuracy of CARD and other deconvolution methods, we performed simulations and each simulation was replicated for 5 times to check variability as well as stability of each methods we compared. All attempts at replication were successful.
Randomization	Randomization is not relevant to this study because each sample or slide was analyzed separately.
Blinding	Blinding is not relevant to our study because we don't compare any case/control groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging