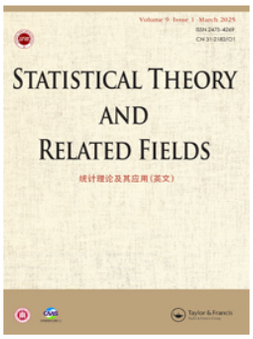

Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: www.tandfonline.com/journals/tstf20

A Bayesian hierarchical model with spatially varying dispersion for reference-free cell type deconvolution in spatial transcriptomics

Xuan Li, Yincai Tang, Jingsi Ming & Xingjie Shi

To cite this article: Xuan Li, Yincai Tang, Jingsi Ming & Xingjie Shi (08 May 2025): A Bayesian hierarchical model with spatially varying dispersion for reference-free cell type deconvolution in spatial transcriptomics, Statistical Theory and Related Fields, DOI: 10.1080/24754269.2025.2495651

To link to this article: https://doi.org/10.1080/24754269.2025.2495651

STATISTICAL THEORY AND RELATED FIELDS
https://doi.org/10.1080/24754269.2025.2495651

Taylor & Francis
Taylor & Francis Group

A Bayesian hierarchical model with spatially varying dispersion for reference-free cell type deconvolution in spatial transcriptomics

Xuan Li[a], Yincai Tang[a], Jingsi Ming[a,b] and Xingjie Shi[a,b]

[a]KLATASDS-MOE, School of Statistics, East China Normal University, Shanghai, People's Republic of China; [b]Academy of Statistics and Interdisciplinary Sciences, School of Statistics, East China Normal University, Shanghai, People's Republic of China

ABSTRACT

A major challenge in spatial transcriptomics (ST) is resolving cellular composition, especially in technologies lacking single-cell resolution. The mixture of transcriptional signals within spatial spots complicates deconvolution and downstream analyses. To uncover the spatial heterogeneity of tissues, we introduce SvdRFCTD, a reference-free spatial transcriptomics deconvolution method, which estimates the cell type proportions at each spot on the tissue. To fully capture the heterogeneity in the ST data, we combine SvdRFCTD with a Bayesian hierarchical negative binomial model with spatial effects incorporated in both the mean and dispersion of the gene expression, which is used to explicitly model the generative mechanism of cell type proportions. By integrating spatial information and leveraging marker gene information, SvdRFCTD accurately estimates cell type proportions and uncovers complex spatial patterns. We demonstrate the ability of SvdRFCTD to identify cell types on simulated datasets. By applying SvdRFCTD to mouse brain and human pancreatic ductal adenocarcinomas datasets, we observe significant cellular heterogeneity within the tissue sections and successfully identify regions with high proportions of aggregated cell types, along with the spatial relationships between different cell types.

ARTICLE HISTORY
Received 18 February 2025
Revised 19 March 2025
Accepted 15 April 2025

KEYWORDS
Spatial transcriptomics; reference-free deconvolution; tissue heterogeneity; spatial pattern; Bayesian hierarchical model

1. Background

Spatial transcriptomics represents a transformative technology that integrates gene expression data with spatial information, allowing for a deeper understanding of tissue organization and function (Ståhl et al., 2016; Williams et al., 2022). Unlike traditional RNA sequencing methods, which provide bulk gene expression profiles, ST captures gene expression at multiple spatial locations within a tissue, revealing cellular heterogeneity and spatial patterns (Larsson et al., 2021). This ability to study gene expression at subcellular and cellular resolution enables the mapping of tissue structures in unprecedented detail (Vickovic et al., 2019). However, a significant challenge in ST is that most datasets lack single-cell resolution, with each spatial spot containing multiple cell types, which complicates the accurate estimation

CONTACT Jingsi Ming jsming@fem.ecnu.edu.cn Academy of Statistics and Interdisciplinary Sciences, East China Normal University, 3663 North Zhongshan Road, Shanghai 200062, People's Republic of China

Supplemental data for this article can be accessed online at http://dx.doi.org/10.1080/24754269.2025.2495651.

of cell type proportions and interactions (Rao et al., 2021). This is particularly problematic when studying tissues with high cellular complexity, where precise knowledge of the cellular composition is crucial for understanding tissue function and disease mechanisms (Levy-Jurgenson et al., 2020).

Deconvolution methods have emerged as a promising solution to this challenge, aiming to resolve the cellular composition at each spatial spot by estimating the proportions of different cell types (Zhang et al., 2023). Most current ST deconvolution methods rely on external references, such as single-cell RNA sequencing (scRNA-seq) data and single-nucleus RNA-seq (snRNA-seq), to provide gene expression profiles for various cell types (B. Li et al., 2022). For instance, RCTD (Cable et al., 2022) utilized a hierarchical Poisson regression model for gene expression to infer cell type profiles from the reference dataset. By incorporating variables that account for platform effects, it effectively addressed estimation issues arising from batch effects. CARD (Ma & Zhou, 2022) introduced spatial effects into cell type proportions and performs deconvolution through an autoregressive-based deconvolution method. SpatialDWLS (Dong & Yuan, 2021) used reference features extracted from scRNA-seq to fit a damped weighted least squares model for inferring cell type composition. In addition, SPOTlight (Elosua-Bayes et al., 2021) and Cell2location (Kleshchevnikov et al., 2022) are also recently developed methods. Y. Li and Luo (2024) proposed a reference-based method for cell type deconvolution by combining spatial relations in ST data via Graph Convolutional Networks (GCN). However, a key assumption of reference-based methods is that cell type-specific gene expression remains constant, with only the cell type composition varying (Teschendorff et al., 2017). Therefore, reference-based methods are limited by several factors, including the availability and quality of reference datasets, as well as platform and batch effects (H. Li et al., 2023). Additionally, reference-based methods are limited by the cell types available in the reference, which may not capture all the cell types present in the target tissue or disease state. As a result, there is a growing need for reference-free methods that do not rely on external scRNA-seq data but instead leverage inherent gene expression patterns and prior knowledge to infer cell type composition.

Recent advancements in reference-free approaches have made progress in addressing these limitations by directly analyzing the spatial transcriptomic data without the need for scRNA-seq references. These methods rely on prior knowledge of marker genes to infer cell type proportions, offering a more flexible and scalable solution (H. Li et al., 2023). STdeconvolve (Miller et al., 2022), as the earliest reference-free deconvolution method, used Latent Dirichlet Allocation (LDA) to simultaneously infer gene expression profiles and cell type proportions for each ST spot. However, it relies only on gene co-expression patterns for cell type decomposition, without utilizing marker gene information, but the external references are still required later to match the estimated cell types. Thus, it cannot be considered a truly reference-free method. RETROFIT (Singh et al., 2023) was developed to decompose the gene expression matrix into latent components and matches cell types using a marker gene list, which is effective. However, it does not take spatial information into account and is unable to classify cell types that lack defined marker genes. Celloscope (Geras et al., 2023) constructed a probabilistic Bayesian framework and incorporates prior qualitative information from marker genes, providing a more detailed statistical description of cell type proportion and gene expression features. SpatialDeX (X. Liu et al., 2025) uses a regression-based approach to estimate cell-type proportions in spatial transcriptomics and performs pan-cancer clustering analysis. However, both Celloscope and SpatialDeX did not account for spatial information. For these reference-free methods, recognizing the latent spatial patterns

and tackling the challenges posed by tissue heterogeneity are still open research questions that need further exploration.

In this study, we introduced SvdRFCTD, a reference-free spatial transcriptomics deconvolution method, which estimates the cell type proportions at each spot on the tissue. To fully capture the heterogeneity in the ST data, SvdRFCTD employs a Bayesian hierarchical negative binomial model with spatial effects incorporated in both the mean and dispersion of the gene expression, which is used to explicitly model the generative mechanism of cell type proportions. The estimation of cell type proportion is achieved through the Markov Chain Monte Carlo (MCMC) method with adaptive Metropolis algorithm. Compared to other reference-free methods, SvdRFCTD not only effectively identifies the underlying spatial patterns of gene expression and provides a robust explanation for heterogeneity, but also accurately characterizes the spatial dependency relationships of gene expression. We illustrate the validity of SvdRFCTD through extensive simulations and applications. For illustration, we apply the method to the anterior section of the mouse brain and human pancreatic ductal adenocarcinomas, and find that SvdRFCTD identifies highly clustered cell types in specific regions, and effectively distinguishes different cell subtypes. And it additionally uncovers numerous spatially co-localized cell types, shedding light on the relationships between cell types.

2. Methods

2.1. Model specification

For cell type deconvolution of the spatial transcriptomics, there are two ways of understanding the spatial effects of RNA transcripts in tissues. One perspective attributes the spatial correlation of neighbouring spots to the spatial correlation of cell type proportions, while another suggests it arises from the spillover effect of gene expression across spots. Existing deconvolution methods have predominantly adopted the first perspective, while no studies have explicitly modelled the second. Here we adopt the second perspective by decomposing gene expression in both mean and scale dimensions to include the spatial effect as a spillover factor in the model.

Let Y_{ij} be the RNA-seq read count for gene $j \in \{1, \ldots, J\}$ and spot $i \in \{1, \ldots, N\}$, where J is the total number of genes and N is the number of spots. Assuming Y_{ij} follows a negative binomial (NB) distribution, then we model Y_{ij} with the following hierarchical model:

$$Y_{ij} \sim f_{\mathrm{NB}}(d_i e_{ij}, \rho_{ij}), \tag{1}$$

where d_i is the total number of cells in spot i, and $d_i e_{ij}$ is the expectation of the distribution, and ρ_{ij} is the overdispersion parameter. Let $E_{ij} = d_i e_{ij}$. The probability mass function of Equation (1) is

$$f(y_{ij}; E_{ij}, \rho_{ij}) = \frac{\Gamma\left(y_{ij} + \rho_{ij}^{-1}\right)}{\Gamma\left(\rho_{ij}^{-1}\right)\Gamma(y_{ij}+1)} \left(\frac{1}{1+E_{ij}\rho_{ij}}\right)^{\rho_{ij}^{-1}} \left(\frac{E_{ij}}{\rho_{ij}^{-1}+E_{ij}}\right)^{y_{ij}}. \tag{2}$$

Then the conditional mean and variance of Y_{ij} are given by

$$\begin{aligned} \mathrm{E}\left(Y_{ij} \mid e_{ij}, \rho_{ij}\right) &= d_i e_{ij} = E_{ij}, \\ \mathrm{Var}\left(Y_{ij} \mid e_{ij}, \rho_{ij}\right) &= E_{ij} + \rho_{ij} E_{ij}^2. \end{aligned} \tag{3}$$

Many studies (Allen et al., 2021; Ma & Zhou, 2022) have shown that spatial correlations exist within tissue domains in spatial transcriptomics. Specifically, both the mean and the dispersion of gene expression within a given tissue may exhibit spatial co-localization patterns to some extent.

Suppose there are a total of K cell types in the tissue, and Y_{ij} is a mixture of K cell type expression profiles. We assume that e_{ij} is a random variable defined by

$$\log e_{ij} = \log\left(\sum_{k=1}^{K} \beta_{ik}\mu_{kj}\right) + \phi_{1i}, \tag{4}$$

where β_{ik} is the cell type proportion for cell type $k \in \{1, \ldots, K\}$ in spot i, μ_{kj} represents the mean gene expression profile for cell type k, and ϕ_{1i} is a spot-level spatial random effect for the mean expression.

To accommodate spatial variation in overdispersion, we assume the following functional form for ρ_{ij}

$$\log\left(\rho_{ij}\right) = \delta_j + \phi_{2i}, \tag{5}$$

where δ_j is a gene-level random effect, which can account for some of the natural variability such as platform effects. ϕ_{2i} is a spot-level spatial random effect for the overdispersion.

The spatial random effects $\boldsymbol{\phi}_i = (\phi_{1i}, \phi_{2i})^\top$ capture potential regional factors that might similarly affect both the mean and dispersion. For example, areas with high gene expression may exhibit increased overdispersion, potentially due to extreme counts. To model this relationship and promote spatial smoothing while sharing information between neighbouring spots within tissues, we adopt a bivariate intrinsic conditional autoregressive (BICAR) prior (Mardia, 1988) distribution for $\boldsymbol{\phi}_i = (\phi_{1i}, \phi_{2i})^\top$.

$$\boldsymbol{\phi}_i \mid \boldsymbol{\phi}_{(-i)}, \Sigma \sim \mathrm{N}_2\left(\frac{1}{m_i}\sum_{l\in\partial_i}\boldsymbol{\phi}_l, \frac{1}{m_i}\Sigma\right), \tag{6}$$

where m_i represents the number of neighbours of spot i, ∂_i is the set of neighbours for region i and $\boldsymbol{\Sigma}$ is a 2×2 variance-covariance matrix of $\boldsymbol{\phi}_i$ conditional on the remaining spatial random effects, $\boldsymbol{\phi}_{(-i)}$. Additionally, a sum-to-zero constraint must be applied to $\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n\}$ to ensure the model is identifiable (Mutiso et al., 2022).

Following Brook's lemma (Banerjee et al., 2014), the joint prior distribution for the $2N \times 1$ spatial random effects matrix $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1^\top, \ldots, \boldsymbol{\phi}_N^\top)^\top$ is proportional to the multivariate normal distribution with formula

$$\boldsymbol{\Phi} \,\Big|\, \boldsymbol{\Sigma} \propto \exp\left\{-\frac{1}{2}\boldsymbol{\Phi}^\top\left[\boldsymbol{Q}\otimes\boldsymbol{\Sigma}^{-1}\right]\boldsymbol{\Phi}\right\}, \tag{7}$$

where $\boldsymbol{Q} = \boldsymbol{M} - \boldsymbol{A}; \boldsymbol{M} = \mathrm{diag}(m_1, m_2, \ldots, m_n)$ and $\boldsymbol{A}$ is $N \times N$ adjacency matrix with $a_{ii} = 0, a_{ij} = 1$ if regions i and j are neighbours and $a_{ij} = 0$ otherwise. Noting that $\boldsymbol{M} - \boldsymbol{A}$ is singular, the joint prior distribution in Equation (7) is improper, but the posterior of $\boldsymbol{\Phi}$ is proper (Mutiso et al., 2022).

To facilitate the update of the binary spatial random effects, we can decompose Equation (7) into two univariate conditional priors $\phi_{hi} (h = 1, 2)$ (Mutiso et al., 2024; Neelon

et al., 2023). Using the properties of conditional multivariate normal theory, the conditional prior for the spatial effect, $\phi_{1i}, i = 1, \ldots, N$, is given by

$$p\left(\phi_{1i} \mid \phi_{2i}, \tau_{1i|2i}\right) \propto \exp\left[-\frac{\tau_{1i|2i}}{2}\left(\phi_{1i} - \mu_{1i|2i}\right)^2\right], \tag{8}$$

where

$$\tau_{1i|2i} = \left[\sigma_{\phi_1}^2\left(1-\rho^2\right)\right]^{-1} m_i,$$
$$\mu_{1i|2i} = \rho\frac{\sigma_{\phi_1}}{\sigma_{\phi_2}}\left(\phi_{2i} - \bar{\phi}_{2i}\right),$$

ρ is the correlation between ϕ_{1i} and ϕ_{2i}, σ_{ϕ_1} and σ_{ϕ_2} are standard deviations for ϕ_{1i} and ϕ_{2i} respectively, and $\bar{\phi}_{2i} = \frac{1}{m_i}\sum_{l\in\partial_i}\phi_{2l}$ is the prior mean for ϕ_{2i}. A similar derivation can be made for the conditional prior of ϕ_{2i}.

Our primary objective is to estimate the proportions of specific cell types across all spots, indicated by B, which is an N by K matrix. Each element β_{ik} represents the proportion of cell type k in spot i, with values ranging from 0 to 1. A row in matrix B, labelled as $B_{i:} = [\beta_{i1}, \ldots, \beta_{iK}]$, illustrates the cell type composition of spot i, and it's clear that the entries of each row sum up to 1.

To elucidate the underlying gene expression patterns in different cell types, we developed a multilayer generation process for the gene expression. We assume the probability of the existence of cell type k in spot i follows a beta distribution (Geras et al., 2023; Yang et al., 2024)

$$\pi_{ik} \sim \text{Beta}\left(\frac{\alpha}{K}, 1\right), \tag{9}$$

where α is a hyperparameter which can represent the average number of cell types present in a spot, and K is the total number of cell types in the tissue.

Let Z_{ik} denote whether the cell type k is present in the spot i, and the distribution of Z_{ik} is as follows

$$Z_{ik} \sim \text{Bernoulli}\left(1, \pi_{ik}\right). \tag{10}$$

Let θ_{ik} denote the unnormalized abundance of type k in spot i, and its distribution depends on Z_{ik}.

$$\begin{aligned}\theta_{ik}|Z_{ik} = 1 &\sim \text{gamma}(a, b),\\ \theta_{ik}|Z_{ik} = 0 &\sim \text{gamma}(a_0, b_0).\end{aligned} \tag{11}$$

The choice of a, b, a_0, b_0 will result in a larger sampling value for $\theta_{ik}|Z_{ik} = 1$ than for $\theta_{ik}|Z_{ik} = 0$.

For spot i, the abundance of cell type k can represent the proportion of this type in spot i. Thus, we can calculate the cell type proportion β_{ik} from θ_{ik}

$$\beta_{ik} = \frac{\theta_{ik}}{\sum_{k=1}^{K}\theta_{ik}}. \tag{12}$$

Moreover, to avoid estimation issues caused by an excessive number of parameters, we split μ_{kj} into different components as

$$\mu_{kj} = \mu_0 + M_{kj}\lambda_{kj}, \tag{13}$$

where M_{kj} is the marker gene indicator, with $M_{kj} = 1$ if gene j is a marker gene for cell type k and $M_{kj} = 0$ otherwise. $\lambda_{kj} > 0$ denotes the average expression level of gene j in cell type k. Here we adopted flat priors for μ_0 and λ_{kj}.

2.2. Parameter inference

For posterior computation, we implement the MCMC sampling with adaptive Metropolis (AM) algorithm (Haario et al., 2001). The AM algorithm is an enhanced version of the Metropolis-Hastings algorithm that dynamically adjusts the proposal distribution's covariance matrix during sampling to improve efficiency, particularly in high-dimensional or complex target distributions. Suppose we have multivariate random variables $x_1, \ldots, x_n$, and the full conditional distribution of x_i is given as $P(x_i \mid \boldsymbol{x}_{-i}) := P(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. In a hierarchical model, the full conditional distribution of a node/variable depends on the distribution of parent, child, and co-parent nodes/variables, which are also called Markov blanket of x_i, denoted as $\mathrm{MB}(x_i)$. Thus, $P(x_i \mid \boldsymbol{x}_{-i}) = P(x_i \mid \mathrm{MB}(x_i))$, which denotes that the conditional distribution of x_i given the values of all other variables equals the conditional distribution given the values of the variables from its Markov blanket. This step is crucial in facilitating the simplification of the derivation process.

Then we can perform iterative sampling procedure to update $x_i, i = 1, \ldots, N$ one by one, up until convergence. For each iteration t, we determine whether the formula $P(x_i \mid \mathrm{MB}(x_i))$ is in closed form and accordingly decide the approach for updating. If it is in a closed form, we can sample a new x_i^* directly from $P(x_i \mid \mathrm{MB}(x_i^{(t-1)}))$. And if the formula is too complex to allow for direct sampling, we can employ the Metropolis-Hastings (MH) accept-reject method. Assume there is a function g proportional to $P(x_i \mid \mathrm{MB}(x_i))$

$$P\left(x_i \mid \mathrm{MB}(x_i)\right) \propto g\left(x_i\right). \tag{14}$$

A candidate value x_i^* is sampled from a predefined proposal distribution $q(\cdot \mid x_i^{(t-1)})$, and then either accepted with probability given by

$$r = \min\left(1, \frac{g\left(x^*\right) q\left(x^{(t-1)} \mid x^*\right)}{g\left(x^{(t-1)}\right) q\left(x^* \mid x^{(t-1)}\right)}\right), \tag{15}$$

and $x_i^{(t)} \leftarrow x_i^*$, or the previous value is held, $x_i^{(t)} \leftarrow x_i^{(t-1)}$. And if we utilize a symmetric proposal distribution, the acceptance probability above can be simplified as

$$r = \min\left(1, \frac{g\left(x^*\right)}{g\left(x^{(t-1)}\right)}\right). \tag{16}$$

After updating x_i, the new value is immediately utilized, allowing us to sequentially sample other variables.

The outline of the MCMC algorithm is provided below.

(1) Updating $\pi_{ik} \mid Z_{ik} \sim \text{Beta}(\pi_{ik}|\, \frac{\alpha}{K} + Z_{ik}, 2 - Z_{ik})$, the posterior is derived as follows

$$\begin{aligned} P(\pi_{ik} \mid Z_{ik}) &\propto P(Z_{ik} \mid \pi_{ik}) P(\pi_{ik}) \\ &\propto \frac{\Gamma(\frac{\alpha}{K}+1)}{\Gamma(\frac{\alpha}{K})\Gamma(1)} \pi_{ik}^{\frac{\alpha}{K}-1} \pi_{ik}^{Z_{ik}} (1-\pi_{ik})^{1-Z_{ik}}] \\ &\propto \pi_{ik}^{\frac{\alpha}{K}+Z_{ik}-1} (1-\pi_{ik})^{1-Z_{ik}}. \end{aligned} \tag{17}$$

(2) Update Z_{ik}. As Z_{ik} is a discrete, binary random variable, it suffices to consider its two possible values, 0 and 1.

$$\begin{aligned} P(Z_{ik} \mid \theta_{ik}, \pi_{ik}) &\propto P(Z_{ik} \mid \pi_{ik})\, P(\theta_{ik} \mid Z_{ik}) \\ &\propto \pi_{ik}^{Z_{ik}} (1-\pi_{ik})^{1-Z_{ik}} \left(\theta_{ik}^{a-1} \mathrm{e}^{-b\theta_{ik}}\right)^{Z_{ik}} \left(\theta_{ik}^{a_0-1} \mathrm{e}^{-b_0\theta_{ik}}\right)^{1-Z_{ik}} \\ &\propto \left(\pi_{ik} \theta_{ik}^{a-1} \mathrm{e}^{-b\theta_{ik}}\right)^{Z_{ik}} \left[(1-\pi_{ik}) \theta_{ik}^{a_0-1} \mathrm{e}^{-b_0\theta_{ik}}\right]^{1-Z_{ik}}, \end{aligned}$$

i.e.,

$$\begin{aligned} P(Z_{ik} = 1 \mid \theta_{ik}, \pi_{ik}) &= \frac{\pi_{ik} \theta_{ik}^{a-1} \mathrm{e}^{-b\theta_{ik}}}{\pi_{ik} \theta_{ik}^{a-1} \mathrm{e}^{-b\theta_{ik}} + (1-\pi_{ik}) \theta_{ik}^{a_0-1} \mathrm{e}^{-b_0\theta_{ik}}}, \\ P(Z_{ik} = 0 \mid \theta_{ik}, \pi_{ik}) &= \frac{(1-\pi_{ik}) \theta_{ik}^{a_0-1} \mathrm{e}^{-b_0\theta_{ik}}}{\pi_{ik} \theta_{ik}^{a-1} \mathrm{e}^{-b\theta_{ik}} + (1-\pi_{ik}) \theta_{ik}^{a_0-1} \mathrm{e}^{-b_0\theta_{ik}}}. \end{aligned} \tag{18}$$

(3) Update θ_{ik}. For a given spot i and cell type k, the target distribution for unnormalized cell type abundance θ_{ik} is given as

$$\begin{aligned} &P(\theta_{ik} \mid \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\delta}, \boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2) \\ &\propto \prod_{j=1}^{J} P\left(Y_{ij} \mid \theta_{ik}, \boldsymbol{\mu}, \boldsymbol{\delta}, \boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2\right) \prod_{k=1}^{K} P(\theta_{ik} \mid Z_{ik}) \\ &\propto \prod_{j=1}^{J} \frac{\Gamma\left(y_{ij} + \rho_{ij}^{-1}\right)}{\Gamma\left(\rho_{ij}^{-1}\right) \Gamma(y_{ij}+1)} \left(\frac{1}{1+E_{ij}\rho_{ij}}\right)^{\rho_{ij}^{-1}} \left(\frac{E_{ij}}{\rho_{ij}^{-1}+E_{ij}}\right)^{y_{ij}} \\ &\quad \times \prod_{k=1}^{K} \left(\theta_{ik}^{a-1} \mathrm{e}^{-b\theta_{ik}}\right)^{Z_{ik}} \left(\theta_{ik}^{a_0-1} \mathrm{e}^{-b_0\theta_{ik}}\right)^{1-Z_{ik}}. \end{aligned} \tag{19}$$

Since $\theta_{ik} > 0$, we choose the truncated normal distribution $\text{TN}(\mu, \sigma)$ for the proposal distribution of θ_{ik} to calculate the acceptance probability in Equation (15), as it allows for effective control over the step size and is well-suited for proposing values for non-negative variables. Then, we obtain the cell type proportion $\beta_{ik} = \theta_{ik} / \sum_{k=1}^{K} \theta_{ik}$.

(4) Update μ_{kj}, which equals to $\mu_0 + M_{kj}\lambda_{kj}$, where M_{kj} is known. The prior of μ_0 is average expression of all non-marker genes, and the prior of λ_{ik} is average expression of all marker genes. Since $\mu_0 > 0$ and $\lambda_{kj} > 0$, the truncated normal proposal distribution is

also used here to calculate the acceptance probability.

$$\begin{aligned} P\left(\mu_0 \mid \boldsymbol{\lambda}, \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2\right) &\propto \prod_{i=1}^{N} f_{\mathrm{NB}}(d_i e_{ij}, \rho_{ij}) P(\mu_0), \\ P\left(\lambda_{kj} \mid \mu_0, \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2\right) &\propto \prod_{i=1}^{N} f_{\mathrm{NB}}(d_i e_{ij}, \rho_{ij}) P(\lambda_{kj}). \end{aligned} \tag{20}$$

(5) Update δ_j. Assuming a normal prior for δ_j, then the proposal distribution is given as

$$P\left(\delta_j \mid \boldsymbol{\lambda}, \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2\right) \propto \prod_{i=1}^{N} f_{\mathrm{NB}}(d_i e_{ij}, \rho_{ij}) P(\delta_j). \tag{21}$$

(6) Update ϕ_{1i}. Given the conditional prior of $\phi_{1i} \mid \phi_{2i}$ by Equation (8), the full conditional of ϕ_{1i} is

$$\begin{aligned} P(\phi_{1i} \mid \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{\mu}, \boldsymbol{\Phi}_2) &\propto \prod_{j=1}^{J} f_{\mathrm{NB}}(d_i e_{ij}, \rho_{ij}) P\left(\phi_{1i} \mid \phi_{2i}\right) \\ &\propto \prod_{j=1}^{J} f_{\mathrm{NB}}(d_i e_{ij}, \rho_{ij}) \exp\left[-\frac{\tau_{1i|2i}}{2}\left(\phi_{1i} - \mu_{1i|2i}\right)^2\right], \end{aligned} \tag{22}$$

and we use a random walk MH step with a symmetric univariate t proposal density centred at the previous ϕ_{1i}. The acceptance ratio in Equation (16) is

$$r_{\phi_{1i}} = \frac{f_{\mathrm{NB}}(d_i e_{ij}^{(p)}, \rho_{ij}) \exp\left[-\frac{\tau_{1i|2i}}{2}\left(\phi_{1i}^{(p)} - \mu_{1i|2i}\right)^2\right]}{f_{\mathrm{NB}}(d_i e_{ij}^{(t)}, \rho_{ij}) \exp\left[-\frac{\tau_{1i|2i}}{2}\left(\phi_{1i}^{(t)} - \mu_{1i|2i}\right)^2\right]}, \tag{23}$$

where $\phi_{1i}^{(p)}$ and $\phi_{1i}^{(t)}$ are proposal and current values of ϕ_{1i} at current iteration t.

(7) Update ϕ_{2i}.

$$\begin{aligned} P(\phi_{2i} \mid \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{\mu}, \boldsymbol{\Phi}_1) &\propto \prod_{j=1}^{J} f_{\mathrm{NB}}(d_i e_{ij}, \rho_{ij}) P\left(\phi_{2i} \mid \phi_{1i}\right) \\ &\propto \prod_{j=1}^{J} f_{\mathrm{NB}}(d_i e_{ij}, \rho_{ij}) \exp\left[-\frac{\tau_{2i|1i}}{2}\left(\phi_{2i} - \mu_{2i|1i}\right)^2\right], \end{aligned} \tag{24}$$

and the acceptance ratio is similar to Equation (23).

(8) Update Σ. Assuming an inverse Wishart distribution $\mathrm{IW}(\nu_0, \boldsymbol{S}_0)$ for the prior of Σ, we update the random effects covariance matrix, $\boldsymbol{\Sigma}$, from a conjugate IW distribution given by

$$\boldsymbol{\Sigma} \mid \boldsymbol{\Phi} \sim \mathrm{IW}\left(\nu_0 + N - 1, \boldsymbol{S}_0 + \boldsymbol{S}_{\boldsymbol{\Phi}^*}\right), \tag{25}$$

where $\boldsymbol{S}_{\boldsymbol{\Phi}^*} = \boldsymbol{\Phi}^{*\top} \boldsymbol{Q} \boldsymbol{\Phi}^*$ and $\boldsymbol{\Phi}^* = [\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2]$ is the $N \times 2$ random effects matrix centred at its mean.

Denote $\boldsymbol{\Omega} = \{\boldsymbol{\pi}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\mu}_0, \boldsymbol{\lambda}, \boldsymbol{\delta}, \boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \boldsymbol{\Sigma}\}$. The MCMC procedure is a combination of the Gibbs Sampler and the Metropolis-Hastings algorithm, and we summarize it in Algorithm 1.

Algorithm 1 MCMC procedure for hierarchical model parameters

Input: Count matrix, Y; number of cells, $\boldsymbol{d}$; marker gene indicator matrix, M; initial values of parameters $\boldsymbol{\Omega}^{(0)}$; number of iterations, T
Output: $\boldsymbol{\Omega}^{(1)}, \ldots, \boldsymbol{\Omega}^{(T)}$

for $t = 1, \ldots, T$ **do**
  sample $\boldsymbol{\pi}^{(t)}$ from $P(\boldsymbol{\pi} \mid \mathrm{MB}(\boldsymbol{\pi}^{(t-1)}))$ showed in Equation (17)
  sample $\boldsymbol{Z}^{(t)}$ from $P(\boldsymbol{Z} \mid \mathrm{MB}(\boldsymbol{Z}^{(t-1)}))$ showed in Equation (18)
  update $\boldsymbol{\theta}^{(t)} \longleftarrow \mathrm{MH}(\boldsymbol{\theta}^{(t-1)})$ based on Equation (19)
  update $\boldsymbol{\mu}_0^{(t)} \longleftarrow \mathrm{MH}(\boldsymbol{\mu}_0^{(t-1)})$ based on Equation (20)
  update $\boldsymbol{\lambda}^{(t)} \longleftarrow \mathrm{MH}(\boldsymbol{\lambda}^{(t-1)})$ based on Equation (21)
  update $\boldsymbol{\delta}^{(t)} \longleftarrow \mathrm{MH}(\boldsymbol{\delta}^{(t-1)})$ based on Equation (22)
  update $\boldsymbol{\phi}_1^{(t)} \longleftarrow \mathrm{MH}(\boldsymbol{\phi}_1^{(t-1)})$ based on Equation (23)
  update $\boldsymbol{\phi}_2^{(t)} \longleftarrow \mathrm{MH}(\boldsymbol{\phi}_2^{(t-1)})$ based on Equation (24)
  update $\boldsymbol{\Sigma}^{(t)} \longleftarrow \mathrm{MH}(\boldsymbol{\Sigma}^{(t-1)})$ based on Equation (25)
end for
return $\boldsymbol{\Omega}^{(1)}, \ldots, \boldsymbol{\Omega}^{(T)}$

In Algorithm 1, the number of cells, $\boldsymbol{d} = \{d_1, \ldots, d_N\}$, can be provided as estimates from external methods or inferred directly within the algorithm. Details are presented in Section 2.4. The $\mathrm{MH}(\cdot)$ step can be performed following Algorithm 2.

Algorithm 2 MH step in Algorithm 1

Input: Current state of parameter, $x^{(t-1)}$
Output: New state of parameter, $x^{(t)}$

Draw proposal sample $x^* \sim q(\cdot \mid x^{(t-1)})$
Evaluate acceptance probability r
Generate $u \sim U(0, 1)$
if $r \leqslant u$ **then**
  $x^{(j)} \leftarrow x^*$
else
  $x^{(j)} \leftarrow x^{(j-1)}$
end if
return $x^{(t)}$

The default parameter settings are listed in Table 1.

2.3. Proposal distribution and adaptive step size

Let $\Phi(x)$ denote the cumulative distribution function of the standard normal distribution $N(0, 1)$, evaluated at x. Additionally, let $q(y \mid x)$ represent the proposal distribution, which

Table 1. Default values of SvdRFCTD's parameters.

Parameter	Default value
(a, b)	(10, 1)
(a, b)	(0.1, 1)
α	$> \frac{K}{K+1}$
Step-size for μ_0	0.05
Step-size for θ	0.1
Step-size for δ	0.1
Step-size for ϕ_1 and ϕ_2	0.17
Burn-in	40,000
Number of iterations	50,000
Thinning	10

specifies the conditional probability of proposing a new state y given that the previous state was x. During the iteration process of the Metropolis-Hastings sampling, we choose the truncated normal distribution $\mathrm{TN}(\cdot \mid \sigma)$ and symmetric t distribution as the proposal distributions. The truncated normal proposal distribution is given as

$$q(y \mid x) = \frac{1}{C} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-x)^2}{2\sigma^2}\right), \quad y > 0, \tag{26}$$

where $C = \Phi(\frac{x}{\sigma})$ is a normalizing constant. By Equation (15), the acceptance ratio is

$$\frac{q(x \mid y)}{q(y \mid x)} = \frac{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-y)^2}{2\sigma^2}\right) \frac{1}{C_1}}{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-x)^2}{2\sigma^2}\right) \frac{1}{C_2}} = \frac{C_2}{C_1} = \frac{\Phi\left(\frac{x}{\sigma}\right)}{\Phi\left(\frac{y}{\sigma}\right)}. \tag{27}$$

The variance parameter σ controls the sampler's step size and impacts convergence speed. Its selection is crucial, as excessively large or small step sizes degrade inference performance. To enhance efficiency, we assign different step sizes to different variables and adjust them during an initial burn-in phase to maintain an optimal acceptance ratio of 0.23 (Brooks et al., 2011; Graves, 2011). Starting with an arbitrary step size, we update it after a fixed number of iterations, depending on the dataset, based on the acceptance ratio. If the ratio falls below the target, we reduce the step size as $\sigma \leftarrow (1-\epsilon)\sigma$; otherwise, we increase it as $\sigma \leftarrow (1+\epsilon)\sigma$, where ϵ controls the adjustment magnitude. This adaptation is applied only during burn-in.

2.4. Cell counting

Recording the results of simulation in Section 3.1, the best choice of d_i in Equation (1) is the number of cells. There are a few approaches commonly used to estimate cell numbers in each spatial spot.

(1) Infer the cell numbers by deconvolution methods of gene expression, such as Cell2location (Kleshchevnikov et al., 2022), SPOTlight (Elosua-Bayes et al., 2021).
(2) Integrate with single-cell data and mapping scores to infer the number of cells per spot, by methods like STalign (Clifton et al., 2023), Tangram (Biancalani et al., 2021), CytoSPACE (Shannon et al., 2003).
(3) Use histological staining or nuclear staining images of the tissue to preform image-based methods, with tools like CellProfiler (Carpenter et al., 2006) or QuPath (Bankhead et al., 2017).

In this article, if the number of cells per spot is not given by the data analyzed, it can either be supplied as estimates derived from external methods or estimated directly within the algorithm. Here, a MH step with TN distribution as proposal is performed for d_i, with the prior distribution set to a constant value.

$$P\left(d_i \mid \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\delta}, \boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2\right) \propto \prod_{j=1}^{J} P\left(Y_{ij} \mid \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\delta}, \boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2\right) P\left(d_i\right). \tag{28}$$

2.5. Performance evaluation metrics

(1) Pearson Correlation Coefficient (PCC). For each cell type k, the PCC value is calculated as follows

$$\mathrm{PCC}_k = \frac{\mathrm{E}\left[\left(\widetilde{\boldsymbol{x}}_k - \tilde{u}_k\right)\left(\boldsymbol{x}_k - u_k\right)\right]}{\tilde{\sigma}_k \sigma_k}, \tag{29}$$

where $\boldsymbol{x}_k$ and $\widetilde{\boldsymbol{x}}_k$ represent the ground truth and predicted proportions of cell type k across N spots, respectively. Similarly, u_k and $\tilde{u}_k$ are their corresponding mean proportions, while σ_k and $\tilde{\sigma}_k$ denote the standard deviations. A higher PCC for cell type k indicates better prediction accuracy.

(2) Root Mean Squared Error (RMSE). We can compute both per-spot RMSE and overall RMSE for estimation of cell type proportion, as defined in Equations (2) and (3), respectively.

$$\mathrm{RMSE}_i = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\tilde{x}_{ik} - x_{ik}\right)^2}, \tag{30a}$$

$$\mathrm{RMSE}_{\mathrm{overall}} = \sqrt{\frac{1}{N \times K}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(\tilde{x}_{ik} - x_{ik}\right)^2}, \tag{30b}$$

where x_{ik} and $\tilde{x}_{ik}$ are the cell type proportion of cell type K in spot i in the ground truth and the predicted result, respectively. A lower RMSE value indicates better prediction accuracy.

(3) Weighted F1 score (F1 score). For each spot, the dominant cell type can be inferred according to the cell type proportion estimated, and F1 score is used to measure the accuracy compared to the true dominant cell type. For each cell type k, calculate F1 score as

$$\begin{aligned} \mathrm{F1}_k &= \frac{2 \cdot \mathrm{Precision}_k \cdot \mathrm{Recall}_k}{\mathrm{Precision}_k + \mathrm{Recall}_k}, \\ \mathrm{Precision}_k &= \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k}, \\ \mathrm{Recall}_k &= \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k}, \end{aligned} \tag{31}$$

where TP_k, FP_k and FN_k are true positives, false positives and false negatives for category k, respectively. And then we can obtain a weighted F1 score

$$F1_{\text{weighted}} = \sum_{k=1}^{K} w_k \cdot F1_k, \tag{32}$$

where $w_k = \frac{N_k}{N_{\text{total}}}$, with N_k being the number of samples in cell type k and N_{total} being the total number of samples.

(4) Moran's I coefficient. The coefficient is used to quantify spatial autocorrelation based on both feature locations and feature values simultaneously. And it indicates whether the variables are spatially distributed in a random pattern or whether they are significantly clustered (positive correlation) or discrete (negative correlation). For a given cell type k, the formula of Moran's I coefficient is given as

$$\begin{aligned} I_k &= \frac{N}{S_0} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \left(x_{ik} - \bar{x}_k\right) \left(x_{jk} - \bar{x}_k\right)}{\sum_{i=1}^{N} \left(x_{ik} - \bar{x}_k\right)^2}, \\ S_0 &= \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}, \end{aligned} \tag{33}$$

where N represents the total number of spatial spots, w_{ij} is the spatial weight between spot i and j, x_{ik} denotes the expression level of cell type k in spot i, and $\bar{x}_k$ is the mean expression level of cell type k across all spots. The coefficient $I_k > 0$ indicates positive spatial autocorrelation, i.e. similar expression values are clustered together. $I_k \approx 0$ indicates no spatial autocorrelation, i.e. random pattern. $I_k < 0$ indicates negative spatial autocorrelation (dispersed pattern), i.e. dissimilar expression values are close together. For a given cell type, Moran's I is computed by considering the expression levels of that cell type across all spatial spots, assessing whether its expression exhibits clustering, randomness, or dispersion.

3. Simulation study

3.1. Explore the appropriate model form in simulation

In order to verify the validity of SvdRFCTD, we conducted a series of simulations, for which we knew the ground truth about the underlying cell type composition. In this article, we used a publicly available single cell RNA-seq (scRNA) data set in mouse kidneys (J. Liu et al., 2023) to construct simulated ST data, which was profiled using the Vizgen Multiplexed Error-Robust Fluorescence in situ Hybridization(MERFISH) platform (Chen et al., 2015). The mouse kidney dataset comprises the expression profiles of 304 genes across 126,241 cells, categorized into eight cell types. Figure 1(a) shows the spatial visualization of single-cell mouse kidney data, with cell types annotated. The marker genes of each cell types are collected from the existing literature (Miao et al., 2021) and the CellMarker 2.0 database (Hu et al., 2023), and will be used to SvdRFCTD and other reference-free deconvolution methods. The marker genes used in simulation study are listed in Supplementary Table S1. The data generated by pooling single cells from original ST data can serve as the gold standard for model evaluation. To simulate spatial transcriptomics (ST) data, we partitioned the single-cell data from the mouse kidney dataset into 2474 spatially contiguous squares. Within each

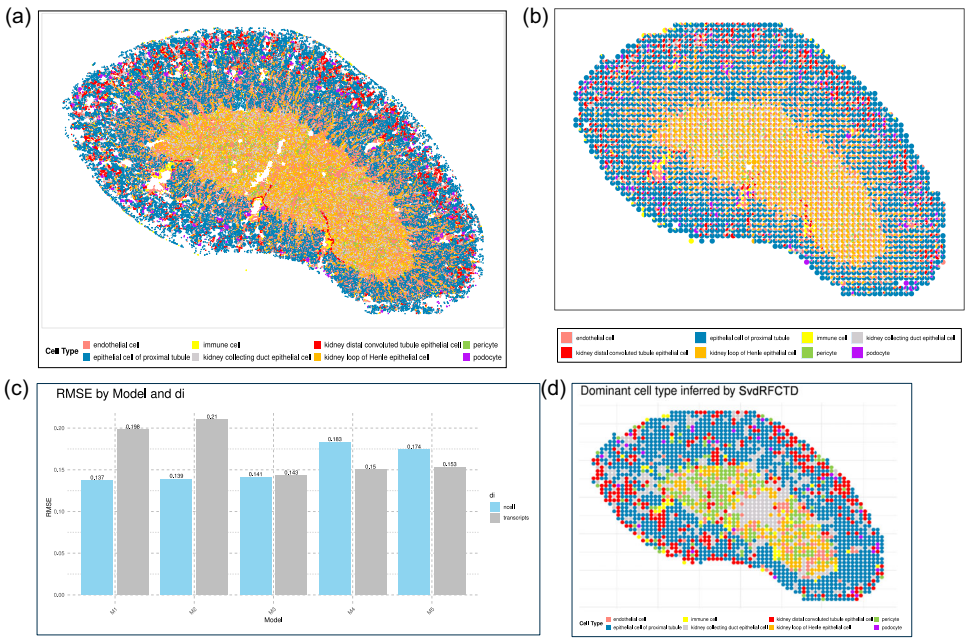

Figure 1. Spatial visualization, cell type proportions, model performance, and inferred dominant cell types on simulated mouse kidney ST data. (a) Spatial visualization of single-cell MERFISH data from the mouse kidney, annotated with cell types. (b) The scatter plot shows the proportion of cell types at each spot on simulated dataset. (c) RMSE for five model formulations, with d_i representing either the number of cells or the transcript count in spot i. (d) The dominant cell type on each spot inferred by SvdRFCTD. The colours corresponding to each cell type are consistent with those in (a) and (b).

square, we aggregated the gene expression of the cells to mimic the spots observed in ST data. Given that the cell type labels for all single cells within each spot are provided, we can get ground truth of cell type proportion on each spot accordingly (see Figure 1(b)). The performance of the models are evaluated by measuring various metrics between the true values and the estimation. In simulation study, we firstly utilize the simulated ST data to explore the appropriate form of the model, and then compare the performance of SvdRFCTD with other seven existing deconvolution methods.

For the first scenario, we consider several model settings, focussing on exploring the decomposition of the mean for gene expression, i.e. Equation (4). Specifically, we examine whether to perform deconvolution of gene expression mean on a linear scale or a log scale, and whether to include non-spatial spot-specific effect, gene-specific effects, or both. The mean model formulas considered includes the following types.

M1. Deconvolve the mean of gene expression on a log scale without non-spatial spot-specific and gene-specific effects. The formula is the same as Equation (4), given by

$$\log e_{ij} = \log\left(\sum_{k=1}^{K} \beta_{ik}\mu_{kj}\right) + \phi_{1i}. \tag{34}$$

M2. Deconvolve the mean of gene expression on a linear scale without non-spatial spot-specific and gene-specific effects. The formula is

$$e_{ij} = \sum_{k=1}^{K} \beta_{ik}\mu_{kj} + \phi_{1i}. \tag{35}$$

M3. Deconvolve on a log scale with non-spatial spot-specific effect ζ_i. The formula is given by

$$\log e_{ij} = \zeta_i + \log\left(\sum_{k=1}^{K} \beta_{ik}\mu_{kj}\right) + \phi_{1i}. \tag{36}$$

M4. Deconvolve on a log scale with gene-specific effect γ_j. The formula is given by

$$\log e_{ij} = \gamma_j + \log\left(\sum_{k=1}^{K} \beta_{ik}\mu_{kj}\right) + \phi_{1i}. \tag{37}$$

M5. Deconvolve on a log scale with non-spatial spot-specific ζ_i and gene-specific effects γ_j. The formula is given by

$$\log e_{ij} = \zeta_i + \gamma_j + \log\left(\sum_{k=1}^{K} \beta_{ik}\mu_{kj}\right) + \phi_{1i}. \tag{38}$$

For each model formula, we consider two settings for d_i, total transcript count or the number of cells in spot i. In the MCMC framework, for models M2–M5, we accordingly adjust the form of the likelihood and sample the newly introduced variables. For the non-spatial spot-specific effect ζ_i, assuming a normal prior, its proposal distribution is given as

$$P\left(\zeta_i \mid \boldsymbol{\lambda}, \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2\right) \propto \prod_{j=1}^{J} f_{\mathrm{NB}}(d_i e_{ij}, \rho_{ij}) P(\zeta_i). \tag{39}$$

Similarly, for gene-specific effect γ_j, assuming a normal prior, its proposal distribution is given as

$$P\left(\gamma_j \mid \boldsymbol{\lambda}, \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2\right) \propto \prod_{i=1}^{N} f_{\mathrm{NB}}(d_i e_{ij}, \rho_{ij}) P(\gamma_j). \tag{40}$$

We randomly select a region of the simulated ST data which contains 599 spots. By evaluating RMSE of the results of each formula (see Figure 1(c)), we find that when the model only includes spot-specific random effects(M1, M2 and M3), setting d_i to the number of cells in spot yields better estimation performance. In contrast, when the model incorporates gene-specific random effects with or without spot-specific random effects(M4 or M5), setting d_i to the transcript count of spot i provides better estimates. When d_i is set to the number of cells, the estimation performance of M1, M2, and M3 is comparable. However, the linear formulation in M2 requires additional constraints on parameters to ensure the mean of gene expression remains positive, while M3 demands more computational time. Overall, M1 in Equation (34) with d_i set as the number of cells, achieves the best performance. In other words, performing deconvolution on a linear scale or incorporating additional redundant random effects reduces the estimation accuracy.

3.2. SvdRFCTD outperformed existing methods in realistic settings

By using simulated ST data from all 2474 spots, we compare SvdRFCTD's performance to seven previously published deconvolution methods: Celloscope, RETROFIT, STdeconvolve, CARD-free, RCTD, SpatialDWLS and CARD. Among these, Celloscope, RETROFIT, STdeconvolve, and CARD-free are reference-free deconvolution methods, whereas the remaining methods, RCTD, SpatialDWLS, and CARD, are reference-based and rely on a reference scRNA-seq dataset to derive gene expression profiles for the analyzed cell types. We categorized STdeconvolve as a reference-free deconvolution method because it does not use marker genes or reference scRNA-seq data for cell type decomposition. However, it still requires reference scRNA-seq data or ground truth information to match cell types in its results, meaning it is not strictly a reference-free method. Reference-based methods often outperform reference-free methods because they leverage additional information from reference scRNA-seq datasets, which provide detailed gene expression profiles for specific cell types. To examine the performance of the above methods in a more realistic situation where an ideal reference dataset is unavailable, we collected single-cell RNA-seq data in mouse kidneys from the Tabula Muris Senis (TMS) cell atlas (Almanzar et al., 2020). The TMS scRNA-seq dataset contains 19,101 cells across 27 cell types. From this, we selected 13,761 cells corresponding to the cell types that matched the simulated dataset to construct the reference dataset. We also obtain marker genes from the dataset as the input for reference-free methods.

For each spot, the dominant cell type can be determined based on the cell type with the highest proportion (see Figure 1(d)). This can be compared with the spatial visualization shown in Figure 1(a,b), which represent the true single-cell MERFISH data with annotated cell types and the proportion of cell types at each spot in the simulated dataset, respectively. The comparison shows that the SvdRFCTD results are, to some extent, consistent with the true data presented in Figure 1(a,b), demonstrating the model's ability to accurately capture the spatial distribution and proportions of cell types.

Based on the results of simulation, SvdRFCTD demonstrated superior performance compared to most methods, ranking as the best among reference-free approaches and achieving results comparable to reference-based methods. The performance of each method was evaluated by RMSE, PCC and F1 score. We evaluated per-spot RMSE between the predicted and ground truth cell type proportions (see Figure 2(a)), where SvdRFCTD has the best performance in reference-free methods and demonstrates the smallest variability among all methods. RMSE across genes and spots is also calculated (see Figure 2(b)) and the results are generally consistent to per-spot RMSE. We calculate F1 score to measure the degree of alignment between the estimated and true values (see Figure 2(c)). In terms of F1 score, reference-based methods perform significantly better than reference-free methods, and SvdRFCTD also outperforms other reference-free methods. In terms of PCC, SvdRFCTD achieved the best performance among all methods (see Figure 2(d)), indicating a strong correlation between the estimated and true cell type proportions. Celloscope and RETROFIT also achieved favourable results in terms of PCC. In contrast, reference-based methods no longer hold an absolute advantage in assessing the similarity of cell type estimations. The performance differences between reference-based and reference-free deconvolution methods can be attributed to several factors. Reference-based methods utilize scRNA-seq data, offering a more comprehensive representation of cell types, which enhances the accuracy of cell-type proportion estimation, as reflected in higher F1 score and lower RMSE. In contrast, reference-free methods depend on predefined gene sets, which may not fully capture

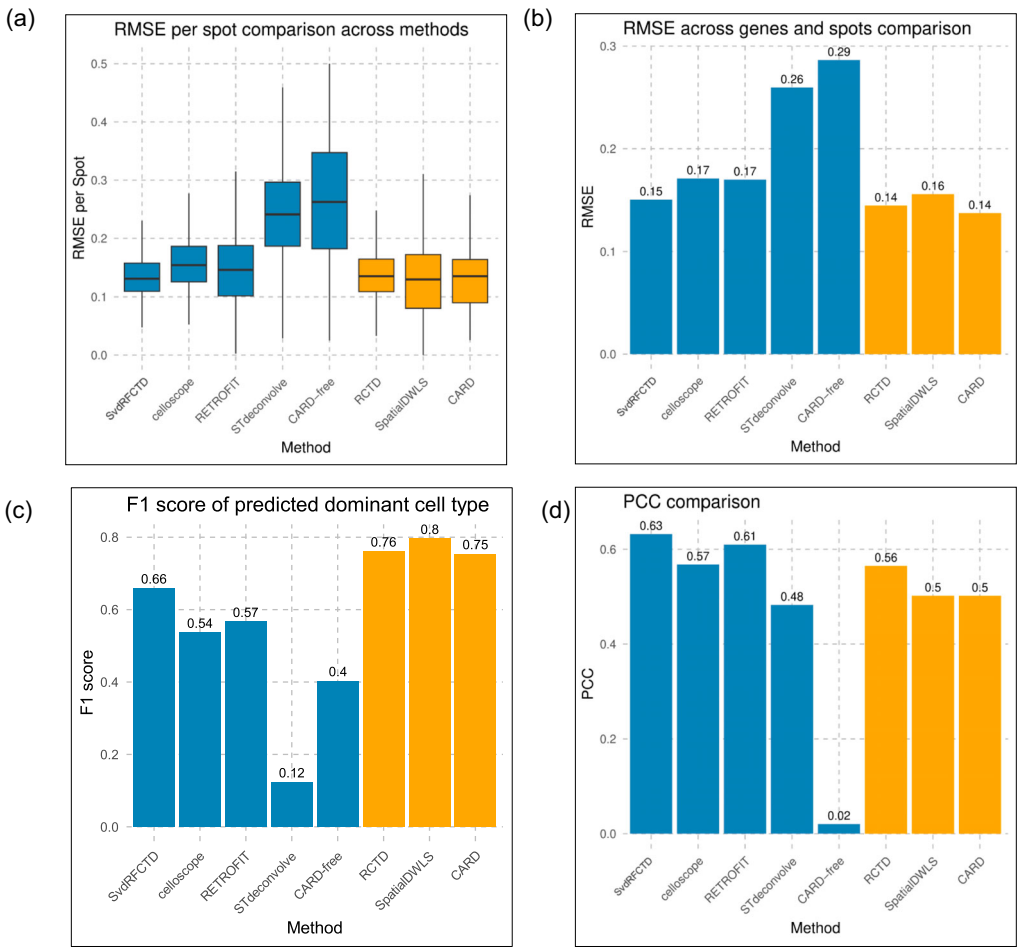

Figure 2. Comparison of SvdRFCTD with other methods on simulated mouse kidney ST data. (a) RMSE per spot between the predicted and the ground truth cell type proportions using different methods. The blue bars represent reference-free methods, and the orange bars represent reference-based methods. (b) RMSE across genes and spots between the predicted and the ground truth cell type proportions using different methods. The blue bars represent reference-free methods, and the orange bars represent reference-based methods. (c) The F1 score between the ground truth and the dominant cell types inferred by different methods. The blue bars represent reference-free methods, and the orange bars represent reference-based methods. (d) PCC between the predicted and the ground truth cell type proportions.

cell-type heterogeneity. However, PCC tends to favour reference-free methods, likely due to their robustness against systematic biases in scRNA-seq data and their reliance on a smaller subset of genes, which helps mitigate the impact of noise. Additionally, reference-based methods may be more susceptible to overfitting or biases arising from incomplete scRNA-seq references, limiting their generalizability. Figure 3(a) shows the heatmap of the PCC between the true cell type proportions and those inferred by SvdRFCTD (see Figure 3(b)), demonstrating a high degree of similarity between the two. For each cell type, we illustrate the cell type proportions across the entire tissue (see Figure 4). At each spot, a stronger red colour indicates a higher proportion of that cell type at that location. From Figure 4, we can also

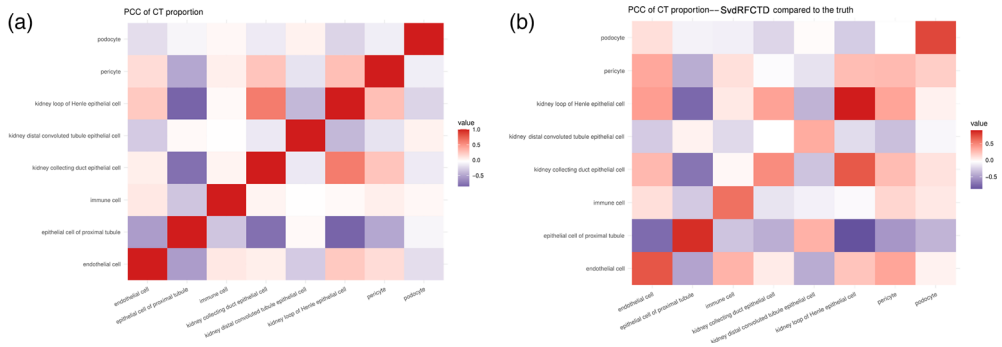

Figure 3. Comparison of true and inferred cell type proportion correlations on simulated mouse kidney ST data. (a) PCC of true cell type proportions. (b) PCC of cell type proportions inferred by SvdRFCTD.

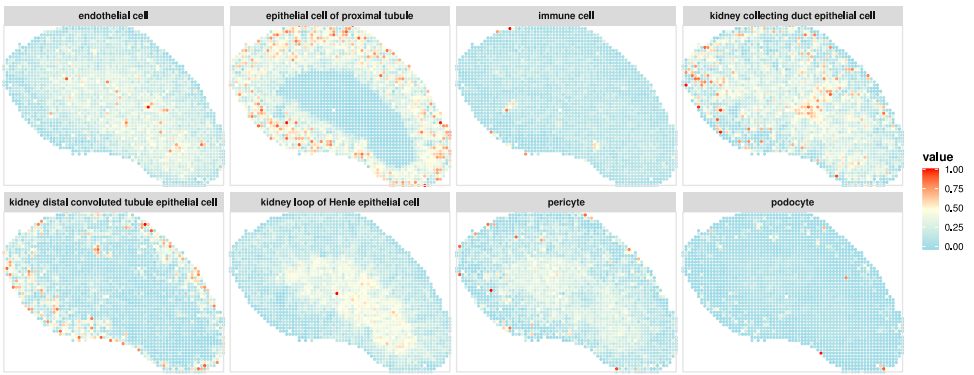

Figure 4. Spatial distribution of inferred cell type proportions by SvdRFCTD on simulated mouse kidney ST data.

clearly observe that different cell types exhibit distinct spatial distribution patterns across the tissue. Thus, the validity of SvdRFCTD is verified.

3.3. Sensitivity analysis of SvdRFCTD for marker genes

To assess the robustness of SvdRFCTD under different marker gene selection strategies, we designed three experimental settings: (1) Setting 1, where marker genes were selected to match those used in Section 3.2 from the TMS marker gene set; (2) Setting 2, where the number of marker genes per cell type was limited to a maximum of 10; and (3) Setting 3, where the number of marker genes per cell type was further restricted to a maximum of 5. We applied SvdRFCTD, Celloscope, and RETROFIT to deconvolve cell type proportions across these settings and evaluated their performance using RMSE, PCC, and F1 score. We did not include STdeconvolve and CARD-free in this comparison. STdeconvolve does not utilize marker gene information during deconvolution and cell type matching, meaning its predictions remain unchanged regardless of the marker gene selection. CARD-free fails when the average number of unique marker genes per cell type is less than 20, making it unusable in Settings 2 and 3.

Across all evaluation metrics, SvdRFCTD outperformed the other methods, achieving the lowest RMSE, the highest and most stable PCC, and the highest F1 score, despite some variability in the latter. Figure 5(a) presents the RMSE per spot under different marker gene settings, where SvdRFCTD consistently demonstrated the best performance and exhibited strong robustness across all three settings. A similar trend is observed in Figure 5(b), which shows RMSE across genes and spots. As the number of marker genes per cell type decreased, the RMSE of SvdRFCTD progressively decreased, mirroring the trend seen with Celloscope. In contrast, RETROFIT displayed a different pattern, reaching its highest RMSE in Setting 2 and exhibiting relatively large fluctuations across settings. Figure 5(c) illustrates the PCC under different marker gene settings. SvdRFCTD not only achieved the highest PCC but also maintained remarkable stability, whereas RETROFIT showed substantial variation across settings. Finally, Figure 5(d) presents the F1 score under different marker gene settings. Although SvdRFCTD exhibited some variability, it achieved the highest F1 score, reflecting its superior classification accuracy in identifying dominant cell types.

As shown in Figure 5, reducing the number of marker genes per cell type led to performance improvements across all metrics for SvdRFCTD and the other two reference-free deconvolution methods. This is likely because the removed marker genes were not the most strongly expressed, cell type-specific genes, and their exclusion helped reduce noise from redundant information.

These results underscore the robustness and effectiveness of SvdRFCTD in reference-free deconvolution. It consistently outperforms existing methods in RMSE, PCC, and F1 score while exhibiting greater stability across different marker gene settings. Notably, although performance improves with optimized marker selection, SvdRFCTD remains the most reliable method regardless of the marker gene constraints, making it well-suited for applications with limited marker information.

4. Case study

4.1. Application of SvdRFCTD to the anterior section of the mouse brain dataset

We applied our method to sagittal mouse brain dataset generated using the 10X Visium protocol. The dataset contains a pair of replicates from the anterior regions of the brain, labelled 'anterior1/2'. Our analysis focuses on the 'anterior1' dataset, which contains 3355 spots in the tissue, with a median of 4772 genes per spot. From this dataset, we selected 2696 spots and 179 marker genes to perform SvdRFCTD, and the marker genes are collected from the previous study (Zeisel et al., 2018), which are listed in Supplementary Table S2. To evaluate the performance of SvdRFCTD, we compare its results with those obtained using other deconvolution methods, including Celloscope, RETROFIT, RCTD, SpatialDWLS and CARD. Here, we do not include STdeconvolve and CARD-free as comparative methods because STdeconvolve requires ground truth of proportions for cell type assignment, and CARD-free exhibits insufficient accuracy. The single-cell reference dataset for these reference-based methods was obtained from the Allen Mouse Brain Atlas, which is different from the spatial transcriptomics dataset analyzed here. Consequently, two cell types present in the marker list of the spatial transcriptomics dataset are not found in the reference dataset, so we simply exclude these two cell types from the comparison.

Unlike the simulated spatial transcriptomics datasets, the realistic dataset does not have ground truth that specifies the exact cell type composition at each location. Therefore, we

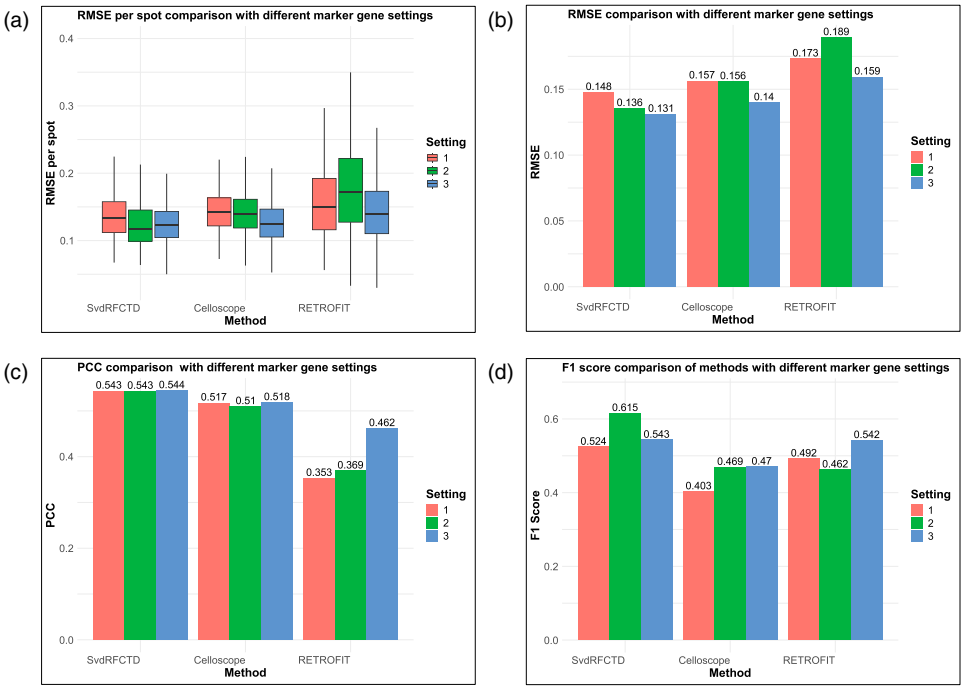

Figure 5. Sensitivity analysis of SvdRFCTD on simulated mouse kidney ST data. (a) RMSE per spot between the predicted and the ground truth cell type proportions using different methods under different marker gene settings. (b) RMSE across genes and spots between the predicted and the ground truth cell type proportions using different methods under different marker gene settings. (c) PCC between the predicted and the ground truth cell type proportions under different marker gene settings. (d) The F1 score between the ground truth and the dominant cell types inferred by different methods under different marker gene settings.

use the clusters identified by the 10X Genomics platform (see Figure 6) for comparison with the deconvolution results. We calculate the PCC values between the cell type proportions estimated by SvdRFCTD and those obtained using other methods (see Figure 7(a)), which reflects the similarity of the results between SvdRFCTD and the other methods. The results show that SvdRFCTD exhibits a higher similarity with reference-free methods, while also maintaining consistency with reference-based methods. Specifically, we examine the similarity between SvdRFCTD and RCTD results for the same cell type and observe a high degree of consistency across most cell types (see Figure 7(b)). This demonstrates that the performance of SvdRFCTD is comparable to existing deconvolution methods.

SvdRFCTD clearly reflects the similarity of spatial location of cell type proportions across the tissue. Figure 7(c) shows the PCC values for all pairs of cell type proportions estimated by SvdRFCTD, which allows us to assess the spatial co-occurrence and exclusivity of different cell types. These cell types can be categorized by function and origin as Neuronal cells(dopaminergic neurons, DOPA; GABAergic neurons, GABA; subtype of GABAergic neurons, GABA-sub; glutamatergic neurons, GLUT), Glial cells(astrocytes, ASC; olfactory ensheathing glia, OEG; oligodendrocytes, OLG; microglia, MG), Endothelial and choroid plexus-associated cells(endothelial cells, EC; choroid plexus epithelial cells, CPC), Vascular

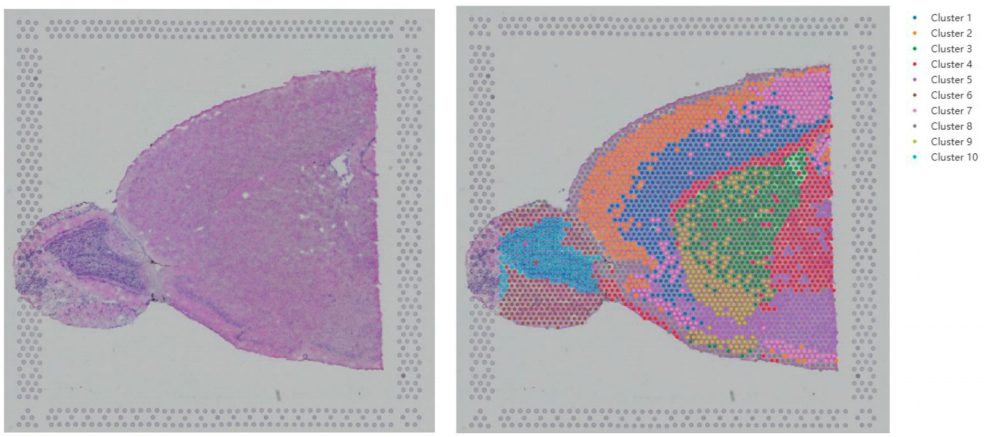

Figure 6. Hematoxylin and eosin (H&E) staining image of the sagittal anterior region (left), and it is coloured by cluster from the 10x Genomics platform(right).

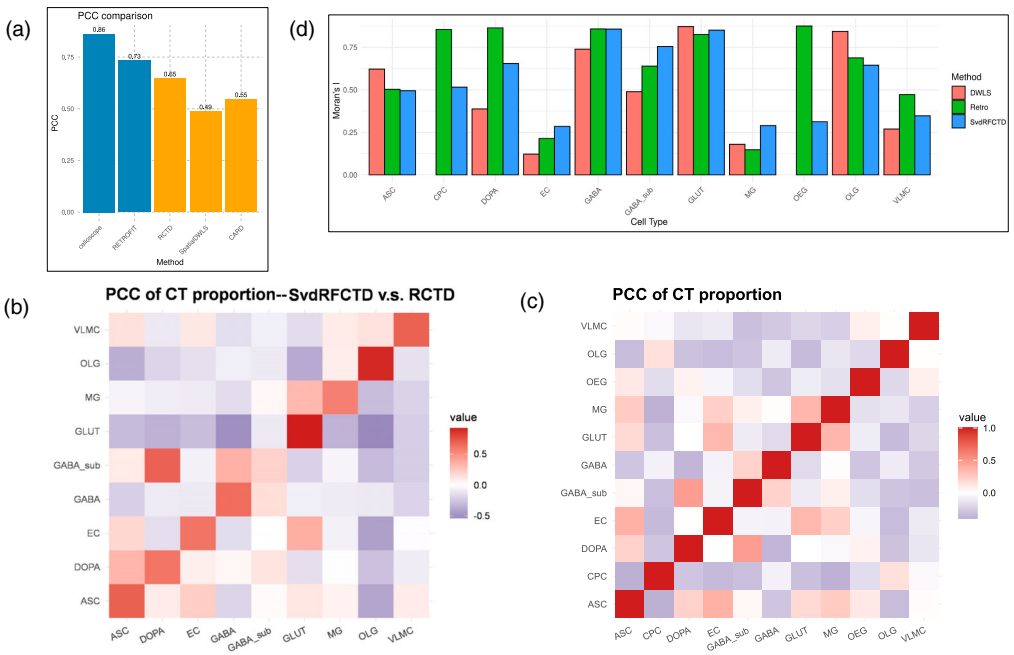

Figure 7. SvdRFCTD is applied to cell type decomposition on the anterior mouse brain (sagittal section). (a) PCC between SvdRFCTD and the other methods. (b) PCC for each pair of cell type proportions estimated by SvdRFCTD and RCTD. (c) PCC for all pairs of cell type proportions estimated by SvdRFCTD. (d) Moran's I coefficient for cell types indicated by SpatialDWLS, RETROFIT and SvdRFCTD. The cell type in this dataset contains ASC, astrocytes; CPC, choroid plexus epithelial cells; DOPA, dopaminergic neurons; EC, endothelial cells; GABA, GABAergic neurons; GABA-sub, GABAergic neurons subtype; GLUT, glutamatergic neurons; OEG, olfactory ensheathing glia; OLG, oligodendrocytes; MG, microglia; VLMC, vascular and leptomeningeal cells.

and perivascular supporting cells(vascular and leptomeningeal cells, VLMC). Figure 7(c) shows that the spatial co-localization of astrocytes (ASC) with dopaminergic neurons (DOPA), endothelial cells (EC), glutamatergic neurons (GLUT), and microglia (MG), is consistent with previous studies (Erö et al., 2018; Langlieb et al., 2023). Astrocytes provide metabolic support, maintain blood-brain barrier (BBB) integrity, and regulate synaptic functions, influencing both excitatory and inhibitory neurons (GLUT and GABA, respectively). The end-feet of astrocytes envelop the outer walls of cerebral blood vessels, working alongside endothelial cells to maintain BBB function. Together, they regulate blood flow and respond to neural activity (Abbott et al., 2006). Under inflammatory or injury conditions, astrocytes interact with microglia (MG), releasing inflammatory mediators to coordinate neural repair (Liddelow et al., 2017). In the case of dopaminergic neurons (DOPA), astrocytes regulate dopamine levels by clearing its metabolic byproducts, thereby maintaining the balance of neural networks (Chinta & Andersen, 2008). Additionally, dopaminergic neurons regulate the activity of glutamatergic neurons (GLUT) and GABAergic neurons (GABA) through dopamine release (Tritsch et al., 2012). Their co-localization in certain regions is also evident in Figure 8. GABAergic neurons (GABA) and glutamatergic neurons (GLUT) are inhibitory and excitatory neurons, respectively. They form well-defined network partitions in different cortical layers, collectively maintaining the balance of excitation and inhibition within neural networks. GABA_sub is a specialized subset of GABAergic neurons, which also exhibit notable spatial co-occurrence (Markram et al., 2004). However, they display significant differences in proportions across different brain regions (see Figure 8).

SvdRFCTD effectively maps different cell types to distinct regions of the brain, as shown in Figures 8 and 9. GLUT, GABA, and ASC are broadly distributed across the cortex, cerebellum, and hippocampus. OEG cells, on the other hand, are predominantly located in the olfactory bulb, where they serve as the primary projection neurons responsible for transmitting olfactory information. OLG cells, which produce myelin to enhance neuronal signal conduction efficiency, are often concentrated in specific brain regions. In contrast, MG cells, as the immune cells of the central nervous system, are tasked with monitoring and clearing pathogens or damage. As a result, they are widely distributed throughout the entire brain. And EC forms the primary component of the blood-brain barrier, regulating the exchange of substances between the bloodstream and brain tissue. Consequently, they are distributed relatively evenly across the brain. To verify the spatial autocorrelation exists or not for each cell type, we also calculate Moran I's coefficient, which takes values from -1 to 1, where -1 indicates perfect dispersion, 0 perfect randomness (no autocorrelation), and 1 signifies perfect clustering of similar values. Consequently, higher Moran's I values suggest that the inferred cell types are spatially clustered. As shown in Figure 7(d), most cell types show high spatial autocorrelation, but the coefficients for EC and MG are relatively low, consistent with our earlier discussion. The dominant cell type on each spot inferred by SvdRFCTD, Celloscope, RETROFIT, RCTD, SpatialDWLS, and CARD is shown in Figure 9, which visually illustrates the cell types aggregated in each region of the brain.

4.2. Application of SvdRFCTD to the coronal section of the mouse brain dataset

Further, we performed SvdRFCTD on coronal mouse brain slice which is orthogonal to the sagittal section analyzed above. The marker genes used in coronal mouse brain dataset are listed in Supplementary Table S3. The hematoxylin and eosin (H&E) stained image of the coronal anterior region is shown in Figure 10, coloured by clusters from the 10x Genomics

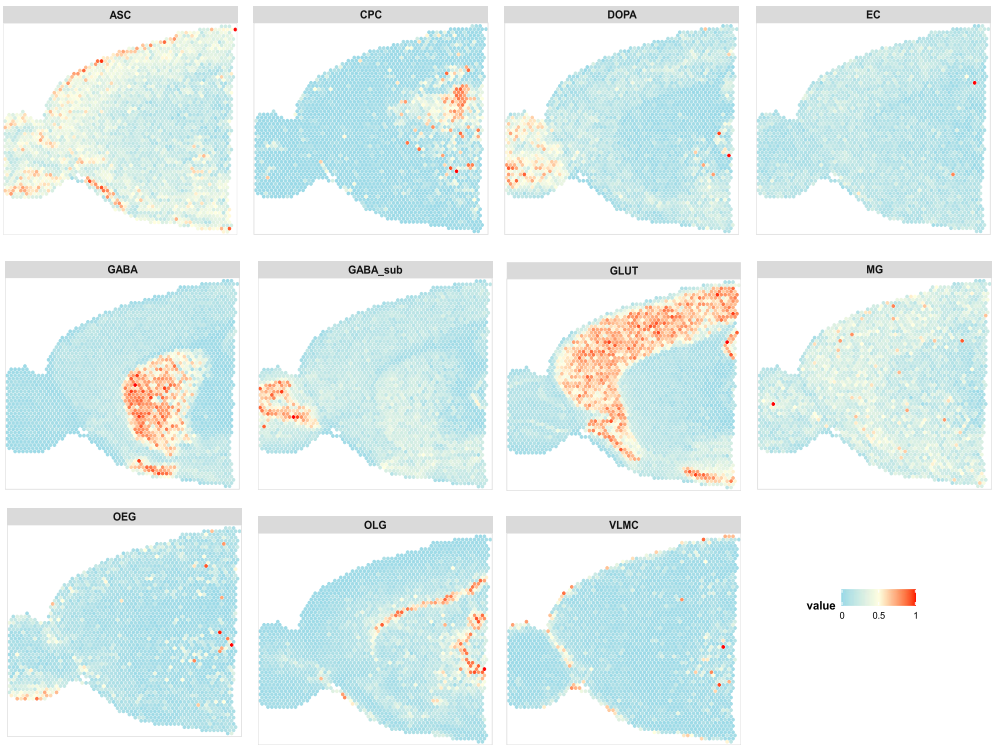

Figure 8. Heatmaps of spatial cell type composition in the anterior mouse brain (sagittal section) estimated by SvdRFCTD.

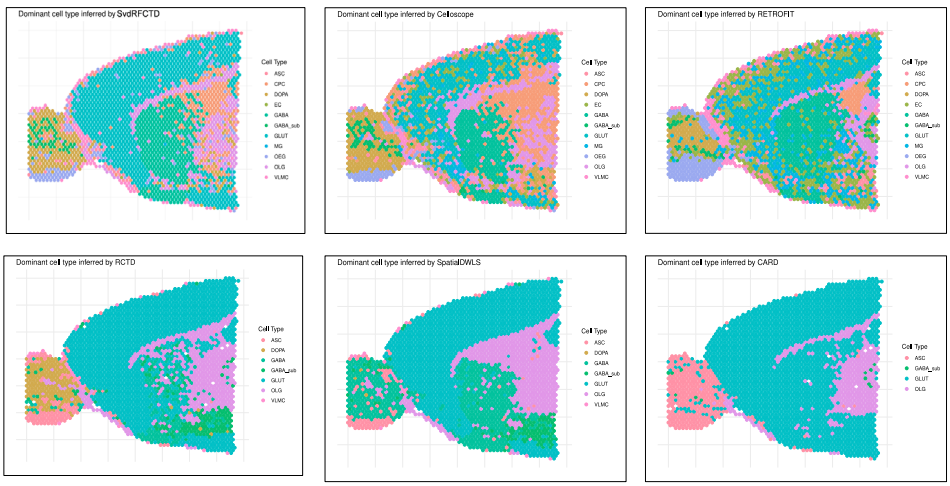

Figure 9. Dominant cell type inference in the anterior mouse brain (sagittal section) by SvdRFCTD and other methods.

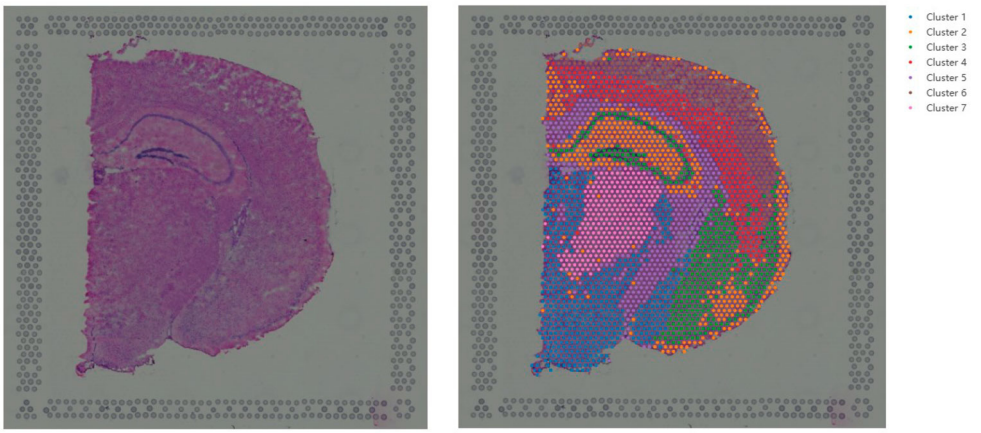

Figure 10. Hematoxylin and eosin (H&E) staining image of the coronal anterior region (left), and it is coloured by cluster from the 10x Genomics platform(right).

platform without explicit cell type annotations. This dataset largely comprises the same cell types as the sagittal data, including oligodendrocytes, astrocytes, GABAergic neurons, glutamatergic neurons, choroid plexus epithelial cells, endothelial cells, microglia, and vascular and leptomeningeal cells. Additionally, we incorporated cholinergic neurons(CHOL), peptidergic cells(PEPTI), granule cells(GRANULE), and the GLUT subtypes GLUT_cortex and GLUTmid. However, the single-cell reference dataset lacks CHOL, PEPTI, and GRANULE cell types, which impacts the comparison. For instance, the absence of granule cells may lead the reference-based methods to overestimate the proportion of GLUTmid cells in the same region. Nevertheless, SvdRFCTD still shows consistency with most methods in terms of cell type proportion estimation (see Figures 11(a) & 13).

SvdRFCTD gives an insight into the spatial co-occurrence relationships of cell types from the PCC values of the estimated cell type pairs (see Figure 11(b)). GLUT, an excitatory neuron, is distributed throughout the brain, while its two subtypes differ in their spatial distribution. GLUT_cortex is concentrated in the cerebral cortex, where it is responsible for higher cognitive functions, while GLUTmid is located in the midbrain and surrounding regions, where it plays a role in motor control and regulation of the dopamine system. From the PCC plot we can see the negative spatial correlation between these two subtypes, which is also clearly visible in the cell type distribution heatmap (see Figure 12). The only cell types that show a positive correlation with the distribution of GLUT_cortex are GABA and granule cells, while other cell types show some degree of different distribution patterns. Relative studies suggest that GLUT_cortex neurons are concentrated in the cortex and provide excitatory signals, while GABA neurons are distributed in different cortical layers and locally modulate the excitatory activity of GLUT_cortex (Murata et al., 2019). Granule cells in the olfactory bulb are inhibitory neurons that primarily regulate the activity of excitatory projection neurons (such as GLUT neurons) via GABA signalling, and are located in the deep layers of the olfactory bulb, close to the excitatory neurons (Erö et al., 2018). CHOL are concentrated in the basal forebrain, where they regulate memory functions in the cortex and hippocampus. PEPTI, located in the hypothalamus, play a key role in regulating neuroendocrine functions. They together form an interconnected regulatory network that influences neural activity

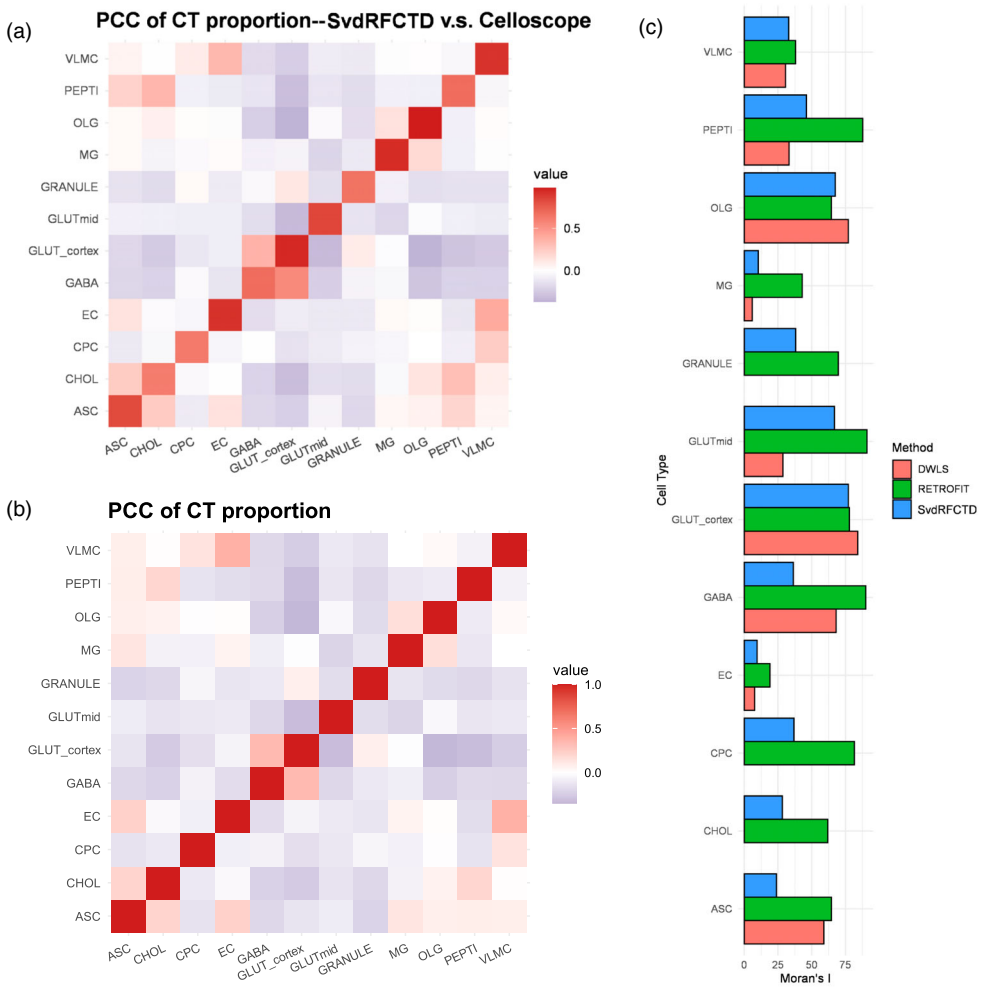

Figure 11. SvdRFCTD is applied to cell type decomposition on the anterior mouse brain (coronal section). (a) PCC for each pair of cell type proportions estimated by SvdRFCTD and Celloscope. (b) PCC for all pairs of cell type proportions estimated by SvdRFCTD. (c) Moran's I coefficient for cell types indicated by SpatialDWLS, RETROFIT and SvdRFCTD. The cell types contains OLG, oligodendrocytes; OEG, olfactory ensheathing glia; ASC, astrocytes; GABA, GABAergic neurons; GLUT, glutamatergic neurons; CPC, choroid plexus epithelial cells; EC, endothelial cells; MG, microglia; VLMC, vascular and leptomeningeal cells; CHOL, cholinergic neurons; PEPTI, peptidergic cells; GRANULE, granule cells.

in multiple regions, including the cortex and hypothalamus (Zaborszky et al., 2012). These findings from existing research are consistent with our results.

Furthermore, SvdRFCTD effectively identifies region-specific cell types and structure of the tissue. Moran's I coefficients for cell types are shown in Figure 11(c), indicating spatial clustering for GLUTmid, GLUT_cortex, and OLG, while EC and MG exhibit spatial dispersion. These results are consistent with the Moran's I analysis in Section 4.1. SvdRFCTD successfully identifies region-specific cell types (see Figure 12). The peptidergic cells were accurately localized in the hypothalamus, consistent well with their role in regulating neuroendocrine function and hormone secretion in this region (Lein et al., 2007). Similarly,

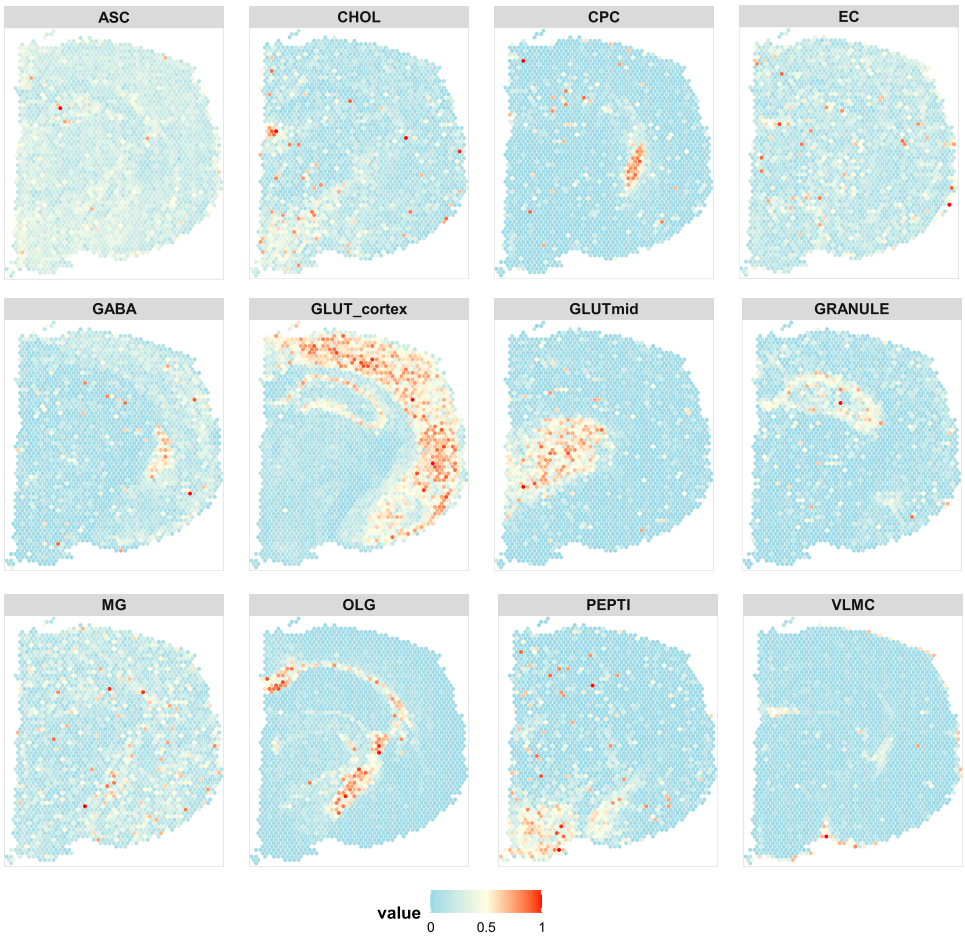

Figure 12. Heatmaps of spatial cell type composition in the anterior mouse brain (coronal section) estimated by SvdRFCTD.

granule neurons were correctly mapped to the dentate gyrus of the hippocampus, a key area involved in memory formation and spatial navigation. Additionally, the medium spiny neurons were precisely mapped to the basal ganglia, consistent with their critical role in motor control and reward processing (Caligiore et al., 2019). SvdRFCTD effectively visualizes the structure of the tissue in the dominant cell type plot (see Figure 13), which highlights the cell type with the highest proportion at each spot. This visualization offers a clear and intuitive depiction of the cellular composition across various brain regions, including fiber tracts, ventricles, cortex, thalamus, and hypothalamus. These regions, annotated based on anatomical images from the Allen Brain Atlas, are distinctly and accurately identified by SvdRFCTD. In comparison, Celloscope provides a less detailed characterization of GRANULE in the corresponding regions and fails to capture the complex structure of the midbrain. Meanwhile, RCTD and CARD significantly underestimate the proportions of GLUTmid and PEPTI in the relevant areas while overestimating those of ASC and OLG. These results highlight the effectiveness of SvdRFCTD in capturing region-specific cellular distributions.

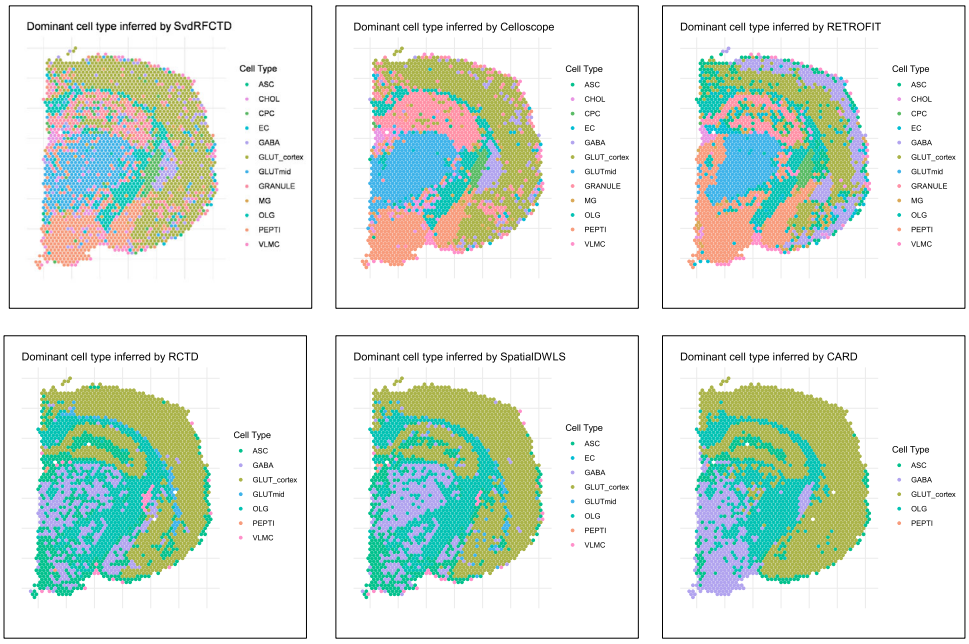

Figure 13. Dominant cell type inference in the anterior mouse brain (coronal section) by SvdRFCTD and other methods.

4.3. Application of SvdRFCTD to human pancreatic ductal adenocarcinomas dataset

Next, we applied our methods to human pancreatic ductal adenocarcinoma (PDAC) dataset (Moncada et al., 2020). The PDAC dataset comprises 428 spots, 1628 genes, and 20 cell types or subtypes. The marker genes were obtained from a sample-matched scRNA-seq dataset generated using the inDrop platform (Moncada et al., 2020), which are listed in Supplementary Table S4. The dataset includes annotations for four main anatomical tissue regions (cancer, pancreatic, ductal, and stromal regions) provided by histologists based on H&E staining and the region annotations are also validated in the original study (Moncada et al., 2020) (see Figure 14(a)).

SvdRFCTD effectively identified highly concentrated cell types in specific regions. By comparing the proportions of cell types in each region with those in other areas, we find that the ductal region exhibits a high proportion of ductal cells, including subtypes such as terminal ductal-like ductal cells, CRISP3-high centroacinar-like ductal cells, MHC Class II-expressing ductal cells, and APOL1-high hypoxic ductal cells (see Figure 14(b)). The pancreatic region shows a significant enrichment of acinar cells (see Figure 14(c)), while the cancer region was dominated by cancer cells, including Cancer Clone A and Cancer Clone B (see Figure 14(d)). These differences were statistically significant, with t-test p-values < 0.05. This pattern is also evident in the heatmaps of cell type distribution (see Figure 15(a)) and the dominant cell type plot (see Figure 15(b)). Specifically, Figure 15(a) reveals a spatially structured distribution of cell types across spots, highlighting distinct regional enrichment patterns. In Figure 15(b), the dominant cell type assigned to each spot by SvdRFCTD aligns well with known anatomical and functional structures, further supporting the method's

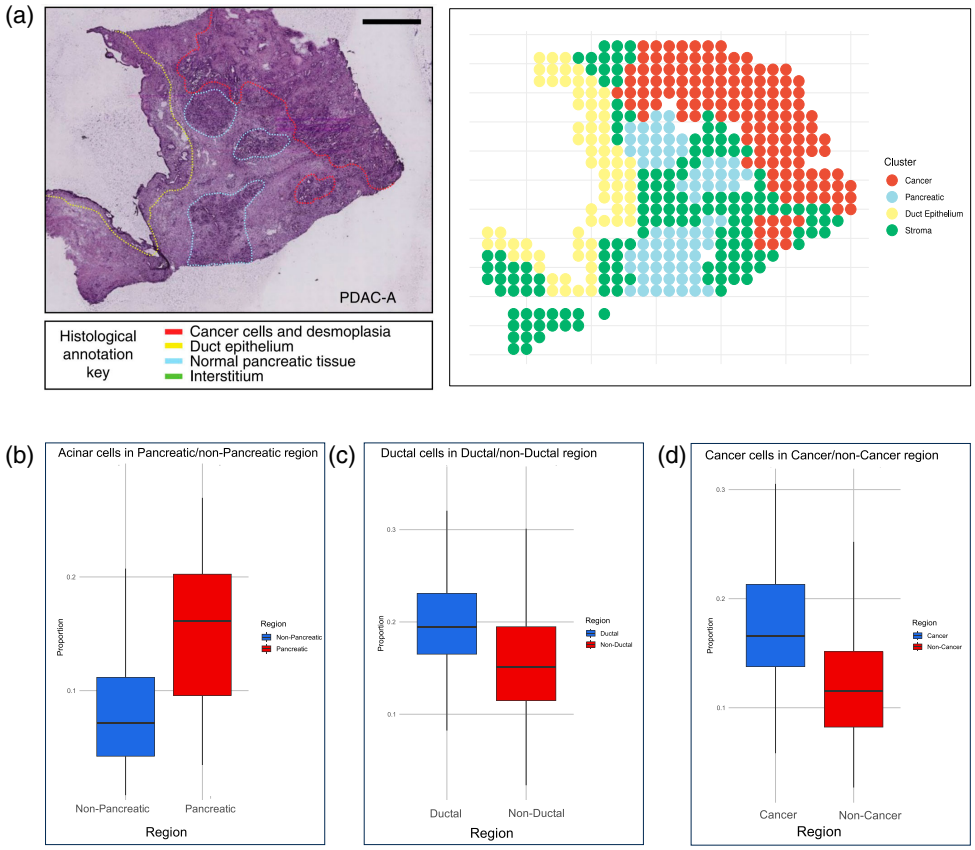

Figure 14. SvdRFCTD is applied to cell type decomposition on human pancreatic ductal adenocarcinomas dataset. (a) Hematoxylin and eosin (H&E) staining image of the PDAC tissue, the regions are annotated by original research. (b) A comparison of the predicted proportion of acinar cells in the spots of the pancreatic region and the spots of the non-pancreatic region. (c) A comparison of the predicted proportion of Ductal cells in the spots of the Ductal region and the spots of the non-Ductal region. (d) A comparison of the predicted proportion of Cancer cells in the spots of the Cancer region and the spots of the non-Cancer region. The cell types contain Acinar cells; Ductal cells; Cancer cells; Fibroblasts; mDcs, myeloid dendritic cells; Tuft cells; pDCs, plasmacytoid dendritic cells; Endocrine cells; Endothelial cells; Macrophages; Mast cells; T cells and natural killer (NK) cells; Monocytes; RBCs, Red blood cells.

accuracy, which represents that SvdRFCTD provides a refined resolution of cell type heterogeneity, particularly in boundary regions where transitions between different cell types occur.

Additionally, we observe that a subset of ductal cells extended into the cancer region. This finding is supported by a comparison of the proportions of ductal subtypes between cancer and non-cancer regions, which shows a slightly higher proportion of ductal high hypoxic cells in the cancer region (see Figure 16(a)). This observation is biologically plausible as some ductal cells may be relocated to the tumour-adjacent regions or transition into cancer-like phenotypes during tumorigenesis. In particular, ductal high hypoxic cells could play a critical role by secreting specific factors, such as VEGF and immunosuppressive molecules,

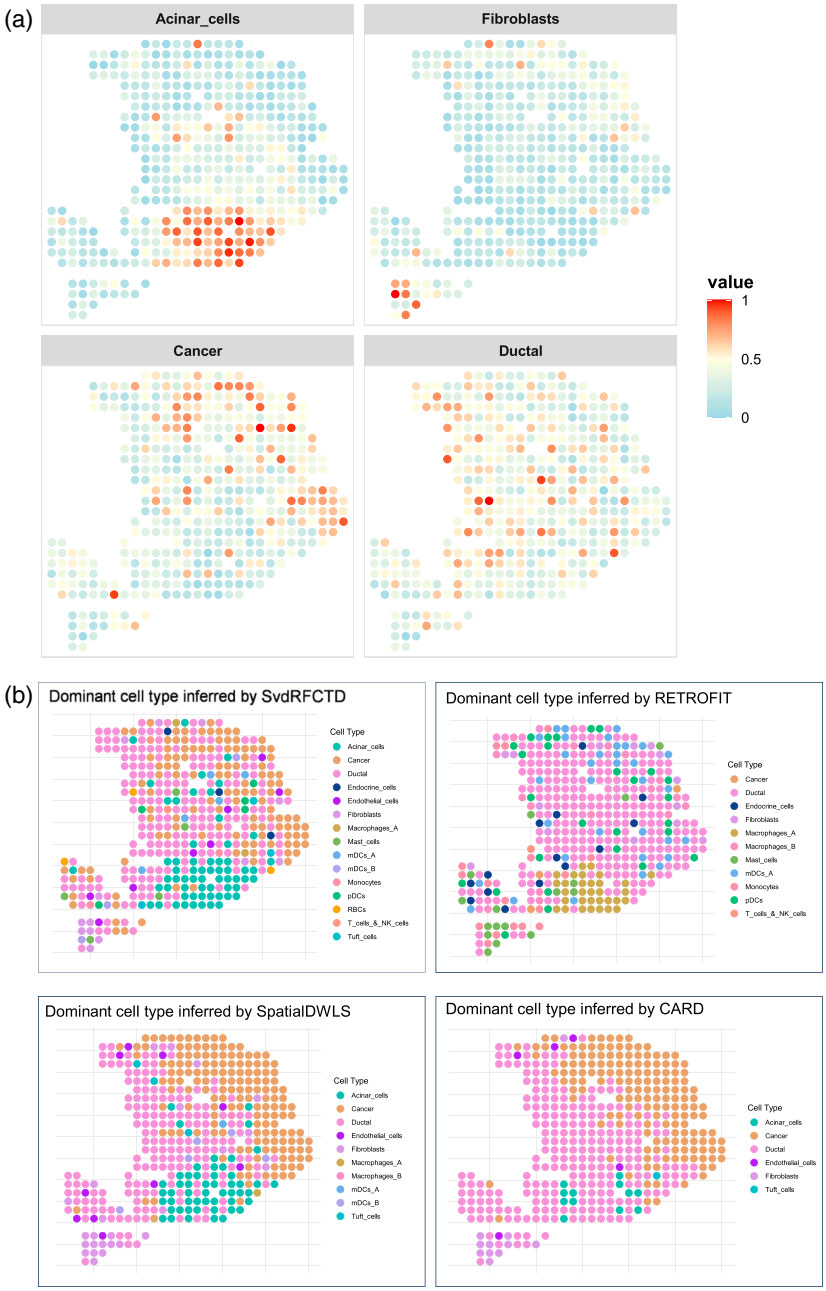

Figure 15. SvdRFCTD is applied to cell type decomposition on human pancreatic ductal adenocarcinomas dataset. (a) Heatmaps represents spatial composition of cell types across spots estimated by SvdRFCTD. (b) The dominant cell type on each spot inferred by SvdRFCTD, RETROFIT, SpatialDWLS and CARD respectively.

to promote nutrient supply, reshape the tumour microenvironment, and regulate immune responses, which ultimately facilitate cancer cell adaptation and survival (Hwang et al., 2024).

SvdRFCTD explores the spatial co-occurrence relationships between cell types (see Figure 16(b)). We observe a co-occurrence of cancer clone A cells with macrophages B and

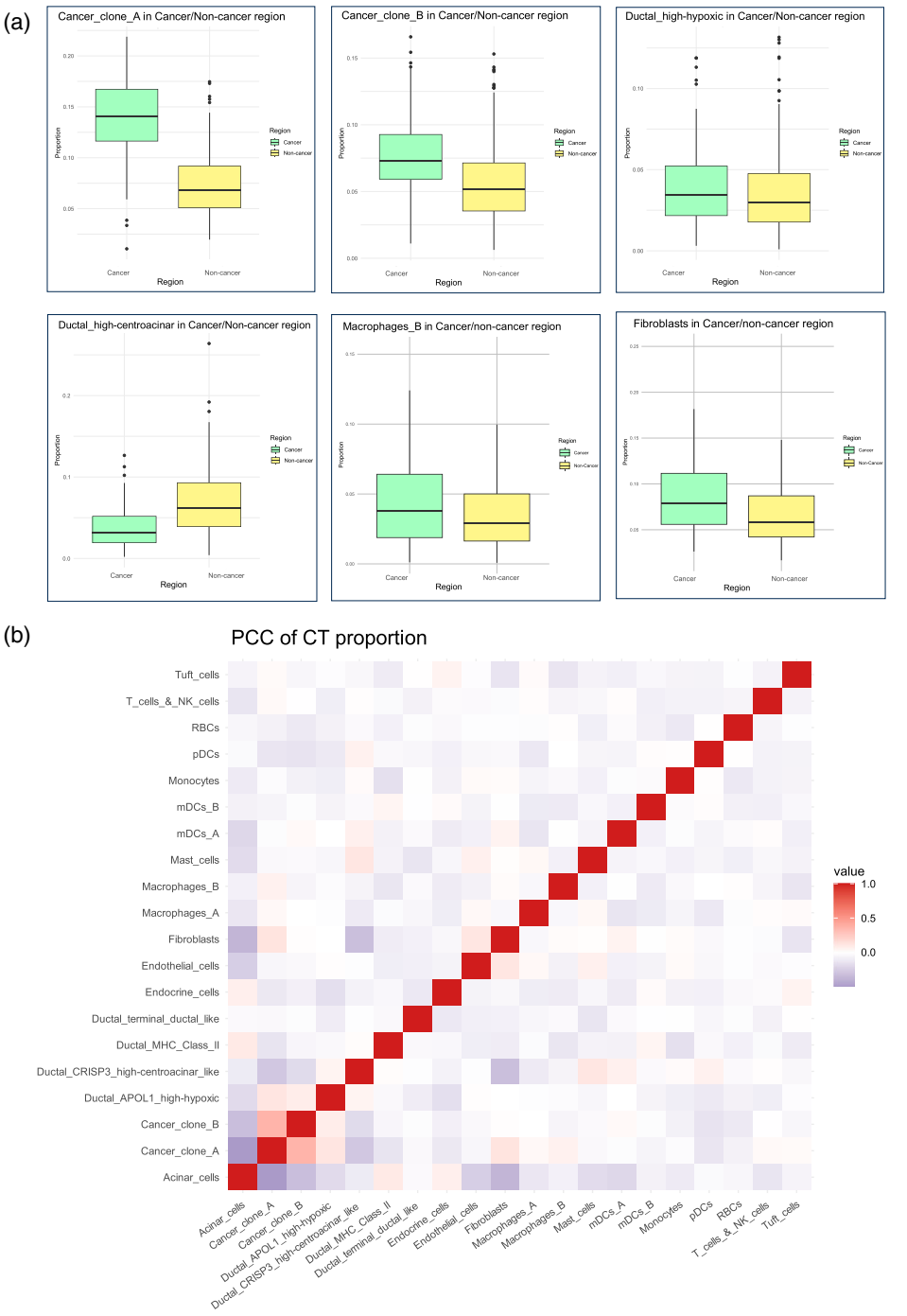

Figure 16. SvdRFCTD is applied to cell type decomposition on human pancreatic ductal adenocarcinomas dataset. (a) PCC for each pair of cell type proportions estimated by SvdRFCTD. (b) Comparisons of cell type proportions inferred by SvdRFCTD in cancer region v.s. non-cancer region.

fibroblasts. This observation is validated by comparing the proportions of macrophages and fibroblasts between cancer and non-cancer regions, showing that these cell types are more abundant in cancer regions (see Figure 16(a)). Macrophages B, are commonly located in the tumour core and are closely associated with both cancer cells and fibroblasts. These macrophages play a critical role in supporting tumour immune evasion by promoting an immunosuppressive microenvironment (Yu et al., 2024). Fibroblasts, on the other hand, contribute to tumour progression by secreting extracellular matrix (ECM) proteins (such as collagen) and pro-inflammatory factors (Du et al., 2024). These interactions enable fibroblasts to collaborate with ductal cells and immune cells to actively remodel the tumour microenvironment (TME). For comparison, when analyzing the PDAC data, the CARD method reached an opposite conclusion regarding the distribution of Macrophages B in cancer and non-cancer regions and failed to capture the spatial distribution of fibroblasts (Ma & Zhou, 2022). We also observe a spatial co-occurrence between endothelial cells and fibroblasts. This is likely due to their coordinated roles in shaping the TME: fibroblasts provide structural support and signalling cues for angiogenesis, while endothelial cells form the vascular network essential for nutrient supply and waste removal in the tumour (Morvaridi et al., 2015; Nielsen et al., 2016). Together, these findings highlight the intricate spatial and functional interplay among cancer cells, immune cells, and stromal components in PDAC tissues.

5. Discussion

SvdRFCTD demonstrates outstanding performance in cell type decomposition through extensive simulations and real data analyses. Using the mouse kidney simulated ST dataset, we investigate the optimal model structure and find that avoiding redundant spatial-specific and gene-specific effects improves performance. Additionally, using the total number of cells in each ST spot, rather than total transcripts, enhances accuracy due to the significant variation in cell numbers across spots. When compared with other deconvolution methods, including Celloscope, RETROFIT, RCTD, SpatialDWLS, and CARD, SvdRFCTD achieves results comparable to reference-based methods across multiple evaluation metrics. Notably, it outperforms all other methods in Pearson Correlation Coefficient (PCC) when compared to true cell type proportions. To further assess the robustness of SvdRFCTD, we conducted a sensitivity analysis under different marker gene selection strategies. The results show that SvdRFCTD consistently outperforms other reference-free methods and remains the most reliable method regardless of marker gene constraints. In analyses of three real ST datasets, SvdRFCTD accurately estimates cell type proportions, consistent with other methods, while clearly capturing tissue anatomical structures and identifying regions with high cellular aggregation. It also reveals spatially coexisting cell types, shedding light on complex inter-cellular relationships.

SvdRFCTD shows its ability to differentiate between cell subtypes when sufficient marker genes are available for definition. For example, in the PDAC dataset, several subtypes of GLUT cells are distinguished. While the overall spatial distribution of GLUT cells shows limited correlation with cancer cells, the ductal high hypoxic GLUT cells are predominantly located in the cancer region of the tissue. These cells play a pivotal role in supporting the development of cancer cells, highlighting their functional significance in the tumour microenvironment.

A key advantage of SvdRFCTD is the ability to account for spatial effects by borrowing information from neighbouring spots and its independence from reference data. By incorporating spillover effects into both the mean and dispersion components, SvdRFCTD effectively captures spatial relationships. The Moran's I tests in the case study also support the similarity of cell type composition between neighbouring locations. While reference-based methods assume constant cell-type-specific gene expression, this assumption is often violated due to batch effects and sample variability, reducing their accuracy. In contrast, reference-free methods like SvdRFCTD avoid this issue.

There are also limitations to SvdRFCTD. First, its use of MCMC for parameter estimation is computationally intensive, making it less scalable for large datasets. Approximate inference methods like INLA could be explored to improve efficiency. Second, the model does not incorporate covariates or account for spatial effects on cell type proportions, focussing only on spillover effects between spots. Expanding its spatial modelling capabilities could enhance performance. Lastly, the accuracy of SvdRFCTD heavily depends on the selection of input marker genes, requiring sufficient prior knowledge on the dataset. Despite its limitations, SvdRFCTD provides a valuable tool for advancing our understanding of the cellular architecture in tissues. Future developments focussing on scalability, covariate integration, and enhanced spatial modelling will further expand its applicability and impact in the field of spatial transcriptomics.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The research is supported by the Natural Science Foundation of China [grant numbers 12201219, 12271168, 12171229] and Natural Science Foundation of Shanghai [grant numbers 23JS1400500, 23JS1400800, 22ZR1420500].

References

Abbott, N. J., Rönnbäck, L., & Hansson, E. (2006). Astrocyte–endothelial interactions at the blood–brain barrier. *Nature Reviews Neuroscience*, *7*(1), 41–53. https://doi.org/10.1038/nrn1824

Allen, C., Chang, Y., Neelon, B., Chang, W., Kim, H. J., Li, Z., Ma, Q., & Chung, D. (2021). *A Bayesian multivariate mixture model for spatial transcriptomics data*. bioRxiv.

The Tabula Muris Consortium, Almanzar, N., Antony, J., Baghel, A. S., Bakerman, I., Bansal, I., Barres, B. A., Beachy, P. A., Berdnik, D., Bilen, B., Brownfield, D., Cain, C., Chan, C. K. F., Chen, M. B., M. F. Clarke, Conley, S. D., Darmanis, S., Demers, A., Demir, K., De Morree, A., … Zou, J. (2020). A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature*, *583*(7817), 590–595. https://doi.org/10.1038/s41586-020-2496-1

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC.

Bankhead, P., Loughrey, M. B., Fernández, J. A., Dombrowski, Y., McArt, D. G., Dunne, P. D., McQuaid, S., Gray, R. T., Murray, L. J., Coleman, H. G., James, J. A., Salto-Tellez, M., & Hamilton, P. W. (2017). QuPath: Open source software for digital pathology image analysis. *Scientific Reports*, *7*(1), Article 16878. https://doi.org/10.1038/s41598-017-17204-5

Biancalani, T., Scalia, G., Buffoni, L., Avasthi, R., Lu, Z., Sanger, A., Tokcan, N., C. R. Vanderburg, Segerstolpe, Å., Zhang, M., Avraham-Davidi, I., Vickovic, S., Nitzan, M., Ma, S., Subramanian, A., Lipinski, M., Buenrostro, J., Brown, N. B., Fanelli, D., … Regev, A. (2021). Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nature Methods*, *18*(11), 1352–1362. https://doi.org/10.1038/s41592-021-01264-7

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC.

Cable, D. M., Murray, E., Zou, L. S., Goeva, A., Macosko, E. Z., Chen, F., & Irizarry, R. A. (2022). Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, *40*(4), 517–526. https://doi.org/10.1038/s41587-021-00830-w

Caligiore, D., Arbib, M. A., Miall, R. C., & Baldassarre, G. (2019). The super-learning hypothesis: Integrating learning processes across cortex, cerebellum and basal ganglia. *Neuroscience & Biobehavioral Reviews*, *100*, 19–34. https://doi.org/10.1016/j.neubiorev.2019.02.008

Carpenter, A. E., Jones, T. R., Lamprecht, M. R., Clarke, C., Kang, I. H., Friman, O., Guertin, D. A., Chang, J. H., Lindquist, R. A., Moffat, J., Golland, P., & Sabatini, D. M. (2006). CellProfiler: Image analysis software for identifying and quantifying cell phenotypes. *Genome Biology*, *7*(10), R100. https://doi.org/10.1186/gb-2006-7-10-r100

Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., & Zhuang, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. *Science (New York, N.Y.)*, *348*(6233), aaa6090. https://doi.org/10.1126/science.aaa6090

Chinta, S. J., & Andersen, J. K. (2008). Redox imbalance in Parkinson's disease. *Biochimica et Biophysica Acta (BBA) – General Subjects*, *1780*(11), 1362–1367. https://doi.org/10.1016/j.bbagen.2008.02.005

Clifton, K., Anant, M., Aihara, G., Atta, L., Aimiuwu, O. K., Kebschull, J. M., Miller, M. I., Tward, D., & Fan, J. (2023). STalign: Alignment of spatial transcriptomics data using diffeomorphic metric mapping. *Nature Communications*, *14*(1), 8123. https://doi.org/10.1038/s41467-023-43915-7

Dong, R., & Yuan, G.-C. (2021). SpatialDWLS: Accurate deconvolution of spatial transcriptomic data. *Genome Biology*, *22*(1), 145. https://doi.org/10.1186/s13059-021-02362-7

Du, W., Xia, X., Hu, F., & Yu, J. (2024). Extracellular matrix remodeling in the tumor immunity. *Frontiers in Immunology*, *14*, Article 1340634. https://doi.org/10.3389/fimmu.2023.1340634

Elosua-Bayes, M., Nieto, P., Mereu, E., & Gut, I. (2021). SPOTlight: Seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Research*, *49*(9), e50–e50. https://doi.org/10.1093/nar/gkab043

Erö, C., Gewaltig, M.-O., Keller, D., & Markram, H. (2018). A cell atlas for the mouse brain. *Frontiers in Neuroinformatics*, *12*, 84. https://doi.org/10.3389/fninf.2018.00084

Geras, A., Shafighi, S. Darvish, Domżał, K., Filipiuk, I., Rączkowska, A., Szymczak, P., Toosi, H., Kaczmarek, L., Koperski, Ł., Lagergren, J., Nowis, D., & Szczurek, E. (2023). Celloscope: A probabilistic model for marker-gene-driven cell type deconvolution in spatial transcriptomics data. *Genome Biology*, *24*(1), 120. https://doi.org/10.1186/s13059-023-02951-8

Graves, T. L. (2011). Automatic step size selection in random walk metropolis algorithms. *arXiv:*1103.5986.

Haario, H., Saksman, E., & Tamminen, J. (2001). An adaptive metropolis algorithm. *Bernoulli*, *7*(2), 223.

Hu, C., Li, T., Xu, Y., Zhang, X., Li, F., Bai, J., Chen, J., Jiang, W., Yang, K., Ou, Q., Li, X., Wang, P., & Zhang, Y. (2023). CellMarker 2.0: An updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Research*, *51*(D1), D870–D876. https://doi.org/10.1093/nar/gkac947

Hwang, E. S., Hyslop, T., Lynch, T., Ryser, M. D., Weiss, A., Wolf, A., Norris, K., Witten, M., Grimm, L., Schnitt, S., Badve, S., Factor, R., Frank, E., Collyar, D., Basila, D., Pinto, D., Watson, M. A., West, R., Davies, L., … Miller, S. (2024). Active monitoring with or without endocrine therapy for low- risk ductal carcinoma in situ: The COMET randomized clinical trial. *JAMA: The Journal of the American Medical Association*, *333*(11), 972–980. https://doi.org/10.1001/jama.2024.26698

Kleshchevnikov, V., Shmatko, A., Dann, E., Aivazidis, A., King, H. W., Li, T., Elmentaite, R., Lomakin, A., Kedlian, V., Gayoso, A., Jain, M. S., Park, J. S., Ramona, L., Tuck, E., Arutyunyan, A., Vento-Tormo, R., Gerstung, M., James, L., Stegle, O., … Bayraktar, O. A. (2022). Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature Biotechnology*, *40*(5), 661–671. https://doi.org/10.1038/s41587-021-01139-4

Langlieb, J., Sachdev, N., Balderrama, K., Nadaf, N., Raj, M., Murray, E., Webber, J., Vanderburg, C., Gazestani, V., Tward, D., Mezias, C., Li, X., Cable, D., Norton, T., Mitra, P. P., Chen, F., & Macosko,

E. (2023). The cell type composition of the adult mouse brain revealed by single cell and spatial genomics. *bioRxiv*.

Larsson, L., Frisén, J., & Lundeberg, J. (2021). Spatially resolved transcriptomics adds a new dimension to genomics. *Nature Methods*, *18*(1), 15–18. https://doi.org/10.1038/s41592-020-01038-7

Lein, E. S., Hawrylycz, M. J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., A. F. Boe, Boguski, M. S., Brockway, K. S., Byrnes, E. J., Chen, L., Chen, L., Chen, T.-M., Chin, M. Chi, Chong, J., Crook, B. E., Czaplinska, A., Dang, C. N., Datta, S., … Jones, A. R. (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, *445*(7124), 168–176. https://doi.org/10.1038/nature05453

Levy-Jurgenson, A., Tekpli, X., Kristensen, V. N., & Yakhini, Z. (2020). Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer. *Scientific Reports*, *10*(1), Article 18802. https://doi.org/10.1038/s41598-020-75708-z

Li, Y., & Luo, Y. (2024). STdGCN: Spatial transcriptomic cell-type deconvolution using graph convolutional networks. *Genome Biology*, *25*(1), 206. https://doi.org/10.1186/s13059-024-03353-0

Li, B., Zhang, W., Guo, C., Xu, H., Li, L., Fang, M., Hu, Y., Zhang, X., Yao, X., Tang, M., Liu, K., Zhao, X., Lin, J., Cheng, L., Chen, F., Xue, T., & Qu, K. (2022). Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nature Methods*, *19*(6), 662–670. https://doi.org/10.1038/s41592-022-01480-9

Li, H., Zhou, J., Li, Z., Chen, S., Liao, X., Zhang, B., Zhang, R., Wang, Y., Sun, S., & Gao, X. (2023). A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics. *Nature Communications*, *14*(1), 1548. https://doi.org/10.1038/s41467-023-37168-7

Liddelow, S. A., Guttenplan, K. A., Clarke, L. E., Bennett, F. C., Bohlen, C. J., Schirmer, L., Bennett, M. L., Münch, A. E., Chung, W.-S., Peterson, T. C., D. K. Wilton, Frouin, A., Napier, B. A., Panicker, N., Kumar, M., Buckwalter, M. S., Rowitch, D. H., Dawson, V. L., Dawson, T. M., … Barres, B. A. (2017). Neurotoxic reactive astrocytes are induced by activated microglia. *Nature*, *541*(7638), 481–487. https://doi.org/10.1038/nature21029

Liu, X., Tang, G., Chen, Y., Li, Y., Li, H., & Wang, X. (2025). SpatialDeX is a reference- free method for cell- type deconvolution of spatial transcriptomics data in solid tumors. *Cancer Research*, *85*(1), 171–182. https://doi.org/10.1158/0008-5472.CAN-24-1472

Liu, J., Tran, V., Vemuri, V. N. P., Byrne, A., Borja, M., Kim, Y. J., Agarwal, S., Wang, R., Awayan, K., Murti, A., Taychameekiatchai, A., Wang, B., Emanuel, G., He, J., Haliburton, J., Pisco, A. Oliveira, & Neff, N. F. (2023). Concordance of MERFISH spatial transcriptomics with bulk and single-cell RNA sequencing. *Life Science Alliance*, *6*(1), e202201701. https://doi.org/10.26508/lsa.202201701

Ma, Y., & Zhou, X. (2022). Spatially informed cell-type deconvolution for spatial transcriptomics. *Nature Biotechnology*, *40*(9), 1349–1359. https://doi.org/10.1038/s41587-022-01273-7

Mardia, K. (1988). Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. *Journal of Multivariate Analysis*, *24*(2), 265–284. https://doi.org/10.1016/0047-259X(88)90040-1

Markram, H., Toledo-Rodriguez, M., Wang, Y., Gupta, A., Silberberg, G., & Wu, C. (2004). Interneurons of the neocortical inhibitory system. *Nature Reviews Neuroscience*, *5*(10), 793–807. https://doi.org/10.1038/nrn1519

Miao, Z., Balzer, M. S., Ma, Z., Liu, H., Wu, J., Shrestha, R., Aranyi, T., Kwan, A., Kondo, A., Pontoglio, M., Kim, J., Li, M., Kaestner, K. H., & Susztak, K. (2021). Single cell regulatory landscape of the mouse kidney highlights cellular differentiation programs and disease targets. *Nature Communications*, *12*(1), 2277. https://doi.org/10.1038/s41467-021-22266-1

Miller, B. F., Huang, F., Atta, L., Sahoo, A., & Fan, J. (2022). Reference-free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data. *Nature Communications*, *13*(1), 2339. https://doi.org/10.1038/s41467-022-30033-z

Moncada, R., Barkley, D., Wagner, F., Chiodin, M., Devlin, J. C., Baron, M., Hajdu, C. H., Simeone, D. M., & Yanai, I. (2020). Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nature Biotechnology*, *38*(3), 333–342. https://doi.org/10.1038/s41587-019-0392-8

Morvaridi, S., Dhall, D., Greene, M. I., Pandol, S. J., & Wang, Q. (2015). Role of YAP and TAZ in pancreatic ductal adenocarcinoma and in stellate cells associated with cancer and chronic pancreatitis. *Scientific Reports*, *5*(1), 16759. https://doi.org/10.1038/srep16759

Murata, K., Kinoshita, T., Fukazawa, Y., Kobayashi, K., Kobayashi, K., Miyamichi, K., Okuno, H., Bito, H., Sakurai, Y., Yamaguchi, M., Mori, K., & Manabe, H. (2019). GABAergic neurons in the olfactory cortex projecting to the lateral hypothalamus in mice. *Scientific Reports*, *9*(1), 7132. https://doi.org/10.1038/s41598-019-43580-1

Mutiso, F., Pearce, J. L., Benjamin-Neelon, S. E., Mueller, N. T., Li, H., & Neelon, B. (2022). Bayesian negative binomial regression with spatially varying dispersion: Modeling COVID-19 incidence in Georgia. *Spatial Statistics*, *52*, 100703. https://doi.org/10.1016/j.spasta.2022.100703

Mutiso, F., Pearce, J. L., Benjamin-Neelon, S. E., Mueller, N. T., Li, H., & Neelon, B. (2024). A marginalized zero-inflated negative binomial model for spatial data: Modeling covid-19 deaths in Georgia. *Biometrical Journal*, *66*(5), e202300182. https://doi.org/10.1002/bimj.v66.5

Neelon, B., Wen, C.-C., & Benjamin-Neelon, S. E. (2023). A multivariate spatiotemporal model for tracking COVID-19 incidence and death rates in socially vulnerable populations. *Journal of Applied Statistics*, *50*(8), 1812–1835. https://doi.org/10.1080/02664763.2022.2046713

Nielsen, M. F. B., Mortensen, M. B., & Detlefsen, S. (2016). Key players in pancreatic cancer-stroma interaction: Cancer-associated fibroblasts, endothelial and inflammatory cells. *World Journal of Gastroenterology*, *22*(9), 2678. https://doi.org/10.3748/wjg.v22.i9.2678

Rao, A., Barkley, D., França, G. S., & Yanai, I. (2021). Exploring tissue architecture using spatial transcriptomics. *Nature*, *596*(7871), 211–220. https://doi.org/10.1038/s41586-021-03634-9

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software nnvironment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Singh, R., Park, A. K., Hardison, R. C., Zhu, X., & Li, Q. (2023). RETROFIT: reference-free deconvolution of cell-type mixtures in spatial transcriptomics. *bioRxiv*.

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg, Å., Pontén, F., P. I. Costea, Sahlén, P., Mulder, J., Bergmann, O., … Frisén, J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science (New York, N.Y.)*, *353*(6294), 78–82. https://doi.org/10.1126/science.aaf2403

Teschendorff, A. E., Breeze, C. E., Zheng, S. C., & Beck, S. (2017). A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics*, *18*(1), 105. https://doi.org/10.1186/s12859-017-1511-5

Tritsch, N. X., Ding, J. B., & Sabatini, B. L. (2012). Dopaminergic neurons inhibit striatal output through non-canonical release of GABA. *Nature*, *490*(7419), 262–266. https://doi.org/10.1038/nature11466

Vickovic, S., Eraslan, G., Salmén, F., Klughammer, J., Stenbeck, L., Schapiro, D., Äijö, T., Bonneau, R., Bergenstråhle, L., Navarro, J. F., Gould, J., Griffin, G. K., Borg, Å., Ronaghi, M., Frisén, J., Lundeberg, J., Regev, A., & Ståhl, P. L. (2019). High-definition spatial transcriptomics for in situ tissue profiling. *Nature Methods*, *16*(10), 987–990. https://doi.org/10.1038/s41592-019-0548-y

Williams, C. G., Lee, H. J., Asatsuma, T., Vento-Tormo, R., & Haque, A. (2022). An introduction to spatial transcriptomics for biomedical research. *Genome Medicine*, *14*(1), 68. https://doi.org/10.1186/s13073-022-01075-1

Yang, C. X., Sin, D. D., & Ng, R. T. (2024). SMART: Spatial transcriptomics deconvolution using marker-gene-assisted topic model. *Genome Biology*, *25*(1), 304. https://doi.org/10.1186/s13059-024-03441-1

Yu, K.-X., Yuan, W.-J., Wang, H.-Z., & Li, Y.-X. (2024). Extracellular matrix stiffness and tumor-associated macrophage polarization: New fields affecting immune exclusion. *Cancer Immunology, Immunotherapy*, *73*(6), 115. https://doi.org/10.1007/s00262-024-03675-9

Zaborszky, L., Den Pol, A. Van, & Gyengesi, E. (2012). The basal forebrain cholinergic projection system in mice. In *The Mouse Nervous System* (pp. 684–718). Academic Press.

Zeisel, A., Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., J. Van Der Zwan, Häring, M., Braun, E., Borm, L. E., Manno, G. La, Codeluppi, S., Furlan, A., Lee, K., Skene, N.,

Harris, K. D., Hjerling-Leffler, J., Arenas, E., Ernfors, P., Marklund, U., … Linnarsson, S. (2018). Molecular architecture of the mouse nervous system. *Cell*, *174*(4), 999–1014.e22. https://doi.org/10.1016/j.cell.2018.06.021

Zhang, Y., Lin, X., Yao, Z., Sun, D., Lin, X., Wang, X., Yang, C., & Song, J. (2023). Deconvolution algorithms for inference of the cell-type composition of the spatial transcriptome. *Computational and Structural Biotechnology Journal*, *21*, 176–184. https://doi.org/10.1016/j.csbj.2022.12.001