Joint cell type identification in spatial transcriptomics and single-cell RNA sequencing data

Agnieszka Geras[1,2], **Kacper Domżał**[2], **and Ewa Szczurek**[2,3]*

[1]Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland
[2]Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Warsaw, Poland
[3]Institute of AI for Health, Helmholtz Center Munich, Neuherberg, Germany

Abstract

Understanding the intricate composition of tissues in complex living organisms is crucial for unraveling the mechanisms underlying health and disease. This study addresses the challenge of dissecting cell types within tissues by integrating information from two powerful experimental techniques: single-cell RNA-sequencing (scRNA-seq) and spatial transcriptomics (ST). While scRNA-seq offers insights into transcriptional heterogeneity at the cellular level, ST provides spatial information within tissues. Current methods for cell-type annotation in scRNA-seq and mixture decomposition in ST data are often conducted independently, resulting in reduced statistical power and accuracy. To bridge this gap, we propose ST-Assign, a novel hierarchical Bayesian probabilistic model that jointly performs cell-type annotation in scRNA-seq data and cell-type mixture decomposition in ST data. ST-Assign accounts for shared variables such as gene expression profiles and leverages prior knowledge about marker genes, amplifying statistical strength and mitigating experimental noise. The model's excellent performance is demonstrated on simulated and real mouse brain data, showcasing accurate cell-type mixture decomposition and cell-type assignment. In comparison to existing tools, ST-Assign demonstrates superior capabilities, particularly in the task of assigning cell types to individual cells. ST-Assign enables exploring the spatial composition of cell types and holds the potential for enhancing our comprehension of diverse biological systems.

probabilistic model | MCMC sampling | spatial transcriptomics data | cell types | single-cell RNA-seq data

Correspondence: ***szczurek@mimuw.edu.pl***

Introduction

Tissues of complex living organisms are composed of millions of cells that come from a large variety of cell types and act in an orchestrated manner (1). The study of individual cell transcriptomes is pivotal to understanding the biological mechanisms behind health and disease (2).

Specifically, single-cell RNA-sequencing (scRNA-seq) is commonly used to examine the extent of transcriptional heterogeneity between cells and identify novel sub-populations within them (3). Cell type annotation is a vital step in the scRNA-seq data analysis pipeline. It is typically conducted through clustering, followed by manual inspection and expert decision-making, which can, unfortunately, be laborious and susceptible to potential biases (4). As a solution, computational tools have been proposed to facilitate automatic, probabilistic cell-type identification, such as CellAssign (5) and scSorter (6).

While scRNA-seq data lack information on cells' position within the tissue, spatial transcriptomics (ST) technology enables its retention. Nonetheless, the mini-bulk nature of some of popular ST protocols, such as (7) Visium (8) presents the challenge of decomposing hidden cell-type mixtures within each spatial location (spot). To tackle this issue, we have previously developed Celloscope (9), specifically designed to infer the prevalence of each cell type from ST data. Celloscope distinguishes itself from earlier methods for cell type deconvolution in ST spots (10–19) by leveraging prior knowledge of marker genes for cell types. These genes are anticipated to exhibit higher expression in their specific cell types than in other cell types.

When scRNA-seq and ST data originate from the same tissue, an identical set of cell types can be assigned to individual cells in scRNA-seq and deconvoluted in each ST spot. Importantly, variables involved in the data-generating process for these two separate datasets overlap. In particular, marker gene expression profiles characteristic of each cell type rule the observations in both scRNA-seq and ST data. Therefore, a model that would integrate scRNA-seq and ST data based on prior knowledge and such shared variables has the potential to enhance the statistical power of cell-type identification in both data sources. This calls for a novel approach for joint single-cell type identification and cell-type mixture decomposition.

To address this need, here we propose a novel hierarchical Bayesian probabilistic model called ST-Assign, combining signals from two different experimental techniques: (i) single-cell RNA-seq and (ii) spatial transcriptomics. The model offers a computational method to jointly perform the two tasks that until now were carried out independently: (i) cell-type annotation in single-cell RNA-seq data and (ii) cell-type mixture decomposition in spatial transcriptomics spots. In this way, ST-Assign preserves single-cell resolution and, at the same time, benefits from spatial information. Moreover, by extracting the same signal from datasets generated on different platforms, each with distinct yet complementary strengths and weaknesses, the model is

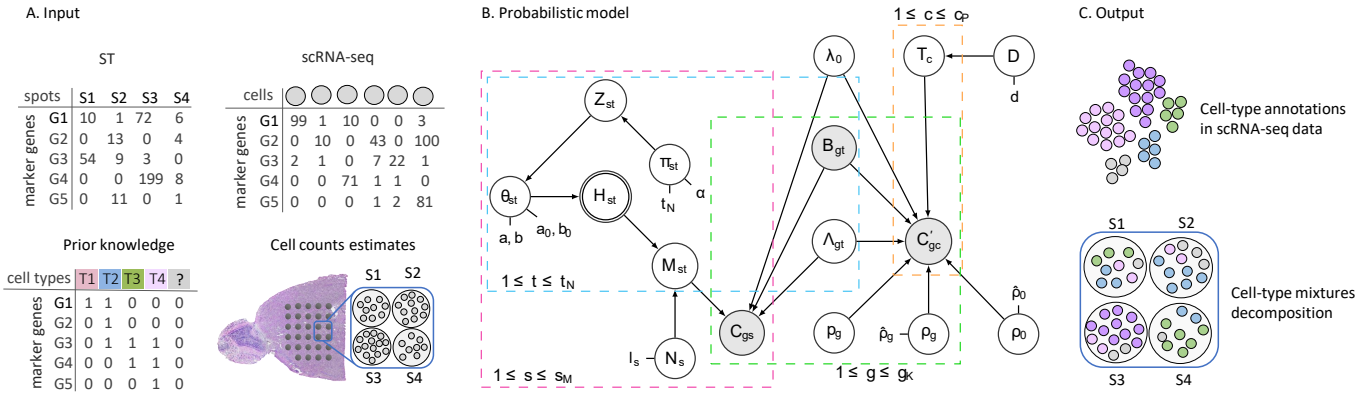

Fig. 1. **A.** ST-Assign processes ST and scRNA-seq data, for selected marker genes. The input also includes prior knowledge about marker genes for cell types (including dummy or unknown type, marked as '?') in the form of a binary matrix, along with the total cell count for each ST spot (optional; estimated from H&E image). H&E image source: 10x Genomics. **B.** The graphical representation of ST-Assign. Double-circled nodes correspond to deterministic variables, gray to observed, and the remaining nodes correspond to hidden (inferred) variables. Arrows represent probabilistic dependencies. ST-Assign variables are described in Supplementary Table S1 and its hyperparameters in Supplementary Table S2. Dotted magenta, blue, green, and orange rectangles (plates) represent variables with different indices but identical distributions in the model. **C.** ST-Assign gives as output cell-type mixture decomposition of ST data and annotations for scRNA-seq data.

less affected by the experimental noise. Importantly, ST-Assign inherits the marker-gene-driven approach to cell type identification from Celloscope (9) and integrates prior domain knowledge about marker genes. This strategy leverages insights derived from multiple independent studies and helps to guide the model inference toward feasible solutions. As a result, ST-Assign effectively analyses unannotated datasets.

We showcase the competitive ST-Assign's performance on real and simulated data, demonstrating the benefit of probabilistic modeling of combined data sources. ST-Assign enables reliable, in-depth studies of single cells of various types of interest, sequenced both using scRNA-seq and ST.

Results

ST-Assign takes as input an unannotated scRNA-seq dataset, an ST dataset taken from the same tissue type, prior knowledge about marker genes encoded as a binary matrix, and, optionally, preliminary estimates for the number of all cells within each ST spot (Fig. 1A).

The probabilistic graphical model behind ST-Assign (Fig. 1B) relies on the expectation that marker genes manifest higher expression in specific cell types across both experiments, facilitating joint decomposition in ST data and cell type assignment in scRNA-seq data. We assume the list of marker genes and their specific cell types are given *a priori*, but gene expression profiles of those cell types are unknown beforehand. As a result, the expression level for each marker gene in each cell type is estimated as a part of the model inference. The model is generative and considers the dependency of the observed gene expression on hidden variables such as the cell type assignment and cell type mixtures' composition. Moreover, ST-Assign is specifically designed to address the effects of differences between the two experimental techniques by introducing dedicated variables. Conditional probability distributions expressing the dependence structure were used to develop the model's inference procedure via Markov chain Monte Carlo, specifically the Metropolis-within-Gibbs algorithm.

The model output constitutes cell-type annotations in scRNA-seq data and the count of each cell type in each ST spot (cell-type decomposition; Fig. 1C).

The study on simulated data proves accurate ST-Assign's performance. To demonstrate the accurate performance of ST-Assign in both cell-type mixture decomposition and annotation of types to single cells, we executed a study utilizing simulated data. The simulated values of the variables, including variables corresponding to gene expression profiles, cross-platform factors, and the overdispersion parameters, were generated using default priors that ensured similarity to real data (Supplementary Methods). To showcase the excellent performance of ST-Assign regardless of the average number of cells per spot, we investigated scenarios with averages of 5, 10, and 25 cells per spot. Furthermore, to evaluate its robustness to small relative differences between the base expression shared across all genes and the over-expression observed for marker genes, we introduced two options for generating the corresponding values: *basic* (mimicking differences observed in real data) and *lower* (where the average over-expression is five times smaller than in the *basic* scenario; see Supplementary Methods). This nuanced approach allows us to assess the model's ability to handle realistic variations in gene expression patterns. For each simulation setting, we generated ten replicates of simulated datasets and evaluated the model's performance in inferring hidden variables. Specifically, to assess the accuracy of estimating the number of cells of each type per spot, we computed the average absolute

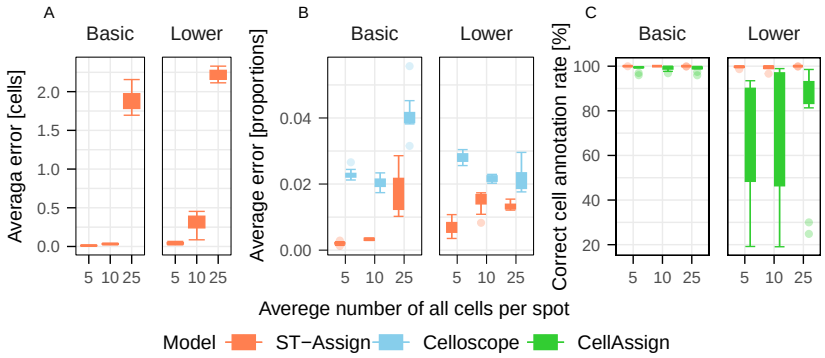

Fig. 2. Excellent performance of ST-Assign on simulated data. Box-plots represent distributions of performance values (y-axis) for different methods (colors) and data simulation scenarios (x-axis and panels). **A.** ST-Assign obtains low average absolute error in cell type count estimation; computed using Equation 1, even for large numbers of cells per spot (x-axis). ST-Assign outperformed Celloscope on simulated data in the cell-type mixture decomposition task measured in proportions (**B**) and CellAssign in the cell-type annotation task (**C**).

error across all spots for each replicate k using the following equation:

$$\epsilon_k = \frac{\sum_{i=1}^{s_M} \sum_{j=1}^{t_N} |M_{ij}^{(k)} - \hat{M}_{ij}^{(k)}|}{t_N s_M}, \tag{1}$$

where $M_{ij}^{(k)}$ represents the true value of the number of cells of type j in spot i for the k-th replicate, $\hat{M}_{ij}^{(k)}$ is the number of cells of type j in spot i for the k-th replicate estimated by ST-Assign, t_N is the number of all considered cell types and s_M is the number of spots in ST data. Notably, ST-Assign is the only approach that estimates $M_{ij}^{(k)}$ and thus was not compared to any previous method in this task. ST-Assign obtained an excellent average error, with a median below 0.5 cell for 5 and 10 cells per spot and below 2.5 cells for 25 cells per spot (Fig. 2A).

Comparison to Celloscope on simulated data. To assess ST-Assign's performance against our previous method Cellsoscope (9) (which infers cell-type proportions instead of the number of cells for each cell type), we normalized the M vectors to obtain proportions (Fig. 2B). ST-Assign outperformed Celloscope in estimating proportions on simulated data, exhibiting significantly lower errors across all considered data simulation setups. This advantage could be attributed to the stronger statistical signal received by ST-Assign, incorporating information from both ST and scRNA-seq data, a feature not present in Celloscope. Furthermore, it is important to highlight that Celloscope relies on the assumption of the negative binomial distribution for the ST data, utilizing two parameters. In contrast, ST-Assign opts for the Poisson distribution, which requires only one. This choice leads to ST-Assign adopting a simpler approach and a distinction in complexity between the two methods in the context of cell-type mixture decomposition.

Cell-type annotation in scRNA-seq data and comparison to CellAssign on simulated data. The excellent performance of ST-Assign was further validated in the task of assigning cell types to individual cells. In this task, ST-Assign demonstrated median accuracy close to 100% for all cells in all the simulation scenarios we considered (Fig. 2C). Importantly, ST-Assign significantly suppressed CellAssign in scenarios characterized by less significant relative differences between base expression and marker gene over-expression (scenario *lower*; Fig. 2C).

ST-Assign enumerates cells of various types in ST and scRNA-seq mouse brain profiling data. To showcase the performance of ST-Assign in concurrent tasks of ST mixture decomposition and cell type assignment on real data, we leveraged data from two distinct sources: (i) ST data from the anterior mouse brain (sagittal section) from (20) previously analyzed using Celloscope (9), and (ii) a subset of scRNA-seq data from an adolescent mouse brain (21). Specifically, ST-Assign was utilized to determine the number of cells belonging to each type in the ST spots and to assign cell types to individual cells in scRNA-seq data. To determine the set of marker genes for each considered cell type we utilized the set of such genes that were identified as markers by (21) and were also expressed in the ST data (20). The subset of single-cell data (ii) was selected to cover a representative, broad range of cell types and serve as a good illustration of ST-Assign's capabilities (see Methods for marker gene and cell type selection).

Cell type identification. Applying ST-Assign to mouse brain data demonstrated its ability to integrate heterogeneous data sources and facilitate a comprehensive analysis of complex biological systems. The obtained cell type decomposition of ST spots (Fig. 3) is in agreement with previous results acquired with Celloscope on the same dataset, although the previously

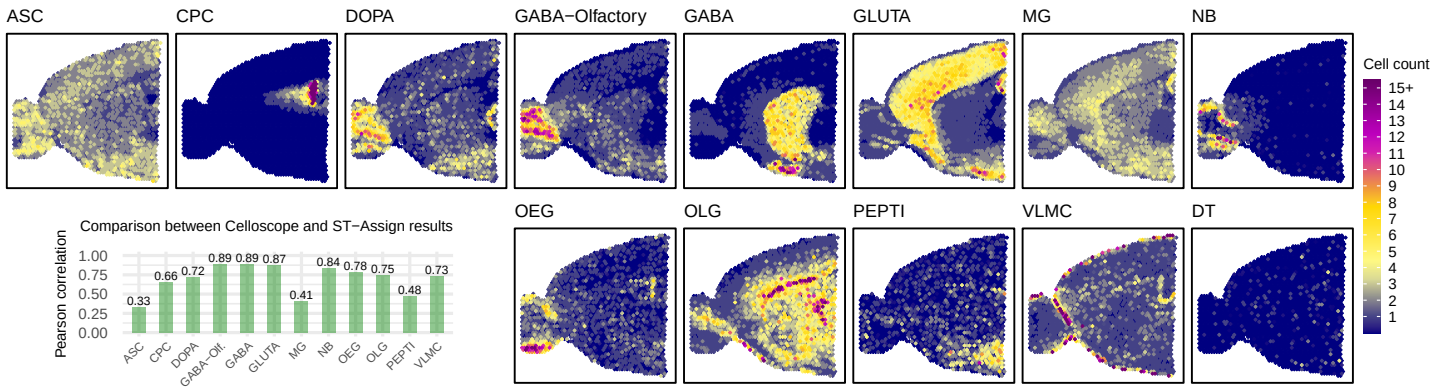

Fig. 3. Spatial composition of cell types in the anterior part of the mouse brain (sagittal section) as inferred by ST-Assign. Heatmaps show the abundance of selected cell types, with dark blue indicating absence, yellow moderate occurrence, and magenta dominance. The cell types are as follows: ASC - Astrocytes, CPC - Choroid plexus epithelial cells, DOPA - Dopaminergic neurons, GABA-Olfactory - GABAergic neurons in the olfactory, GABA - GABAergic neurons in the striatum, GLUT - Glutamatergic neurons, MG - Microglia, NB - Neuroblasts, OEG - Olfactory ensheathing glia, OLG - Oligodendrocytes, PEPTI - Peptigeneric neurons, VLMC - Vascular and leptomeningeal cells, DT - dummy type. The barplot displays the Pearson correlation coefficient between results obtained with ST-Assign and Cellocope for each cell type.

used gene markers and cell types were a bit different (9). ST-Assign was able to identify known mouse brain structures, effectively differentiate between major neuron types (GABAergic and Glutamatergic), and accurately map the spatial distribution of various cell types, including non-neuronal (such as astrocytes and oligodendrocytes).

As another part of the ST-Assign inference, we performed cell type annotation in the accompanying scRNA-seq dataset (Fig. 4A). To present our findings, we employed an independent t-SNE visualization of the single-cell dataset (Fig. 4). The standard single-cell data analyses pipeline and visualization were carried out using Seurat package (22), incorporating to the dataset marker genes associated with the considered cell types from the binary matrix encoding prior knowledge. The cells were color-coded based on the outcomes of the ST-Assign analysis (Fig. 4A), CellAssign results (Fig. 4B), and annotations sourced from the (21) study, which we regard as the ground truth (Fig. 4C).

The results of ST-Assign (Fig. 4A) generally aligned with the visible scRNA-seq clusters. Moreover, the majority of cells were annotated by ST-Assign to their corresponding cell types (Fig. 5A), as defined by (21). However, a portion of cells (507 out of 38,081) was annotated as the dummy type (Fig. 5A), especially among neuroblasts (204) and vascular and leptomeningeal cells (131), potentially indicating separate sub-types.

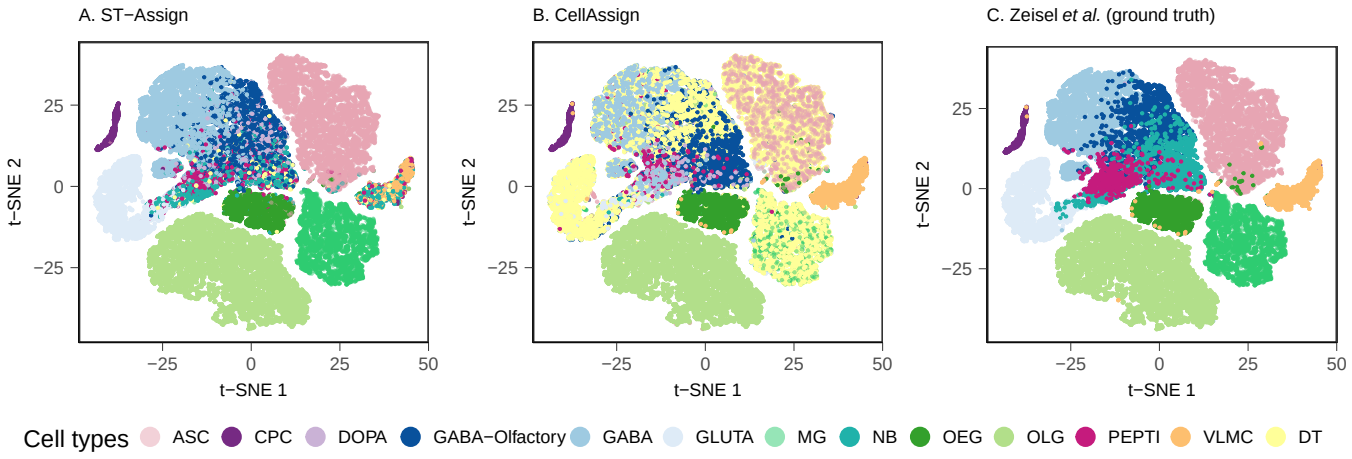

Fig. 4. An independent clustering results and t-SNE visualization, with colors indicating the cell type annotation for each cell by **A.** ST-Assign **B.** CellAssign (5), **C.** (21).

Comparison to Celloscope on real mouse brain data. To compare the performance of ST-Assign and Celloscope (9), we applied Celloscope on the same mouse brain dataset (Supplementary Fig. S1). The Pearson correlation coefficient was computed to assess the concordance between cell counts estimated by ST-Assign and cell-type proportions estimated by Celloscope (Fig. 3). Overall, a high correlation level was observed, ranging approximately between 0.7 and 0.9. However, specific cell types, such as astrocytes (0.33), microglia (0.41), and peptigeneric neurons (0.48), exhibited lower correlation coefficients. Noteworthy, discrepancies in spatial distribution were evident for astrocytes. For instance, Celloscope showed a low frequency of astrocytes in the olfactory bulb, where ST-Assign indicated their presence. ST-Assign's results appear more plausible in this context, considering that astrocytes are expected to be distributed throughout the brain.

Comparison to CellAssign on real mouse brain data. To compare ST-Assign to its competitor in the task of assigning types to single cells, we applied CellAssign to the same single-cell data (Fig. 4B). CellAssign suffered from both a high rate of assignment of cells to the dummy type (40%) and misclassifications between cell types (Fig. 5B). ST-Assign outperformed CellAssign in correctly identifying glutamatergic neurons and microglia (Fig. 5A). For instance, out of 3,312 cells previously labeled as glutamatergic neurons in (21), 2,721 were unexpectedly annotated as dummy type by CellAssign. For the same cell type, ST-Assign correctly identified 2,944 cells and assigned only 49 of them to the dummy type. In contrast to ST-Assign, the performance of CellAssign was particularly low for astrocytes: 6,946 cells were annotated as astrocytes in (21), CellAssign misclassified 4,933 as dummy type. CellAssign, in turn, exhibited commendable performance for vascular and leptomeningeal cells, aligning with annotations from (21) for 98% of the cells. However, when considering the overall performance, as measured by the Adjusted Rand Index (ARI) with reference to annotations from (21), CellAssign fell short with an ARI of 0.43, while ST-Assign displayed exceptional accuracy with an ARI of 0.93.

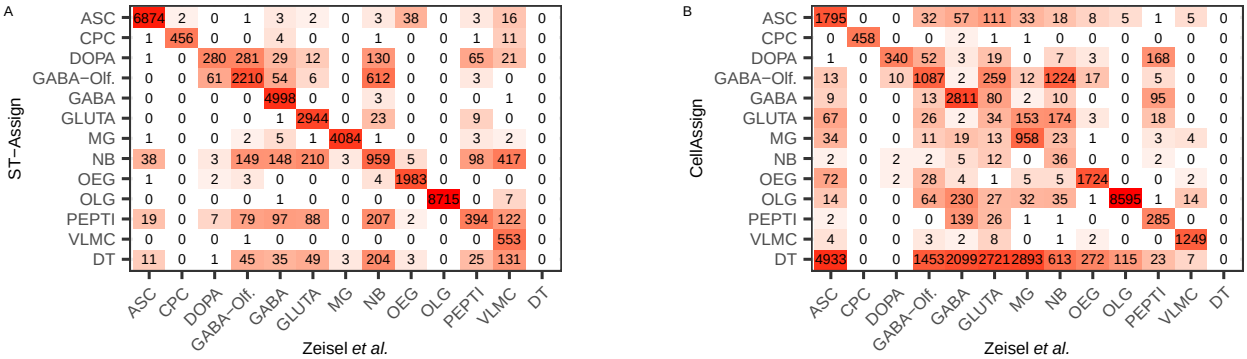

Fig. 5. Confusion matrices compare cell type annotations among the ground truth from the study of (21), and ST-Assign (**A**), and CellAssign (**B**). The cell types are denoted by abbreviations, while the counts indicate the number of discrepancies.

Methods

ST-Assign represents all variables of interest as random variables (Supplementary Table S1). We use a graph-based representation to express the conditional dependence structure between them (Fig. 1B). Since ST-Assign integrates data from the two sources, naturally, the model variables can be divided into three components: (i) variables modeling ST data, (ii) variables modeling scRNA-seq data, and (iii) variables shared between the two components mentioned above.

Let $s \in \mathcal{S} = \{s_1, s_2, \ldots, s_M\}$ index spots, $c \in \mathcal{C} = \{c_1, \ldots, c_P\}$ cells and $g \in \mathcal{G} = \{g_1, g_2, \ldots g_K\}$ genes. We assume t_N cell types are present in the considered tissue, indexed by $t \in \mathcal{T} = \{t_1, t_2, \ldots, t_N\}$. A binary matrix B encodes prior knowledge about marker genes, and B_{gt} takes the value 1 if gene g is a marker for cell type t and 0 otherwise. We additionally consider a so-called *dummy type*, a helper type that does not have any marker genes. For the dummy type t^{dt}, the value $B_{g,t^{dt}} = 0$ for all marker genes g. Consequently, the dummy type gathers all types that have other marker genes than considered *a priori*. Λ_{gt} denotes the over-expression of gene g when it is a marker for type t, and λ_0 represents the base expression profile shared across all the genes. Similarly to our previous model, Celloscope, we assume the increased expression of the marker genes in their specific cell types compared to other cell types. Therefore, we posit that the expected expression measured in a single cell c of type t equals:

$$\lambda_0 + \Lambda_{gt} B_{gt}. \tag{2}$$

Consequently, if gene g is a marker gene for cell type t, its expression in cell c equals $\lambda_0 + \Lambda_{gt}$. Otherwise, $B_{gt} = 0$, and the expected expression is equal to λ_0. Notably, variable Λ_{gt} is shared between the two components modeling ST and scRNA-seq data. Let us notice that ST-Assign considers only a limited subset of genes; therefore, it is easy to check that those considered genes are expressed in the two considered data sets.

ST-Assign serves two purposes: delineating cell-type composition of spots in ST data, represented by the hidden variable M_{st}, and cell-type annotation in scRNA-seq data, defined by the hidden variable T_c. Therefore, M_{st} and T_c are the two central variables in the model. M is a matrix with s_M rows and t_N columns. The value of an element M_{st} is the number of cells in spot s that are of type t. One row of the matrix M, denoted as $M_{s:} = [M_{st_1}, \ldots, M_{st_N}]$ represents the hidden composition of spot s. The number of all cells of all types within spot s is denoted N_s. Clearly, a given row's entries sum up to N_s:

$$N_s = \sum_{t \in \mathcal{T}} M_{st}, \text{ where } N_s \sim \text{Pois}(l_s),$$

and l_s denotes the preliminary estimates of the number of all cells for each spot (either from counting cells in spots using H&E (23) slides or using a common expected number of cells across spots, given their size).

When it comes to modeling scRNA-seq data, T_c is a categorical variable indicating the cell type of cell c:

$$T_c \mid D \sim \text{Categorical}(D) \tag{3}$$

on which we impose Dirichlet prior $D \sim \text{Dirichlet}(d)$ with the concentration parameter d. D and d are vectors of sizes equal to the number of cell types t_N: $D = [D_{t_1},\, D_{t_2},\, \ldots, D_{t_N}]$, $d = [d_{t_1},\, d_{t_2},\, \ldots, d_{t_N}]$.

For gene g, its expression in spot s denoted C_{gs}, is measured as the count of reads from spot s that map to gene g. Likewise, the expression of gene g in cell c, denoted as C'_{gc}, is measured as the count of reads that map to gene g from cell c. The gene expression matrices C and C' are modeled as observed variables. A row $C_{g:} = [C_{gs_1}, \ldots, C_{gs_M}]$ represents the expression of gene g across spots and a column $C_{:s} = [C_{g_1 s}, \ldots, C_{g_K s}]$ represents the gene expression in spot s across marker genes. Analogously, a row $C'_{g:} = [C'_{gc_1}, \ldots, C'_{gc_P}]$ represents the expression of gene g across cells and a column $C'_{:c} = [C'_{g_1 c}, \ldots, C'_{g_K c}]$ represents expression of cell c across marker genes.

Let us denote C^t_{gs} as the number of counts coming from a cell of type t. We assume that the random variable C_{gs} depends on the hidden cell-type composition in spot s, modeled as the number of cells of each type in the spot $[M_{st_1}, \ldots, M_{st_N}]$. Therefore, C_{gs} can be expressed as a sum of independent random variables:

$$C_{gs} \overset{d}{=} M_{st_1} C^{t_1}_{gs} + M_{st_2} C^{t_2}_{gs} + \cdots + M_{st_n} C^{t_N}_{gs},$$

where $\overset{d}{=}$ denotes equality in distribution. Therefore:

$$C_{gs} \mid M_{s:}, B_{g:}, \Lambda_{g:}, \lambda_0 \sim \text{Pois}\left(\sum_t M_{st}(\lambda_0 + B_{gt}\Lambda_{gt})\right). \tag{4}$$

To model M_{st}, we use the multinomial distribution with the number of trials given as the number of all cells in spot s (N_s) and the event probability parameter $H_{s:}$

$$M_{s:} \mid N_s, H_s \sim \text{Multinomial}(N_s, H_{s:}), \tag{5}$$

where $H_{s:}$ is deterministic and computed based on the unnormalized abundance of cell types within spot s:

$$H_{s:} = \frac{\theta_{s:}}{\sum\limits_{t=1}^{t_N} \theta_{st}}. \tag{6}$$

Let us note that $H_{s:}$ can be interpreted as proportions of cell types within spot s and is the primary variable in Celloscope (9). The remaining three parameters, corresponding to the feature allocation model that determines the binary presence of each cell type in each spot: θ_{st}, Z_{st} and π_{st} are defined in precisely the same manner as in Celloscope (9), and identical probability distributions are imposed on them (Supplementary Methods).

To model gene expression counts in scRNA-seq data, we employ the negative binomial distribution, where p_g serves as the success probability parameter for gene g. This parameter accounts for the over-dispersion in gene expression data and is dependent solely on the gene in question. We posit:

$$\begin{aligned} &C'_{gc} \mid T_c = t,\, B_{gt},\, \Lambda_{gt},\, \lambda_0,\, p_g, \rho_g, \rho_0 \sim \\ &NB\left(\frac{1-p_g}{p_g}\left(\rho_0\lambda_0 + \rho_g \Lambda_{gt} B_{gt}\right),\, p_g\right), \end{aligned} \tag{7}$$

where ρ_g denotes cross-platform factor accounting for differences in sequencing depth between the two techniques. Here, we distinguish between gene-dependent factors, different for each marker gene (ρ_g) and a shared factor for non-markers (ρ_0). We assume a Normal prior on ρ_g, with the mean $\hat{\rho}_g$ estimated from the data:

$$\hat{\rho}_g = \frac{\sum\limits_c C'_{gc} \mathbf{1}_{\{\mathbf{C'_{gc}} > \mathbf{0}\}}}{\sum\limits_c \mathbf{1}_{\{\mathbf{C'_{gc}} > \mathbf{0}\}}} \left(\frac{\sum\limits_s \frac{C_{gs}}{l_s} \mathbf{1}_{\{\mathbf{C_{gs}} > \mathbf{0}\}}}{\sum\limits_s \mathbf{1}_{\{\mathbf{C_{gs}} > \mathbf{0}\}}}\right)^{-1}.$$

This corresponds to the ratio of the average nonzero expression of gene g across single cells in scRNA-seq to the average

Cell type	Clusters in (21)	Size
Astrocytes (ASC)	ACTE1, ACTE2	$6,946$
Choroid plexus epithelial cells (CPC)	CHOR	458
Dopaminergic neurons (DOPA)	OBDOP1	354
GABAergic neurons in the striatum (GABA)	MSN1, MSN2, MSN3, MSN4, MSN5, MSN6	$5,375$
GABAergic neurons in the olfactory (GABA-Olfactory)	OBINH1, OBINH2, OBINH3, OBINH4, OBINH5	$2,771$
Glutamatergic neurons (GLUTA)	TEGLU1, TEGLU6, TEGLU10, TEGLU11	$3,312$
Microglia (MG)	MGL1, MGL2	$4,090$
Neuroblasts (NB)	OBNBL1, OBNBL2, OBNBL3	$2,147$
Olfactory ensheathing glia (OEG)	OEC	$2,031$
Oligodendrocytes (OLG)	MOL1	$8,715$
Peptigeneric neurons (PEPTI)	DEINH5, HYPEP1, HYPEP3	601
Vascular and leptomeningeal cells (VLMC)	VLMC1, VLMC2	$1,281$

Table 1. Cell types (their abbreviations in brackets) and their respective clusters in (21) study. In the (21) study, the measured single cells were clustered according to gene expression. These clusters, interpreted as cell types or sub-types, were annotated based on deferentially expressed genes. Columns: analyzed cell types, their respective clusters in (21), and the number of cells in the respective clusters in (21)

nonzero expression of gene g in ST data. Similarly, a Normal prior is imposed on ρ_0, with the mean computed as

$$\hat{\rho}_0 = \frac{\sum_c C'_{gc}}{c_p} \left(\frac{\sum_s \frac{C_{gs}}{l_s}}{s_M} \right)^{-1}.$$

This corresponds to the ratio of the average overall expression of gene g across all single cells to the average expression of gene g in ST data (including 0 expression). The term $\frac{1-p_g}{p_g}$ in Equation 7 serves as a rescaling factor, ensuring that the assumption from Equation 2 holds.

To infer the model's hidden variables, we use the Metropolis-within-Gibbs sampler (Supplementary Methods). Sampling steps (Algorithm 1, Supplementary Methods), the model's equations for full conditional distributions and proposal distributions used in the implementation are provided in Supplementary Methods.

Mouse brain data and marker genes selection. We used ST data from a sagittal section of the anterior mouse brain provided by (20) (accessed via `SeuratData` package (24)). We considered $2,696$ spots. Across them, a minimum of 8, on average 108.5, and a maximum of 144 genes had non-zero expression. The minimal total expression per spot was 16, the maximal was $13,745$, and the average was $1,041.6$.

We employed a scRNA-seq dataset that covered all regions of an adolescent mouse brain (21). We downloaded gene expression data from `http://mousebrain.org/adolescent/`, which also featured information on cell types (clusters) and their specific marker genes. We omitted cells with a total expression below 20, resulting in considering a total of $38,081$ cells. Across cells, a minimum of 4, an average of 28.6, and a maximum of 89 genes had non-zero expression. The maximal total expression per cell was $9,202$, while the average was 307.6.

To define the set of analyzed cell types, we extended the set of cell types selected based on (25), which were previously analyzed in (9) using Celloscope. Here, we included additional types (clusters) from (21). Clusters from (21) that represented brain parts not profiled with ST (e.g., spinal cord) were excluded. The remaining clusters were selected based on the coherent expression of marker genes in ST data (Table 1). A total of 194 genes were chosen and modeled as markers. Marker genes were primarily sourced from (21), supplemented with genes previously by (9).

The scRNA-seq dataset under examination displayed an elevated percentage of zero entries (approximately 85%) compared to ST data (around 44%). The main reason behind the observed differences in the number of zero entries should be attributed to the fact that the analyzed 10x Genomics Visium ST data consist of *mini-bulk* measurements and as such, may exhibit less technical noise compared to scRNA-seq data. Additionally, this difference may arise from various technical and biological factors, including differences in sequencing depth and biological variability. This justifies using the Poisson distribution to model counts in ST data rather than the negative binomial distribution.

Discussion

Here, we introduce a novel probabilistic model for concurrent cell-type annotation in scRNA-seq and mixture decomposition in ST data, based on prior knowledge of marker genes. The evaluation carried out on simulated data demonstrated the ST-Assigns's accurate performance in both tasks. Furthermore, the application of ST-Assign to real mouse brain data showed its utility in the analysis of biological samples.

ST-Assign extends the functionalities of Cellocope (9) to accommodate scRNA-seq data. While leveraging the foundation laid by Cellocope, we have implemented significant modifications, particularly concerning how gene expression given by mixtures of cells in ST data is modeled. Instead of inferring proportions of cell types in each ST spot, modeled as continuous values ranging from 0 to 1, in ST-Assign, we explicitly account for the number of cells (a discrete value) of every type in each ST spot. To achieve this, we utilize the multinomial distribution to model the cell-type composition. Notably, ST-Assign adopts the Poisson distribution to model ST data counts, deviating from Celloscope's use of the negative binomial distribution. This choice results in a representation with fewer parameters. Importantly, compared to Celloscope, the new model is supplied with additional information on marker gene expression measured in single cells. This new information boosted the performance of inferring gene expression profiles for each cell type. In the task of cell-type annotation, ST-Assign surpassed its rival, CellAssign, as evidenced by its notably higher ARI.

Nevertheless, integrating data from two sources may pose challenges due to technical discrepancies. Cross-platform effects, such as unequal effectiveness with which genes are captured or only partial overlap between sequenced cell types and genes, must be addressed. To this end, we introduced dedicated random variables. Specifically, we added cross-platform factor ρ that alleviates unequal sequencing depth, with which genes are measured.

In conclusion, ST-Assign emerges as a valuable tool for the joint analysis of scRNA-seq and ST data, providing a robust solution for simultaneous cell-type annotation and mixture decomposition. Its innovative approach and enhanced capabilities make it a promising asset for understanding complex biological systems at the single-cell level.

Data and code availability. The ST-Assign's implementation and data used in this manuscript are freely accessible on GitHub: https://github.com/szczurek-lab/ST-Assign. The results were visualized using ggplot2 (26).

Funding. This project has received funding from the Polish National Science Centre PRELUDIUM grant no 2021/41/N/ST6/03619. E.S. acknowledges the support from the Polish National Science Centre SONATA BIS grant No. 2020/38/E/NZ2/00305.

Ethics approval and consent to participate. Not applicable.

Competing interests. Projects in Szczurek lab are co-funded by Merck Healthcare. Otherwise, the authors declare that they have no competing interests.

Bibliography

1. E. Armingol et al. Deciphering cell–cell interactions and commun from gene expression. *Nat Rev Gen*, 22(2), 2021. doi: 10.1038/s41576-020-00292-x.
2. D. Lähnemann et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1), 2020.
3. E. Papalexi et al. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat Rev Immunl*, 18(1), 2018.
4. M.D. Luecken et al. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular systems biol*, 15(6), 2019.
5. A.W. Zhang et al. Probabilistic cell–type assignment of scRNA-seq for tumor microenvir profiling. *Nat Met*, 16(10), 2019.
6. H. Guo and J. Li. scSorter: assigning cells to known cell types according to marker genes. *Genome Biology*, 22(1), 2021.
7. 10x Genomics, 2023. www.10xgenomics.com/products/spatial-gene-expression.
8. P. Ståhl et al. Visualization and analysis of gene expression in tissue sections by st. *Science*, 353, 2016.
9. A. Geras et al. Celloscope: a probabilistic model for marker-gene-driven cell type deconvolution in spatial transcriptomics data. *Genome Biology*, 24(1), 2023.
10. D. Cable et al. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotech*, 40(4), 2022.
11. A. Andersson et al. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Communications Biology*, 3(1), 2020.
12. M. Elosua-Bayes et al. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Research*, 49(9), 2021.
13. P. Danaher et al. Advances in mixed cell deconvolution enable quantif of cell types in ST data. *Nat Com*, 13(1), 2022.
14. V. Kleshchevnikov et al. Cell2location maps fine-grained cell types in spatial transcrip. *Nat Biotech*, 40(5), 2022.
15. J. Hu et al. Statistical and machine learning methods for spatially resolved transcriptomics with histology. *Computational and Structural Biotechnology Journal*, 19, 2021.
16. R. Dong et al. SpatialDWLS: accurate deconvolution of spatial transcriptomic data. *Genome Biology*, 22(1), 2021.
17. T. Biancalani et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nature Methods*, 18(11), 2021.
18. T. Chu et al. Cell type and gene expression deconvol with BayesPrism enables Bayesian integrative analysis across bulk and scRNA-seq in oncology. *Nat Can*, 3(4), 2022.
19. B.F. Miller et al. Reference–free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data. *Nature Commun*, 13(1), 2022.
20. 10x Genomics. Mouse brain serial section 1 (sagittal-anterior), 2022. https://support.10xgenomics.com/spatial-gene-expression/datasets/1.1.0/V1_Mouse_Brain_Sagittal_Anterior, Accessed: 2022-04-16.
21. A Zeisel et al. Molecular Architecture of the Mouse Nervous System. *Cell*, 174(4), 2018.
22. Y. Hao et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13), 2021.
23. Animgoeth. szczurek-lab/qupath-spot-utils, 2023.
24. Rahul Satija, Paul Hoffman, and Andrew Butler. *SeuratData: Install and Manage Seurat Datasets*, 2019.
25. M. Ximerakis et al. Single-cell transcriptomic profiling of the aging mouse brain. *Nature Neuroscience*, 22(10), 2019.
26. H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, 2016. ISBN 978-3-319-24277-4.

Joint cell type identification in spatial transcriptomics and single-cell RNA sequencing data

Supplementary Information

Agnieszka Geras[1,2], Kacper Domżał[2], Ewa Szczurek[2,3*]

[1]Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland,
[2]Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Warsaw, Poland,
[3]Institute of AI for Health, Helmholtz Center Munich, Neuherberg, Germany, Neuheberg, Germany
*szczurek@mimuw.edu.pl

Supplementary Tables

Observed variables		
C_{gs}	Negative binomial	gene expression for gene g in spot s
C'_{gc}	Negative binomial	gene expression for gene g in cell c
B_{gt}	-	indicator whether gene g is a marker for cell type t
Hidden variables		
θ_{st}	Gamma	unnormalised abundance of cell type t in spot s
λ_0	-	base gene expression level, shared across all genes and cell types
Λ_{gt}	-	over-expression for a marker gene g in its specific cell type t
N_s	Poisson	number of cells in spot s
ρ_g	-	gene-dependent cross-platform factor for gene g
ρ_0	-	shared cross-platform factor for non-markers
N_s	Poisson	number of cells in spot s
$M_{s:}$	Multinomial	cell type composition in spot s
p_g	-	success probability parameter for gene g
Z_{st}	Bernoulli	indicator variable stating presence of cell type t in spot s
π_{st}	Beta	prior for the probability of presence of cell type t in a spot s
T_c	Categorical	cell type of cell c
D	Dirichlet	prior distribution on T_c
Deterministic variables		
h_{st}	Dirichlet	fraction of cell type t in a spot s

Table S1: Observed, hidden, and deterministic variables in ST-Assign, their corresponding probability distributions, and their interpretations.

Hyperparameters	
α	adjustment for the average number of cell types present in ST spot
a	rate parameter for cell types present in ST spot
b	scale parameter for cell types present in ST spot
a_0	rate parameter for cell types absent in ST spot
b_0	scale parameter for cell types present in ST spot
l_s	estimate for the total number of cells in ST spot
d	concentration parameters for cell types in scRNA-seq data

Table S2: Description of hyperparameters of ST-Assign variables.

Supplementary Methods

The Metropolis-within-Gibbs sampler is a Markov chain Monte Carlo (MCMC) algorithm which is a combination of the Gibbs sampler and the Metropolis-Hastings method. Suppose the graphical model contains variables $x_1, \ldots, x_n$. Let $MB(x_i)$ denote the Markov Blanket of x_i, i.e., the set containing its parents, children, and co-parents. Then:

$$p(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = p\left(x_i \mid \mathrm{MB}(x_i)\right), \tag{S1}$$

i.e., 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.

The iterative sampling procedure is as follows: firstly, starting values for $x_1^{(0)}, \ldots, x_n^{(0)}$ are randomly initialized, then, in a given iteration j ($j = 1, 2, \ldots, J$, where J denotes the number of iterations), we take every variable x_i, $i = 1, ..., n$ one by one, in some arbitrary ordering and for each given variable x_i, its value is sampled given the values of the variables from the Markov blanket $\mathrm{MB}\left(x_i^{(j-1)}\right)$ from the previous iteration. As a result each x_i is updated iteratively, up until convergence. There are two options for updating the value of x_i in the j-th iteration:

1. If $p\left(x_i \mid \mathrm{MB}\left(x_i^{(j-1)}\right)\right)$ can be expressed in a closed form, a new value x_i^* is sampled directly from $p\left(x_i \mid \mathrm{MB}\left(x_i^{(j-1)}\right)\right)$ and $x_i^{(j)}$ is set to x_i^*.

2. In case we only known a function g proportional to $p\left(x_i \mid \mathrm{MB}(x_i)\right)$

$$p\left(x_i \mid MB\left(x_i^{(j-1)}\right)\right) \propto g(x_i),$$

   we perform a single Metropolis-Hastings accept-reject step (MH-single-step procedure in Algorithm 1). A candidate value x_i^* is sampled from a predefined proposal distribution $q\left(\cdot \mid x_i^{(j-1)}\right)$, and then either accepted with probability given by

$$r = \min\left(1, \frac{g\left(x^*\right) q\left(x^{(j-1)} \mid x^*\right)}{g\left(x^{(j-1)}\right) q\left(x^* \mid x^{(j-1)}\right)}\right) \tag{S2}$$

   and $x_i^{(j)} \leftarrow x_i^*$, or the previous value is held: $x_i^{(j)} \leftarrow x_i^{(j-1)}$. After updating x_i, we immediately use the new value for sampling other variables.

Inference of the model's parameters

For inferring ST-Assign's hidden variables, we use the Metropolis-within-Gibbs sampler. Sampling steps for ST-Assign are provided in Algorithm 1, where variables correspond to the variables in the model (described in the main text), and additionally, variable J corresponds to the number of iterations.

Below, we provide the model's equations for full conditional distributions and proposal distributions used in the implementation. When it comes to parameters π, Z, D, T, we are in possession of the explicit formulas to sample from. In the case of the variables: Λ, λ_0, θ, ρ, ρ_0 p, N, M, only functions proportional to the full conditional distributions are known. These functions are used to compute the acceptance probability. To avoid numerical errors, all variables are sampled in the log-space.

Algorithm 1 Metropolis-Hastings within Gibbs sampler for inferring ST-Assign parameters

Require: B, C, C', J

procedure MH-WITHIN-GIBBS(B, C, C', J)
  for $j \leftarrow 1$ to J **do**
    Sample $\pi^{(j)} \sim p(\pi \mid \mathrm{MB}(\pi^{(j-1)}))$
    Sample $Z^{(j)} \sim p(Z \mid \mathrm{MB}(Z^{(j-1)})$
    $\Theta^{(j)} \leftarrow$ MH-SINGLE-STEP($\Theta^{(j-1)}$)
    $[M^{(j)}, N^{(j)}] \leftarrow$ MH-SINGLE-STEP($M, N \mid M^{(j-1)}, N^{(j-1)}$)
    $\Lambda^{(j)} \leftarrow$ MH-SINGLE-STEP($\Lambda^{(j-1)}$)
    $\lambda_0^{(j)} \leftarrow$ MH-SINGLE-STEP($\lambda_0^{(j-1)}$)
    $p^{(j)} \leftarrow$ MH-SINGLE-STEP($p^{(j-1)}$)
    $\rho_g \leftarrow$ MH-SINGLE-STEP($\rho_g^{(j-1)}$)
    $\rho_0 \leftarrow$ MH-SINGLE-STEP($\rho_0^{(j-1)}$)
    Sample $T_c^{(j)} \sim p(T_c \mid \mathrm{MB}(T_c^{(j-1)}))$
    Sample $D^{(j)} \sim p(D \mid \mathrm{MB}(D^{(j-1)}))$
  end for
  return $\pi, Z, \Theta, \Lambda, \lambda_0, p, N, M, T, D, \rho, \rho_0$
end procedure

procedure MH-SINGLE-STEP($x^{(j-1)}$)
  Draw proposal sample $x^* \sim q(. \mid x^{(j-1)})$
  Evaluate acceptance probability r
  Generate $u \sim U[0, 1]$
  if $u \leq r$ **then**
    Accept proposed sample: $x^{(j)} \leftarrow x^*$
  else
    Reject proposed sample and replicate previous sample: $x^{(j)} \leftarrow x^{(j-1)}$
  end if
  return $x^{(j)}$
end procedure

Sampling π

$$\pi_{st} \sim Beta\left(\frac{\alpha}{t_N}, 1\right) \tag{S3}$$

$$P\left(\pi_{st} \mid Z_{st}\right) \propto \text{Beta}\left(\pi_{st} \mid \frac{\alpha}{t_N}, 1\right) \text{Bernoulli}\left(Z_{st} \mid 1, \pi_{st}\right) = \text{Beta}\left(\pi_{st} \mid \frac{\alpha}{t_N} + Z_{st}, 2 - Z_{st}\right) \tag{S4}$$

Sampling Z

$$Z_{st} \sim Bernoulli(1, \pi_{st}) \tag{S5}$$

As Z_{st} is a binary random variable, it suffices to consider its two possible values: 0 and 1.

$$P(Z_{st} \mid \theta_{st}, \pi_{st}) \propto P(Z_{st} \mid \pi_{st}) P(\theta_{st} \mid Z_{st}) = \text{Bernoulli}(Z_{st} \mid 1, \pi_{st}) \Gamma(\theta_{st} \mid a, b)^{Z_{st}} \Gamma(\theta_{st} \mid a_0, b_0)^{1-Z_{st}}$$

$$P(Z_{st} = 1 \mid \pi_{st}, \theta_{st}) \propto \pi_{st} \theta_{st}^{a-1} e^{-b\theta_{st}} = A$$

$$P(Z_{st} = 0 \mid \pi_{st}, \theta_{st}) \propto (1 - \pi_{st}) \theta_{st}^{a_0-1} e^{-b_0\theta_{st}} = B$$

$$P(Z_{st} = 1 \mid \pi_{st}, \theta_{st}) = \frac{A}{A+B} \quad P(Z_{st} = 0 \mid \pi_{st}, \theta_{st}) = 1 - \frac{A}{A+B} = \frac{B}{A+B}$$

Sampling Θ

$$\begin{aligned} \theta_{st} \mid Z_{st} = 1 &\sim Gamma(a, b) \\ \theta_{st} \mid Z_{st} = 0 &\sim Gamma(a_0, b_0) \end{aligned} \tag{S6}$$

We update each spot independently. For a selected spot s:

$$p(\Theta_{s:} \mid N_s, M_{s:}, Z_{s:}) \propto \text{Multinom}\left(M_{s:} \mid N_s, H_{s:}\right) \prod_{\substack{t \in \mathcal{T}: \\ Z_{st} \neq 0}} \Gamma(\theta_{st} \mid a, b) \prod_{\substack{t \in \mathcal{T}: \\ Z_{st} = 0}} \Gamma(\theta_{st} \mid a_0, b_0). \tag{S7}$$

Full conditional distribution for $[N_s,\ M_s]$

Recall that the relationship between the variables N_s and M_n is expressed by the following formula:

$$N_s = \sum_{t \in \mathcal{T}} M_{st}, \tag{S8}$$

therefore, these variables are updated together in one block. For a selected spot s:

$$p(N_s, M_{s:}, \mid C_{:s}, B, \Theta_{s:}, \Lambda, l_s) \propto p(N_s \mid l_s) p(M_{s:} \mid N_s, \Theta_{s:}) \prod_{g \in \mathcal{G}} p(C_{gs} \mid \Lambda_{g:}, \lambda_0, B_{g:}, M_{s:}) =$$

$$\text{Pois}(N_s \mid l_s)\ \text{Multinom}\left(M_{s:} \mid N_s, H_{s:}\right) \prod_{g \in \mathcal{G}} \text{Pois}\left(C_{gs} \mid \sum_t M_{st}(\lambda_0 + B_{gt}\Lambda_{gt})\right).$$

Sampling T_c

For a selected cell c:

$$p(T_c \mid C'_{:c}, B, D_c, \Lambda, \lambda_0, p, \rho) \propto p(T_c \mid D) \prod_{g \in \mathcal{G}} p(C'_{gc} \mid \Lambda, \lambda_0, p_g, B, T_c, \rho_g, \rho_0) =$$

$$\text{Cat}(T_c \mid D) \prod_{g \in \mathcal{G}} \text{NB}\left(C'_{gc} \mid \frac{1-p_g}{p_g}(\rho_0\lambda_0 + \rho_g\Lambda_{gT_c}B_{gT_c}), p_g\right) =$$

$$D_{T_c} \prod_{g \in \mathcal{G}} \text{NB}\left(C'_{gc} \mid \frac{1-p_g}{p_g}(\rho_0\lambda_0 + \rho_g\Lambda_{gT_c}B_{gT_c}), p_g\right).$$

Technically, to sample from the categorical distribution while avoiding leaving the log-space, we use the Gumbel-max trick [1].

Sampling D

We leverage the conjugate prior property of the Dirichlet distribution with respect to the categorical distribution, and we employ the Categorical-Dirichlet model.

$$p(D \mid T_:, d) \propto \text{Dirichlet}(D \mid d) \prod_{c \in \mathcal{C}} \text{Categorical}(T_c \mid D)$$
$$= \text{Dirichlet}(D_{t_1}, D_{t_2}, \ldots, D_{t_N} \mid d_{t_1}, d_{t_2}, \ \ldots, d_{t_N}) \prod_{c \in \mathcal{C}} \text{Categorical}(T_c \mid D_{t_1}, D_{t_2}, \ldots, D_{t_N}) \propto$$
$$\prod_{j=1}^{t_N} D_{t_j}^{d_{t_j}-1} \prod_{c \in \mathcal{C}} \prod_{j=1}^{t_N} D_{t_j}^{\mathbb{1}\{T_c = t_j\}} = \prod_{j=1}^{t_N} D_{t_j}^{d_{t_j}-1+\sum_{c \in \mathcal{C}} \mathbb{1}\{T_c = t_j\}}$$

Therefore,

$$p(D \mid T_:, d) = \text{Dirichlet}(d'_{t_1}, \ d'_{t_2}, \ \ldots, d'_{t_N}),$$

where:

$$d'_{t_j} = d_{t_j} + \sum_{c \in \mathcal{C}} \mathbb{1}\{T_c = t_j\}.$$

Full conditional distribution for ρ_g

For a gene g:

$$p(\rho_g \mid C'_{g:}, \Lambda, \lambda_0, p_g, B, T, \rho_0, \hat{\rho}_g) \propto N(\hat{\rho}_g, 0.1) \prod_{c \in \mathcal{C}} \text{NB}\left(C'_{gc} \mid \frac{1-p_g}{p_g}(\rho_0 \lambda_0 + \rho_g \Lambda_{gT_c} B_{gT_c}), p_g\right).$$

Full conditional distribution for ρ_0

$$p(\rho_0 \mid C', \Lambda, \lambda_0, p_g, B, T, \rho_g, \hat{\rho}_0) \propto N(\hat{\rho}_0, 0.1) \prod_{\substack{g \in \mathcal{G} \\ c \in \mathcal{C}}} \text{NB}\left(C'_{gc} \mid \frac{1-p_g}{p_g}(\rho_0 \lambda_0 + \rho_g \Lambda_{gT_c} B_{gT_c}), p_g\right).$$

Full conditional distribution for Λ_{gt}

$$p\left(\Lambda_{gt} \mid C_{gt}, C'_{gt}, B_{g:}, \Lambda_{g:}, \lambda_0, p_g, M, \rho, \rho_0, T\right) \propto$$
$$\prod_{s \in \mathcal{S}} p(C_{gs} \mid M_{s:}, \Lambda_{g:}, \lambda_0, B_{g:}) \prod_{c \in \mathcal{C}:\, T_c = t} p(C'_{gc} \mid \Lambda_{gt}, \lambda_0, p_g, B_{gt}, \rho_g, \rho_0) =$$
$$\prod_{s \in \mathcal{S}} \text{Pois}\left(C_{gs} \mid \sum_{t'} M_{st'}(\lambda_0 + \Lambda_{gt'} B_{gt'})\right) \prod_{c \in \mathcal{C}:\, T_c = t} \text{NB}\left(C'_{gc} \mid \frac{1-p_g}{p_g}(\rho_0 \lambda_0 + \rho_g \Lambda_{gt} B_{gt}), p_g\right). \tag{S9}$$

Full conditional distribution for λ_0

$$p\left(\lambda_0 \mid C, C', B, \Lambda, p, M, T, \rho, \rho_0\right) \propto$$
$$\prod_{g\in\mathcal{G}}\left[\prod_{s\in\mathcal{S}} p(C_{gs} \mid M_{s:}, \Lambda_{g:}, \lambda_0, B_{g:}) \prod_{c\in\mathcal{C}} p(C'_{gc} \mid \Lambda_{gT_c}, \lambda_0, p_g, B_{gT_c}, \rho_g, \rho_0)\right] =$$
$$\prod_{g\in\mathcal{G}}\left[\prod_{s\in\mathcal{S}} \mathrm{Pois}\left(C_{gs} \mid \sum_t M_{st}(\lambda_0 + \Lambda_{gt}B_{gt})\right) \prod_{c\in\mathcal{C}} \mathrm{NB}\left(C'_{gc} \mid \frac{1-p_g}{p_g}(\rho_0\lambda_0 + \rho_g\Lambda_{gT_c}B_{gT_c}), p_g\right)\right]. \tag{S10}$$

Full conditional distribution for p_g

For a gene g:

$$p\left(p_g \mid C'_{g:}, B_{g:}, \Lambda_{g:}, \lambda_0, T, \rho_g, \rho_0\right) \propto \prod_{c\in\mathcal{C}} p(C'_{gc} \mid \Lambda_{gT_c}, \lambda_0, p_g, B_{gT_c}, T_c, \rho_g, \rho_0).$$

The proposal distributions are analogous to proposal distributions used in Celloscope [2] except for N_s and $M_{s:}$ variables, for which new values have to be proposed together due to the relationship between them (Equation S8).

Proposal distribution and Hastings ratio for [N_s, M_s]

We propose a three-step procedure for proposing new values of $M_{s:}$ and N_s. First, to sample a new value N_s^* for the variable N_s, we choose the ceiling of the truncated normal distribution as the proposal distribution:

$$N_s^* \mid N_s^{\mathrm{prev}} \sim \lceil T \rceil,\ T \sim \mathrm{TN}(N_s^{\mathrm{prev}}, \sigma_N),$$

where σ_N controls step size and N_s^{prev} is the value from the previous iteration.
Let $f_{\lceil \mathrm{TN} \rceil}(\cdot \mid N_s^{\mathrm{prev}}, \sigma_N)$ denote the density of $\lceil T \rceil$. Second, we sample a random vector $\epsilon \sim N(0, \sigma_M \mathbb{I})$, where σ_M controls the step size. Let us denote

$$\mu_s(M_{s:}, \epsilon) = \frac{M_{s:} + \epsilon_t}{\sum_{t\in\mathcal{T}}(M_{st} + \epsilon_t)}.$$

Third, we sample new values $M_{s:}^*$ of the variable $M_{s:}$ from the multinomial distribution with the number of trials parameter equal to N_s^* and the event probability parameter computed as a normalized cell-type composition from the previous iteration perturbed with ϵ:

$$M_{s:}^* \sim \mathrm{Multinomial}(N_s^*, \mu_s(M_{s:}^{\mathrm{prev}}, \epsilon)).$$

To sum up, the procedure to acquire new values for $[N_s, M_{s:}]$ is as follows:

1. Draw a proposal sample $N_s^* \sim f_{\lceil \mathrm{TN} \rceil}(\cdot \mid N_s^{\mathrm{prev}}, \sigma_N)$.
2. Draw $\epsilon \sim \mathrm{N}(0, \sigma_M \mathbb{I})$.
3. Draw $M_{s:}^* \sim \mathrm{MN}(N_s^*, \mu_s(M_{s:}^{\mathrm{prev}}, \epsilon))$.

Finally, the Hastings ratio needs to be computed:

$$\frac{q(M_{s:}^{\text{prev}}, N_s^{\text{prev}} \mid M_{s:}^*, N_s^*)}{q(M_{s:}^*, N_s^* \mid M_{s:}^{\text{prev}}, N_s^{\text{prev}})} =$$
$$\frac{f_{\lceil \text{TN} \rceil}(N_s^{\text{prev}} \mid N_s^*, \sigma_N)}{f_{\lceil \text{TN} \rceil}(N_s^* \mid N_s^{\text{prev}}, \sigma_N)} \frac{f_{\text{MN}}\left(M_{s:}^{\text{prev}} \mid N_s^{\text{prev}}, \mu_s(M_{s:}^*, \epsilon^{\text{prev}})\right)}{f_{\text{MN}}\left(M_{s:}^* \mid N_s^*, \mu_s(M_{s:}^{\text{prev}}, \epsilon^*)\right)},$$

where f_{MN} denotes the probability mass function of the multinomial distribution, ϵ^* and ϵ^{prev} denote the values of ϵ, which perturbed the event probability parameter in the current and previous iteration, respectively.

Synthetic data simulation

We set $s_M = 500$ (number of spots), $c_P = 2500$ (number of cells), $t_N = 7$ (number of cell types including the dummy type, which corresponds to the last type on the list of types), $a = 10$, $b = 1$, $a_0 = 0.1$, $b_0 = 1$, $\lambda_0 = 0.5$, $\rho_0 = 1.2$, $D = (1, 1, 1, 1, 1, 1, 0.2)$ (i.e., the concentration parameter for the dummy type is relatively smaller, as we expect that a minority of cell types is unknown), $\alpha = 0.5t_N$ and the total number of marker genes $g_K = 160$, with their numbers across seven distinct cell types delineated as $11, 15, 20, 27, 37, 50, 0$, respectively. To sample gene expression profiles Λ, we first calculate the average gene expression of marker genes for cell types found in a scRNA-seq dataset on the mouse brain cortex [3]. The obtained values were bootstrapped to get values of Λ for all marker genes in the *basic* scenario. To sample Λ in the scenario with *lower* relative difference between λ_0 and Λ, Λ values were sampled to be, on average, five times smaller than those in the *basic* scenario, while keeping the same λ_0 values. ρ_g was influenced by the gene expression profiles: genes with higher average expression levels were assigned higher ρ_g values. Additionally, a threshold was applied to ensure the generated ρ_g values remained within a biologically plausible range. The values for p_g were sampled from $Unif(0, 1)$, π according to Equation S3, Z according to Equation S5 and Θ according to Equation S6. We computed H based on Θ, from Equation 6 (main text). The values of M and N were sampled according to Equation 5 (main text), and the mean value of N was dependent on the data simulating scenario, each corresponding to the expected total number of cells in each spot, and equal to 5, 10, 25, respectively. T_c was simulated according to the distribution provided by Equation 3 (main text). After all values for parameters have been established, for each of the six considered scenarios, ten replicates were sampled from the generative model: Equation 4 (main text) and Equation 7 (main text). As a result, we obtained 60 synthetic datasets.

Supplementary Text

In the following section, we outline the run settings that serve for the utilization of ST-Assing, Celloscope, and CellAssign.

Running ST-Assign

ST-Assign's run settings were configured as follows: 150,000 iterations were executed, with a burn-in period of 120,000 iterations. The specific parameter values were set as $a = 10$, $b = 1$, $a_0 = 0.1$, $b_0 = 1$, and $\alpha = 5$. To determine the initial values and step sizes for the Λ parameter, a data-driven approach was employed. For each gene, the initial value of $\Lambda_{g:}$ was set as maximum out of 2 and the average of non-zero entries in $C_{g:}$ divided by 10. The step sizes for Λ parameters were then calculated as one-fifth of these initial values. The step sizes for the rest of the parameters are provided in Table S3. To impose shrinkage prior to the proportion of the *dummy type*, we manually set $Z_{\cdot t} = 0$, where t here denotes *the dummy type*.

variable	step size
θ	0.09
N_s	3
$M_{s:}$	2
λ_0	0.5
p_g	0.5
ρ_0	1
ρ_g	5

Table S3: The values of step sizes used for proposal distributions in updating variables through a single accept-reject Metropolis-Hastings step.

Running Celloscope

The run settings for Celloscope mirrored those used for the sagittal mouse brain in our previous work about Celloscope. However, only one chain was executed (100,000 iterations with 80,000 burn-in period).

Running CellAssign

We ran CellAssign using the following commands:

```
calculated_sum_factors = calculateSumFactors(C_gc)
fit <- cellassign(exprs_obj = C_gc, marker_gene_info = B,
    s = calculated_sum_factors, learning_rate = 10^(-1), shrinkage = TRUE).
```

`C_gc` denotes scRNA-seq data. `B` denotes a binary matrix with prior knowledge about marker genes. The learning rate was set to 0.1.

Supplementary Figures

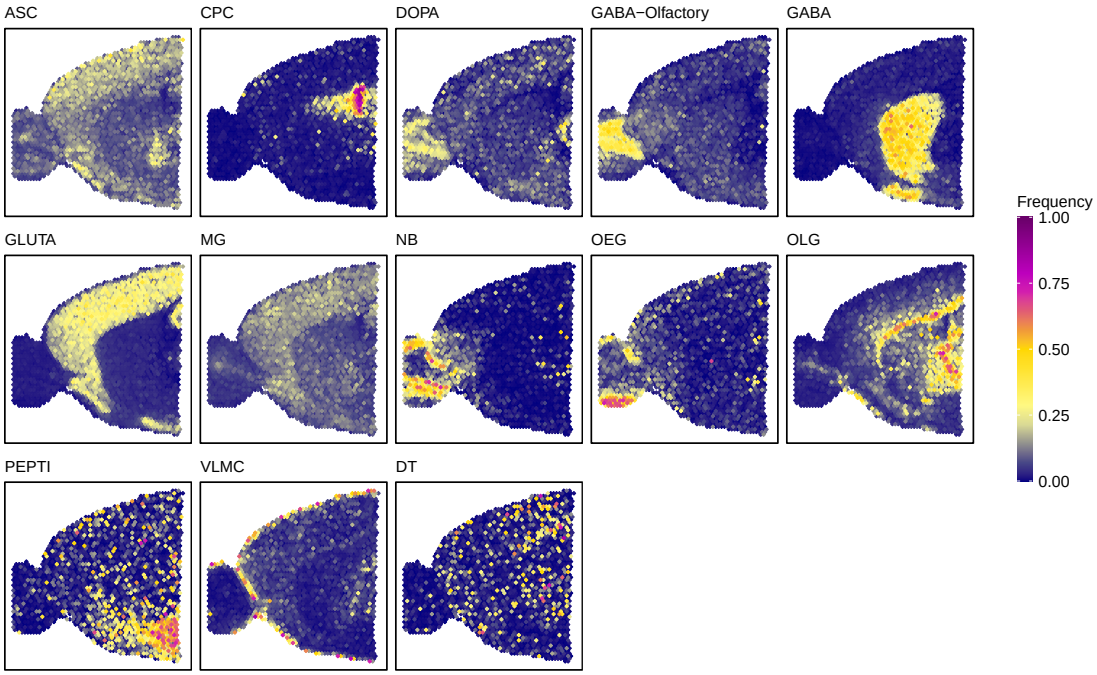

Figure S1: **Spatial composition of cell types in the anterior part of the mouse brain (sagittal section) as inferred by Celloscope.** Heatmaps show the abundance of selected cell types, with dark blue indicating absence, yellow moderate occurrence, and magenta dominance. The cell types are as follows: ASC - Astrocytes, CPC - Choroid plexus epithelial cells, DOPA - Dopaminergic neurons, GABA-Olfactory - GABAergic neurons in olfactory, GABA - GABAergic neurons, GLUT - Glutamatergic neurons, MG - Microglia, NB - Neuroblasts, OEG - Olfactory ensheathing glia, OLG - Oligodendrocytes, PEPTI - Peptigeneric neurons, VLMC - Vascular and leptomeningeal cells, DT - dummy type.

References

[1] R. Adams. The gumbel-max trick for discrete distributions, 2023. `https://lips.cs.princeton.edu/the-gumbel-max-trick-for-discrete-distributions/`, Accessed: 2023-04-21.

[2] A. Geras et al. Celloscope: a probabilistic model for marker-gene-driven cell type deconvolution in spatial transcriptomics data. *Genome Biology*, 24(1), 2023.

[3] B. Tasic et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neurs*, 19(2), 2016.