Patterns

MUSTANG: Multi-sample spatial transcriptomics data analysis with cross-sample transcriptional similarity guidance

Highlights

- MUSTANG is a spot deconvolution tool for multi-sample ST
- MUSTANG is a reference-free spot deconvolution tool
- MUSTANG allows both intra-sample and inter-sample information sharing
- MUSTANG outperforms existing single-sample ST data analysis tools

Authors

Seyednami Niyakan, Jianting Sheng, Yuliang Cao, ..., Ling Wu, Stephen T.C. Wong, Xiaoning Qian

Correspondence

xqian@ece.tamu.edu (X.Q.), stwong@houstonmethodist.org (S.T.C.W.)

In brief

Spatial transcriptomics offers a revolutionary approach to studying gene expression patterns within tissues by integrating spatial information with traditional transcriptomics sequencing technologies. There is a vast amount of spot deconvolution tools for spatial transcriptomics data that aim to dissect the spot-level aggregated gene expression signals. However, these tools are limited to single-sample analysis. This paper presents a new multi-sample spot deconvolution method that allows for efficient and accurate cross-sample and within-sample information sharing, drastically improving the deconvolution performance.

Niyakan et al., 2024, Patterns *5*, 100986
May 10, 2024
https://doi.org/10.1016/j.patter.2024.100986

CellPress

Patterns

CellPress
OPEN ACCESS

Article

MUSTANG: Multi-sample spatial transcriptomics data analysis with cross-sample transcriptional similarity guidance

Seyednami Niyakan,[1,4] Jianting Sheng,[2,4] Yuliang Cao,[2] Xiang Zhang,[3] Zhan Xu,[3] Ling Wu,[3] Stephen T.C. Wong,[2,*] and Xiaoning Qian[1,5,*]

[1]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA
[2]Department of System Medicine and Bioengineering, Houston Methodist Neal Cancer Center, Houston, TX 77030, USA
[3]Lester and Sue Smith Breast Center, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA
[4]These authors contributed equally
[5]Lead contact
*Correspondence: xqian@ece.tamu.edu (X.Q.), stwong@houstonmethodist.org (S.T.C.W.)
https://doi.org/10.1016/j.patter.2024.100986

THE BIGGER PICTURE Spatial transcriptomics (ST) enables the localization of cell types and their associated gene expression within tissue samples. In multi-cellular resolution ST, a tissue is divided into spots consisting of several cells, and this sometimes creates difficulties for cell characterization and identification in complex tissue samples. There are several methods for spot deconvolution, but most are limited to single-sample analysis and require a reference cellular profile. Here, we present MUSTANG (MUlti-sample Spatial Transcriptomics data ANalysis with cross-sample transcriptional similarity Guidance), a data analysis framework that permits multi-sample spot cellular deconvolution without a reference expression profile.

SUMMARY

Spatially resolved transcriptomics has revolutionized genome-scale transcriptomic profiling by providing high-resolution characterization of transcriptional patterns. Here, we present our spatial transcriptomics analysis framework, MUSTANG (MUlti-sample Spatial Transcriptomics data ANalysis with cross-sample transcriptional similarity Guidance), which is capable of performing multi-sample spatial transcriptomics spot cellular deconvolution by allowing both cross-sample expression-based similarity information sharing as well as spatial correlation in gene expression patterns within samples. Experiments on a semi-synthetic spatial transcriptomics dataset and three real-world spatial transcriptomics datasets demonstrate the effectiveness of MUSTANG in revealing biological insights inherent in the cellular characterization of tissue samples under study.

INTRODUCTION

Recent advances in single-cell RNA sequencing (scRNA-seq) have enhanced our knowledge of different cellular development processes and can help better characterize heterogeneity of cell types in many complex tissues.[1–3] However, in original scRNA-seq approaches spatial information is not retained when preparing samples with tissue dissociation and cell isolation.[4] Thus, scRNA-seq technologies lack the spatial resolution, which can be crucial for characterizing cellular heterogeneity in the spatial context when investigating tissue organizations.[5,6] To address this limitation, spatial transcriptomics (ST) technologies can measure gene expression at a variety of spatial locations (spots) in a tissue sample while preserving the source position of each expression datapoint.[7] Since the processes by which cells evolve into tissue compartments and interact with each other depend on interactions with the environment around it, spatial information that is naturally preserved by ST technologies presents ample opportunities for enhancing our understanding of disease progression and tissue development.[8]

Despite the rapid development of ST technologies, many of them still lack single-cell resolutions, such as Visium,[9] Slide-seq,[10] and HDST.[11] In these approaches, each tissue is divided into a grid or lattice of spots, with each spot in the grid typically being 50–100 μm wide, covering around 10–60 cells. These ST technologies output a high-dimensional, spatially localized

gene expression count vector for each spot, representing an aggregated gene expression of the cells in the spot.[12] As a result of the accumulated measurement at each detected spot, the measured signal is generally a mixture of multiple homogeneous or heterogeneous cell types, which may make it difficult to explore the spatial distribution of cell types in complex tissues.[13] Spot deconvolution methods aim to separate the contribution of different cell types in each spot, allowing for cell-type identification and characterization. This enables the analysis of cell-type-specific gene expression patterns and functional annotations, which is necessary for understanding the heterogeneity and cellular composition of complex tissues.[14] As a result of crucial need for methods capable of deconvolving cell-type fractions for each spot to improve interpretability and analysis of gene expression patterns, recently several spot deconvolution tools have been developed such as CARD,[14] BayesTME,[12] STdeconvolve,[15] Cell2location,[16] DestVI,[17] RCTD,[18] EnDecon,[13] SPOTlight,[19] and UniCell.[20]

One of the limitations of many existing spot deconvolution methods is the requirement for a reference profile of cell-type expression. Previous studies of RNA-seq data deconvolution algorithms have shown that choice of reference is more important than methods of choice in determining deconvolution performance. A reference-free spot deconvolution pipeline that does not rely on pre-existing reference atlases or datasets assures an unbiased analysis of ST data.[21] Recently, two reference-free tools, STdeconvolve and BayesTME, have been developed to deconvolve underlying cell types comprising multi-cellular spot resolution ST datasets.[12,15] STdeconvolve is based on latent Dirichlet allocation (LDA), a generative statistical model commonly used in natural language processing for discovering latent topics in collections of documents.[15] On the other hand, BayesTME is a Bayesian hierarchical generative model capable of performing spot deconvolution for aggregated gene expression measurements at spots in ST datasets, explicitly modeling the aggregated counts via a Bayesian factorized model formulation.[12]

While many of these ST analysis methods focus on analyzing individual ST samples, recent advances in high-throughput sequencing technologies, coupled with spatially resolved experimental techniques, have facilitated the generation of multi-sample ST datasets, enabling data integration and statistical modeling for more robust comparisons, validation, and identification of spatially regulated gene expression patterns.[22–24] For example, multi-sample ST allows more comprehensive investigation of gene expression spatial dynamics across different conditions (e.g., knockout versus wild type) or experimental settings (e.g., treatment responders versus non-responders).[25] In addition, Comparative analysis between samples offers insights into the spatial regulation of gene expression, unveiling spatial clusters and coordinated gene modules that would be overlooked in single-sample ST analysis. However, despite the ample opportunities that multi-sample ST data analysis may offer, to the best of our knowledge there are no available spot deconvolution tools for integrative analysis of multi-sample ST datasets. Recently, a hybrid machine learning and Bayesian statistical modeling framework called MAPLE has been developed for spot clustering of multi-sample ST data but does not perform spot cell-type deconvolution, which is crucial for the characterization of tissue samples.[25]

To fill these gaps, we introduce MUSTANG (MUlti-sample Spatial Transcriptomics data ANalysis with cross-sample transcriptional similarity Guidance), a multi-sample ST data analysis framework, to simultaneously derive the spot cellular deconvolution of multiple tissue samples without the need for reference cell-type expression profiles. MUSTANG is designed based on the assumption that the same or similar cell types exhibit consistent gene expression profiles across samples. This assumption is reasonable in practice. For example, there are several studies, including Joglekar et al.,[26] suggesting cell types such as excitatory neurons or inhibitory interneurons, and glial cells (astrocytes and oligodendrocytes) often tend to display relatively consistent gene expression patterns across different regions of the central nervous system. However, regional identity can, although rarely, override cell-type specificity. There are some cell types such as immune cell populations that can display region-specific gene expression profiles within a tissue but still these cells have shared consistent transcriptional patterns to some extent, which assures the practicality of our assumption even in these rare cases. In addition, MUSTANG adjusts for potential batch effects as crucial multi-sample experimental considerations to enable cross-sample transcriptional information sharing to aid in parameter estimation. With that, spatial correlation in gene expression patterns within samples is further accommodated by constructing and employing a spot "similarity" graph that includes both transcriptional and spatial similarity edges between spots across samples. By aligning and integrating multiple tissue samples, MUSTANG can effectively leverage shared information and increase the robustness of joint spot cell-type deconvolution analysis across multiple ST samples. In summary, our key technical contributions include the following:

(1) MUSTANG, to the best of our knowledge, is the first reference-free spot deconvolution method for multi-sample ST data analysis.
(2) MUSTANG allows both intra-sample and inter-sample information sharing by introducing a new spot similarity graph.
(3) Besides modeling spot spatial dependency, MUSTANG implements batch correction across ST samples in the workflow to avoid obscuring inherent biological signals when sharing transcriptional information.

To demonstrate the capability of MUSTANG for revealing the true underlying spot-level cell-type proportions in multi-sample ST datasets, we have applied MUSTANG to a simulated semi-synthetic and three real-world ST datasets of different tissue properties and show that it can be effectively used for unveiling the inherent biological signal in tissue architectures.

RESULTS

Model overview

Given gene count matrices of all spots across tissue samples and spatial coordinates for spot centroid positions, MUSTANG performs spot cellular deconvolution for multi-sample ST data.

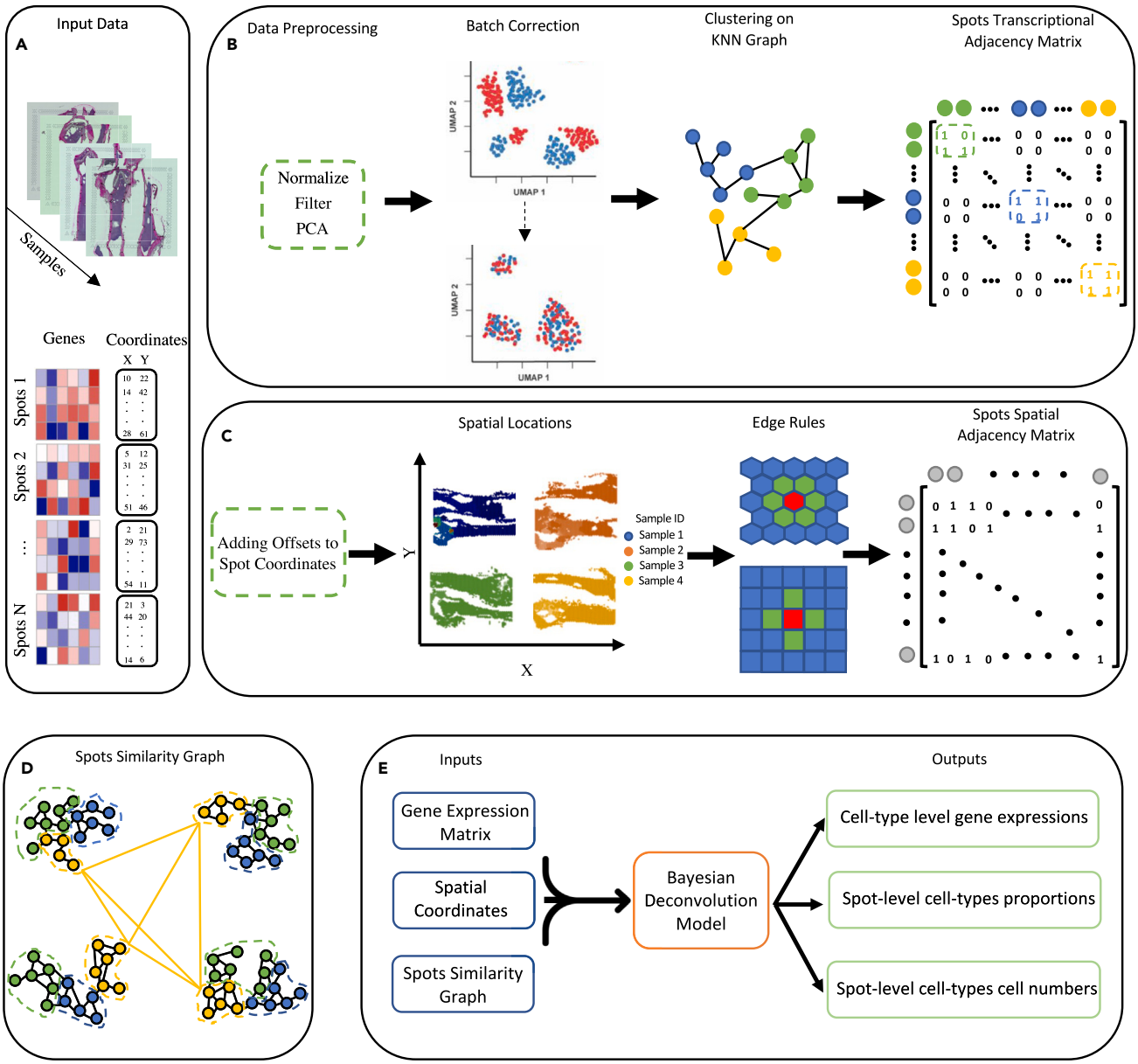

Figure 1. The MUSTANG framework to analyze multi-sample spatial expression data

(A) MUSTANG requires gene expression matrices of all the spots across tissue samples as well as the spatial coordinates of the spots. The gene expression matrices are concatenated to form a single expression matrix of genes for all spots.

(B) MUSTANG performs standard scRNA-seq data preprocessing steps such as normalization, gene filtering, and then dimension reduction of gene expression matrices of the combined spots across samples via principal-component analysis (PCA). The top principal components are batch corrected to remove any unwanted technical confounders. Then MUSTANG performs Louvain clustering on the K-nearest neighbor graph constructed based on the batch corrected top PCs to get the clusters of similar spots. The spot transcriptional adjacency matrix is then constructed based on the resulted spot cluster memberships.

(C) MUSTANG adds different offset values to the spatial coordinates of the spots from different ST samples so that they can be aligned properly. Depending on the sequencing technology layout (e.g., lattice or hexagonal), the spots spatial adjacency matrix is determined.

(D) The spot similarity graph is constructed by MUSTANG based on the summation of spots spatial and transcriptional adjacency matrices. Spots are colored by their corresponding transcriptional clusters. The edges in black indicate the spatial neighboring connection between two spots and the yellow-colored edges demonstrate the transcriptional similarity between yellow-colored spots.

(E) Final step of MUSTANG corresponds to joint Bayesian deconvolution analysis based on raw concatenated gene expression matrix, spatial coordinates with added offsets, and the spot similarity graph.

The overall workflow of MUSTANG is presented in Figure 1. MUSTANG includes four main steps: (1) construction of spot transcriptional adjacency matrix of expression-based information sharing across tissue samples after batch effect correction, (2) construction of a spot spatial adjacency matrix to allow spatial correlation between physically neighboring spots within the samples, (3) construction of the spot *similarity* graph based on the spot transcriptional and spatial adjacency matrices, and (4) deconvolution of aggregated spot-level gene expression measurements to signals coming from different cell types based on a Bayesian hierarchical model. Here, we discuss each step in more detail.

Spot transcriptional adjacency matrix

MUSTANG first identifies the common genes across multiple input tissue samples and then concatenates the spot count matrices of all samples $\{1, \ldots, N\}$ over the common genes (Figure 1A). Then, MUSTANG performs the common data preprocessing steps similar to typical scRNA-seq data analysis, such as normalization, feature selection, and dimension reduction. First, the combined gene expression matrix of all tissue samples are log transformed and normalized using library size. Then, the top 2,000 (optional) highly variable genes are selected based on the variance of the log-expression profiles. We further perform principal-component analysis on the normalized expression profiles of selected top highly variable genes across all the spots from tissue samples. Then, the reduced-dimension transcriptional matrix of all spots by top 50 principal components (PCs) is retained to capture as much variation as possible while scaling up with complexity of analyzing high-dimensional data. To remove any unwanted technical batch effect from the analysis such as the case where tissue samples are from different sequencing technologies or samples are generated from multiple experiments or across different laboratories, MUSTANG performs batch effect correction on the retained top PCs. One powerful method for batch correction is the Harmony algorithm.[27] MUSTANG uses Harmony to adjust for batch effects from the PCs and ensures that the subsequent analyses are not confounded by technical variability. Later, based on the batch corrected top 50 PCs, the K-nearest neighbor (KNN) graph of spots is constructed. Basically, in the KNN graph the nodes represent spots across ST samples and two spots are connected with an edge if they are within the k-most transcriptionally similar spots from each other for user-selected resolution parameter k. We measure the transcriptional similarity between spots by calculating the Euclidean distance of the batch corrected top 50 PC scores. Here, in MUSTANG we suggest selecting k to be 50 considering computation performance trade-off. In addition, we weigh the edges between two spots i and j in the KNN graph by $\frac{1}{1+Dist(i,j)}$ where $Dist(i,j)$ is the corresponding PC-based Euclidean distance between the two spots. This way, the edges between spots that are transcriptionally more similar will be weighed with higher values. Then, MUSTANG applies unsupervised graph-based Louvain clustering on the weighted KNN graph to get clusters of spots that are transcriptionally similar.[28] Lastly, MUSTANG constructs the spot transcriptional adjacency matrix based on the spot membership in the resulted Louvain clustering results. If $\mathbf{T}$ is the cross-sample spot transcriptional adjacency matrix, then the value $\mathbf{T}_{ij} = \mathbf{T}_{ji} = 1$ at spots i and j means that i and j are in same transcriptional Louvain clustering class of spots and they are not within a same tissue sample (Figure 1B).

Spot spatial adjacency matrix

The next step in MUSTANG constructs a spot spatial adjacency matrix. In this step MUSTANG only uses the coordinates of all the spots. Initially, we add different constant values to all spot coordinates of different samples so that it could be possible to overlay the physical locations of spots from different samples on a single layout without spots from different samples getting overlapped or neighbored as shown in Figure 1C. Then, based on the geometric representations of spots in ST sequencing technologies, such as lattice layouts (e.g., Slide-seq[10]) or hexagonal layouts (e.g., Visium[9]), neighbors can be identified for each spot based on shared edges. This edge rule leads to four and six neighbors for non-boundary spots in lattice and hexagonal layouts, respectively. Finally, MUSTANG constructs the spots spatial adjacency matrix based on the described edge rule. If we call the spots spatial adjacency matrix $\mathbf{S}$, then the value $\mathbf{S}_{ij} = \mathbf{S}_{ji} = 1$ means that i and j have a shared edge between them (Figure 1C).

Spot similarity graph

After deriving both spot transcriptional and spatial adjacency matrices, MUSTANG constructs the overall spot similarity graph. The adjacency matrix of the spot similarity graph is a binary matrix, which is resulted after taking the logic "OR" operation between pairwise indices of spot transcriptional and spatial matrices $\mathbf{T}$ and $\mathbf{S}$. More specifically, if we denote the spot similarity graph adjacency matrix by $\mathbf{A}$, $\mathbf{A}_{ij} = \mathbf{T}_{ij} \vee \mathbf{S}_{ij}$, where $\vee$ indicates the OR operator. Figure 1D shows an example of how a spot similarity graph might look like for an ST dataset with four tissue samples. In this figure, spots are colored based on their transcriptional cluster labels. The black-colored edges are the edges according to the spot spatial adjacency matrix. On the other hand, the yellow-colored edges indicate the transcriptional similarity between yellow-colored spots. Note that, for simplicity, only the transcriptional edges between yellow-colored spots are drawn and transcriptional edges between blue and green spots are not shown in the figure. In addition, it worth mentioning that each yellow edge between a pair of yellow spots in the corresponding clusters is representative of all edges from spots of one cluster to another in Figure 1D.

Joint Bayesian deconvolution analysis

The last step of our MUSTANG workflow corresponds to joint Bayesian deconvolution analysis of a raw concatenated gene expression matrix to preserve information in the original ST data, together with the spot similarity graph and spatial coordinates with added offsets. Our joint Bayesian deconvolution model is based on the Poisson discrete deconvolution model recently introduced in BayesTME for single-sample analysis of ST data.[12] More precisely, in this Poisson model, the raw aggregated expression measurement of gene g at spot s, denoted as Y_{sg}, are factorized as the summation of k (i.e., number of cell types) different Poisson distributed read counts Y_{sgk}. In

fact, each of these reads models the total expression count of gene g in the cells of type k that are at spot s. Thus, based on this factorization we can explicitly model the raw ST counts Y_{sg}:

$$Y_{sg} = \sum_k Y_{sgk} \sim \text{Pois}\left(\sum_k \beta_k d_{sk} \varphi_{kg}\right), \quad \text{(Equation 1)}$$

where the rate parameter of the Poisson distributions is controlled with three parameters β_k, d_{sk}, and φ_{kg}. The cell-type-dependent parameter β_k quantifies the expected total count for cell type k and d_{sk} represents the number of cells of type k that are at spot s. The parameter φ_{kg} captures the normalized gene expression profile of gene g in cell type k. This way of modeling gene expression in ST data assures biological considerations such as a monotonic relationship between the number of cells and aggregated read measurement in each spot as well as different expression profiles for each gene in various cell types. To complete the Poisson discrete deconvolution model, Dirichlet and gamma distribution priors are imposed on φ_k and β_k parameters, respectively. In addition, the prior on d_{sk} is constructed hierarchically based on the heavy-tailed Bayesian variant of the graph-fused binomial tree as described in Tansey et al.[29] In this binomial tree model, the cell-type assignment probabilities in each spot are decomposed into a series of binomial decisions where the prior on each binomial probability encourages spatial smoothness across spots. Specifically, such spatial smoothness on cell-type assignment probabilities is achieved by imposing the sparsity inducing grouped horseshoe distribution[30] over the graph fussed LASSO[31] (i.e., zeroth-order graph trend filtering) penalized cell-type assignment probabilities:

$$\begin{aligned} &D_s \sim \text{Binom}(n_{max}, 1 - \sigma(\theta_{s0})), \\ &d_{sk} \sim \text{Binom}\left(D_s - \sum_{r=1}^{k-1} d_{sr}, \sigma(\theta_{sk})\right), \forall 1 < k < K \quad \text{(Equation 2)} \\ &(\Delta_{Spatial}\Theta)_j \sim \text{Grouped Horseshoe}(\lambda). \end{aligned}$$

In Equation 2, n_{max} is the maximum possible number of cells in each spot, for which its default value is set to be 100. The logistic function is noted by σ and the parameter D_s is the total number of cells in spot s out of possible n_{max} cells and θ_{sk} captures the cell type k probability proportions at spot s. Lastly, $\Delta_{Spatial}$ is the edge-oriented zeroth-order graph trend filtering matrix of the spot spatial graph with a hyperparameter λ controlling the global degree of smoothness.

Here, in our joint Bayesian deconvolution model while performing multi-sample ST data analysis in MUSTANG, we further allow information sharing across tissue samples in the Poisson discrete deconvolution model. We take advantage of the prior knowledge inherited in the spot similarity graph that we constructed in the MUSTANG workflow as detailed in the previous section. Specifically, we include transcriptional similarity in addition to the spatial similarity to take into consideration the biological belief that spots that have similar batch-corrected transcriptional profiles might also have similar cell-type composition as well. This is done by taking advantage of the zeroth-order graph trend filtering matrix of the spot similarity graph in the hierarchical prior in Equation 2. In MUSTANG, we impose the grouped horseshoe distribution over the graph fussed LASSO penalized cell-type assignment probabilities based on the spot similarity graph as:

$$(\Delta_{Similarity}\Theta)_j \sim \text{Grouped Horseshoe}(\lambda). \quad \text{(Equation 3)}$$

This results in inferring both transcriptionally and spatially smooth cell-type proportions, allowing to borrow signal strengths from both inter-sample and intra-sample spots for effective joint analysis of multiple tissue samples in a given ST dataset.

The posterior inference procedure of the joint Bayesian deconvolution model in MUSTANG is based on Gibbs sampling. The full derivations for all complete conditionals and Gibbs sampling-based updates are similar to Zhang et al.[12] and detailed in the supplemental information. During the inference process, we use Markov chain thinning, with five thinning steps between each sample. We collect 100 Markov chain Monte Carlo (MCMC) samples after 1,200 burn-in iterations for our consequent analyses and evaluation.

Experiments

We have evaluated our MUSTANG for analysis of multi-sample ST data from semi-synthetic ST data as well as three real-world ST datasets generated by the 10X Genomics Visium platform.[9] First, a semi-synthetic multi-sample ST data generation is described and then the simulated samples are analyzed with MUSTANG and other state-of-the-art cell-type deconvolution tools to comprehensively quantify and benchmark the performances of these tools across different metrics. Specifically, the results clearly showcase the MUSTANG superiority in accurate deconvolution of aggregated signals in ST data in most of the settings. Then, a mouse brain ST dataset having nearby brain tissue areas bisected to paired anterior and posterior sections is analyzed with MUSTANG to showcase its capability in identifying cell types that have consistent patterns across neighboring tissue regions from different paired sections. The results match the known anatomical brain regions from the Allen Brain Atlas.[32,33] We also apply MUSTANG on a human brain ST dataset to further quantitatively benchmark the spot deconvolution performance. Specifically, the significance of different components in MUSTANG enabling multi-sample ST analysis will be demonstrated in this ablation study compared with BayesTME and a simpler version of MUSTANG that does not take spot transcriptional adjacency matrix into account. We then analyze a mouse bone marrow tissue ST dataset to characterize the tumor microenvironment (TME). The matched immunofluorescence (IF) staining images are used to validate the findings by analyzing bone tissue samples with MUSTANG.

Semi-synthetic data

To benchmark the performance of MUSTANG on accurately deconvolving the aggregated signals in ST data, we apply it to the recently published ST benchmark datasets.[34] As the ground-truth cell-type compositions are not available for multi-cell per spot ST datasets, following the instructions in Li et al.,[34] we have generated four ST dataset samples from the STARmap data of mouse primary visual cortex tissue, termed "Dataset 10" in the original benchmark study,[34] as shown in Figure 2A.

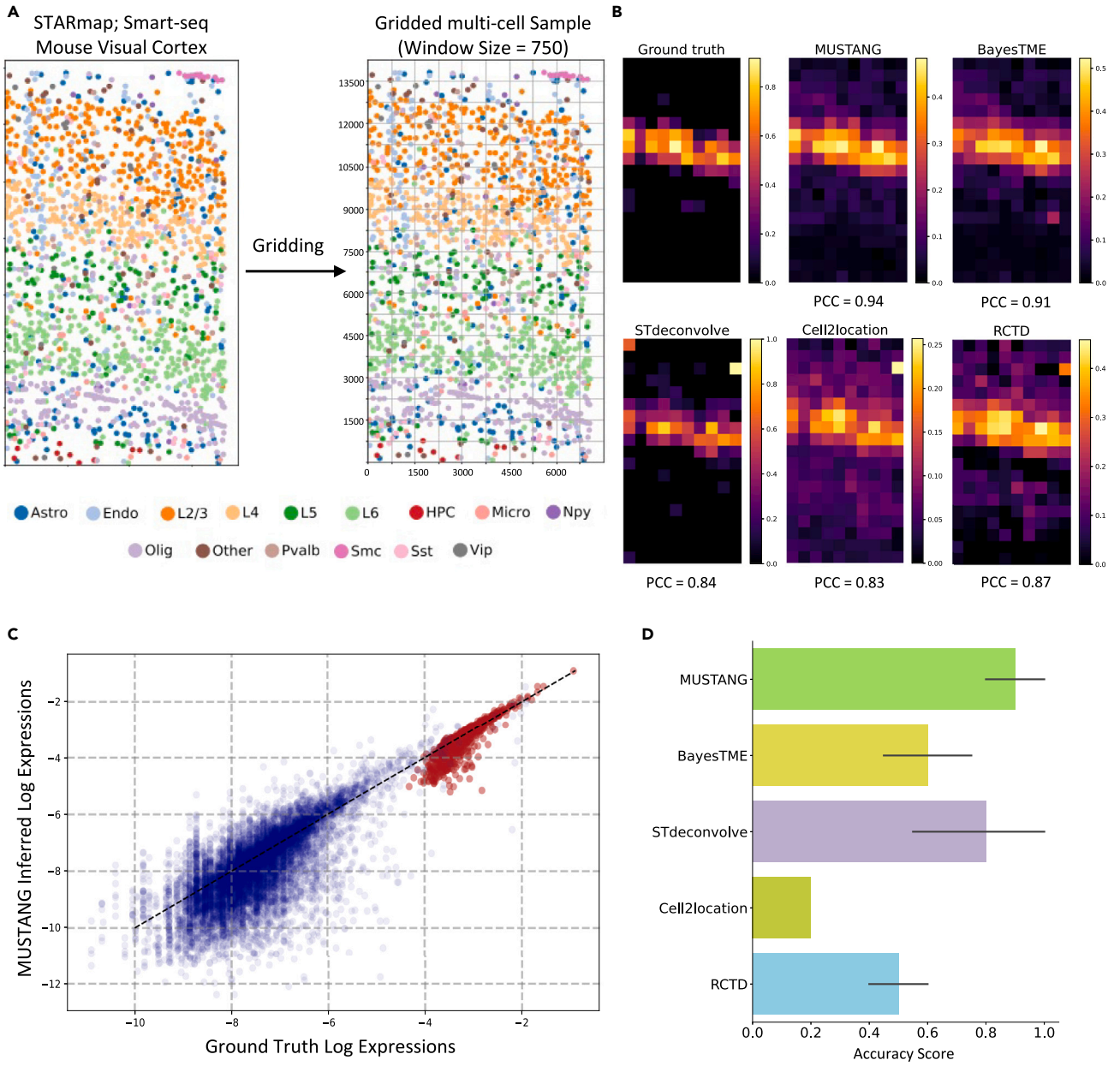

Figure 2. Comparing the cell-type deconvolution performance of MUSTANG and other deconvolution methods on semi-synthetic ST data

(A) Left: a STARmap slide of mouse visual cortex tissue, with cells annotated by cell types. Right: an example of a simulated gridded multi-cell sample with a window size of 750 pixels where each grid represents a simulated spot containing multiple cells.

(B) The proportion of L4 excitatory neurons in the spots simulated in the gridded sample with a window size of 750 pixels, including the ground truth and the predicted results of deconvolution tools.

(C) MUSTANG-inferred cell-type-level expression profiles for all clusters and genes compared against the ground truth (n = 10,854). As an example, the expression signatures of L4 excitatory neurons are colored in red (PCC = 0.98, n = 882).

(D) MUSTANG outperforms all other existing tools in the cell-type deconvolution task for all clusters in all four simulated samples from the mouse visual cortex data in terms of the accuracy score aggregated from PCC, SSIM, RMSE, and JSD metrics.

The original STARmap data have the spatial position and gene expression information of the 1,549 cells, corresponding to 15 cell types. To generate a semi-synthetic multi-sample ST data from the STARmap mouse visual cortex data, we partition the original tissue slide into grids and each grid simulates an ST spot with known cell-type composition. Then, the corresponding gene count expression matrix is generated by taking the sum of the expression profile of all the cells in each spot. To generate the four simulated ST samples with potentially ambiguous cell-type compositions, we consider four different grid window sizes of 600, 650, 700, and 750 pixels to partition the original STARmap data. The generated samples are shown in

Figures 2A and S1A. The simulated multiple ST samples had different numbers of spots ranging from 189 to 276 spots (details in supplemental information).

We jointly analyze the generated multiple ST samples by MUSTANG and compare its cell-type deconvolution performance with both reference-free (BayesTME and STdeconvolve) and reference-based (Cell2location and RCTD) single-sample cell-type deconvolution state-of-the-art tools. As an example, in Figure 2B we visualize the ground-truth proportions of the L4 excitatory neurons across spots in the simulated ST sample with grid window size of 750 and compared it with the estimated proportion of different methods for this cell type. As indicated in the figure, MUSTANG performs better in terms of Pearson correlation coefficient (PCC) values (0.94), followed by BayesTME (0.91), RCTD (0.87), STdeconvolve (0.84), and Cell2location (0.83). In addition, as MUSTANG is able to estimate the normalized cell-type-level gene expressions (i.e., φ_{kg} in Equation 1), we compare the mean expression of genes from the single-cell reference for each cell type with the MUSTANG-inferred expression signatures in Figure 2C. For better visualization, we have plotted the expressions in log10 space. As an example, the expression profiles of genes in L4 excitatory neurons are colored in red with PCC values of 0.98, confirming the accuracy of MUSTANG to estimate the cell-type-level normalized gene profiles.

Following Li et al.,[34] to more comprehensively quantify the MUSTANG cell-type deconvolution performance and those of the other state-of-the-art ST data cell-type deconvolution tools, we calculate three other metrics besides PCCs for each cluster in all four simulated samples: structural similarity index (SSIM), root-mean-square error (RMSE), and Jensen-Shannon divergence (JSD). Then, to simplify the evaluation of the accuracy, the accuracy score (AS), which is the normalized average rank of the four metrics (with the highest AS score of 1), is derived. As plotted in Figure 2D, MUSTANG outperforms all other tools in terms of ASs in the deconvolution task, highlighting the power of multi-sample ST data analysis with effective inter- and intra-sample information sharing implemented in MUSTANG to aid the parameter estimation procedure. The detailed benchmarking of the methods across all four metrics are demonstrated in the supplemental information. Overall, the results on this multi-sample semi-synthetic data analysis experiment suggest that MUSTANG has dominant performance in most of the adopted evaluation metrics consistently across all clusters in the four samples but, as also previously noted in other benchmark studies, no method is able to obtain superior performance in all settings.[12,34]

Mouse brain data

The brain tissue in an adult mouse is composed of myriad cell types in a highly organized and coordinated manner for normal neurological functions through well-defined molecular mechanisms.[32,33] To validate the MUSTANG capability of appropriately deconvolving the aggregated gene expression signals from spatial sequencing technologies on complex tissue architectures, we use the four anterior and posterior sections of mouse brain tissues on the sagittal plane. These adult mouse brain tissue sections are sequenced by the 10X Visium platform[9] and the generated spatially resolved transcriptomics data made publicly available by 10X Genomics. Specifically, these mouse brain tissue data consist of two biological replicates of paired anterior and posterior sections on the sagittal plane. Figure 3A shows the four tissue slices placed on the 10X Visium gene expression slides. Due to the presence of nearby brain tissue regions in different tissue sections (in either anterior or posterior slices) at bisection areas in this ST dataset, applying MUSTANG multi-sample analysis helps validate the effectiveness of MUSTANG by checking whether the neighboring tissue regions from different sections have consistent cell-type deconvolution properties or not.

To catalog the spatial organizations of various brain areas in brain tissue and thus provide a holistic view of gene expression patterns at whole-brain level, we simultaneously analyze the ST data of four brain tissue sections with MUSTANG. Following the same MUSTANG workflow steps indicated in detail in the model overview section, we first construct the spot similarity graph and then fit our joint deconvolution model to the concatenated ST data. As matched ground truth annotations are not available for these data, for picking the number of brain regions K, we follow the known anatomy of mouse brain tissue publicly available by the Allen Mouse Brain Atlas,[32,33] which is the most comprehensive genome-wide atlas of mouse brain tissue. Based on this reference annotation of mouse brain regions, we select K to be 11, corresponding to the 11 major brain regions, including the olfactory bulb, cortex, striatum, pallidum, hippocampus, thalamus, hypothalamus, midbrain, pons, medulla, and cerebellum regions (Figure S2 in the supplemental information).

The spatial scatter pie chart of the MUSTANG-inferred brain region probabilities in Figure 3B indicates that the deconvolution analysis by MUSTANG accurately reconstructs the layered and segmented structure of mouse brain anatomy. Matching the reference anatomy of mouse brain tissue from the Allen Brain Atlas (Figure S2) and the MUSTANG spatial scatter pie chart demonstrates a clear mapping between identified sub-populations by MUSTANG and known major mouse brain anatomical regions in both anterior and posterior regions. For instance, the brain area 4 found in anterior sections of samples 1 and 2, corresponds clearly to the olfactory bulb region of mouse brain. Likewise, in the posterior sections, brain area 2 corresponds to the cerebellum region. In addition, some regions such as brain area 1 are more heterogeneous as they cover both striatum and pallidum brain areas in the anterior slices.

A closer inspection of deconvolution analysis results by MUSTANG in Figure 3B clearly demonstrates the capability of MUSTANG in identification of brain tissue areas that are shared in all posterior and anterior sections. Particularly, MUSTANG detects brain areas 3 and 5, which represent hypothalamus and cortex regions that are bisected by the sagittal plane for division of anterior-posterior sections in the experimental design. Furthermore, the continuous spatial patterns and consistency of the inferred brain area probabilities for these areas that are at bisection regions of the paired anterior-posterior sections highlight the distinct advantage of jointly analyzing ST samples with accurate cross-section information sharing implemented in MUSTANG over the non-integrative ST data analysis tools.

In addition to evaluating the MUSTANG performance on inferring the cell-type probabilities, we also examine the cell-type cell-count values learned by our deconvolution model. The left

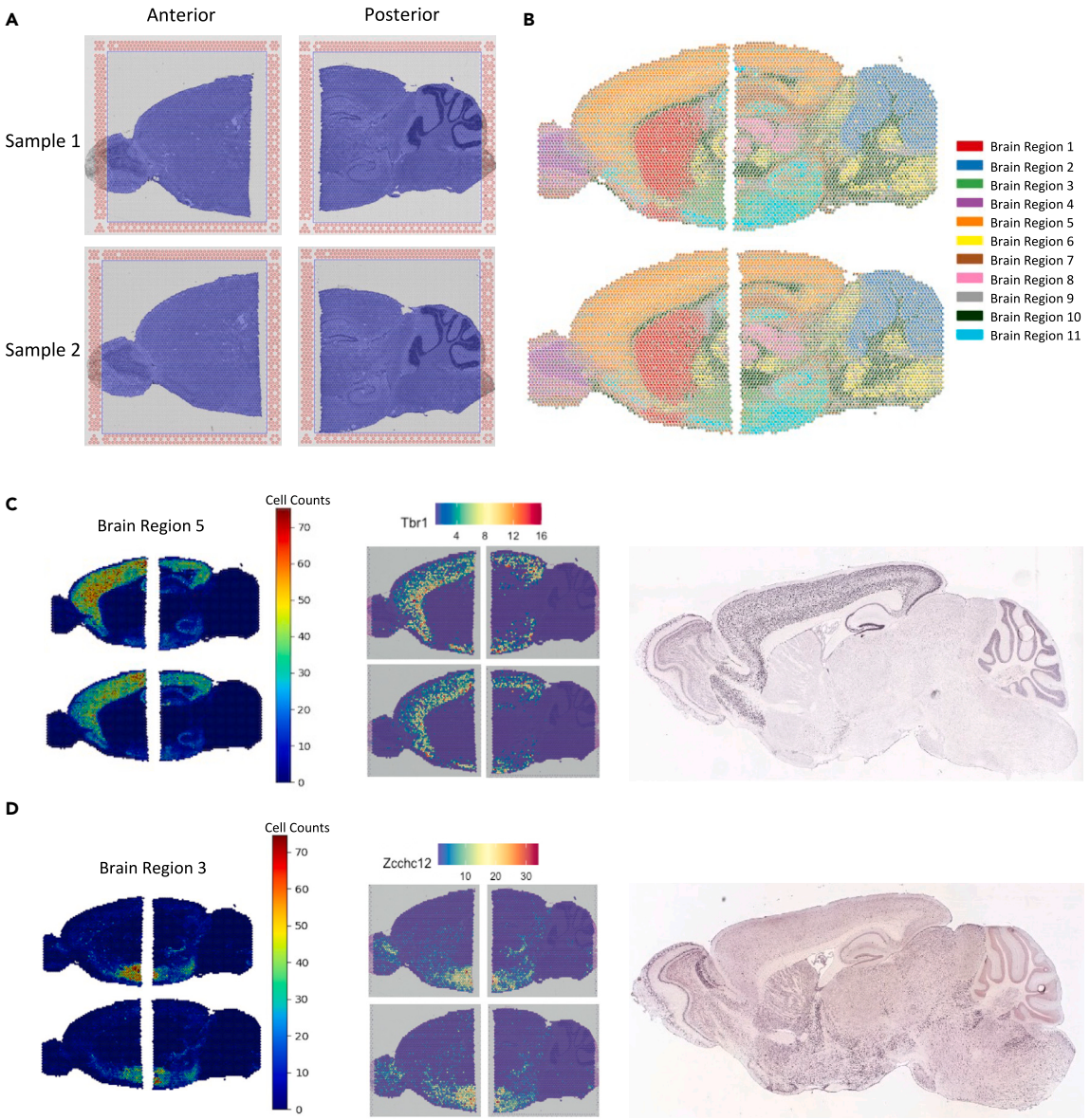

Figure 3. Analysis of four anterior and posterior sections of mouse brain tissue on sagittal plane with MUSTANG

(A) Paired anterior-posterior slices placed on the 10X Visium gene expression slides.

(B) Spot-based spatial pie charts of MUSTANG-inferred brain region proportions for all four mouse brain tissue sections.

(C) Left: MUSTANG-inferred cell numbers for brain region 5 matching the spatial pattern of the cortex anatomical brain region. Middle: spot-level expression visualization of the known cortex layer marker gene Tbr1. Right: the ISH images of this marker gene from the Allen Brain Atlas.

(D) Left: MUSTANG-inferred cell numbers for brain region 3 matching the spatial pattern of the hypothalamus anatomical brain region. Middle: spot-level expression visualization of the known hypothalamus layer marker gene Zcchc12. Right: the ISH images of this marker gene from the Allen Brain Atlas.

panels of Figures 3C and 3D visualize the spatial pattern of MUSTANG-inferred brain regions 5 and 3 cell counts across the four tissue sections. As we can see, the spatial patterns of inferred brain region cell counts similar to brain area probabilities clearly match the cortex and hypothalamus regions from the Allen Brain Atlas annotations. To further examine these cell-type mappings, we visualize the raw gene expression spatial patterns of two known marker genes *Tbr1* and *Zcchc12* for the mouse brain areas cortex and hypothalamus from the Allen Mouse Brain Atlas[32,33] in the middle panels of Figures 3C and 3D. In addition, for more accurate validation of the predicted brain area spatial distribution within the brain structure, we extract the reference *in-situ* hybridization (ISH) image data for these two known brain region gene markers from the Allen Mouse Brain Atlas and plot them in the right panels of Figures 3C and 3D. As we can see in Figures 3C and 3D, there is high correlation between the raw gene expression and ISH image spatial patterns of the cortex and hypothalamus brain region gene markers and the inferred cell counts for their matching brain area from MUSTANG, highlighting the accurate simultaneous segmentation and

deconvolution analysis of all four brain sections done by MUSTANG.

Overall, analyzing the four anterior and posterior mouse brain ST data with MUSTANG clearly showcases the important advantage of our proposed multi-sample data analysis tool in identifying both section-specific brain regions as well as shared areas between all tissue sections by jointly analyzing these sections. Furthermore, the inferred deconvolution parameters for the brain regions present at the areas close to the bisection plane of paired anterior-posterior sections are also consistent, illustrating the deconvolution accuracy of our MUSTANG in nearby tissue regions.

Human brain data

In a recent study,[35] spatial expression profiles of 12 dorsolateral prefrontal cortex (DLPFC) tissue samples were generated. Based on the selected DLPFC layer-specific gene makers and cytoarchitecture consideration, six cortical layers (i.e., L1-L6) and white matter (WM) for each brain tissue sample were annotated. Here, we use the ST expression profiles of four samples (sample IDs: 151673 to 151676) from this dataset to showcase the benefits of simultaneously denconvolving tissue samples using our proposed MUSTANG.

Figure 4A shows the hematoxylin and eosin (H&E) staining images of four DLPFC tissue samples from the human brain ST dataset as well as the cortical layers and WM reference annotations for sample 151673 from the original study. Following our MUSTANG workflow, we first start analyzing the samples by constructing spot transcriptional and spatial adjacency matrices. As shown in Figure 4B, we derive the spot spatial adjacency matrix by adding offsets to spatial coordinates of DLPFC tissue samples and overlaying them on the ST grid space based on the Visium platform. In the transcriptional space, we follow the data preprocessing steps previously described in the MUSTANG model overview section to derive the dimension-reduced top 50 PCs for spot-aggregated gene expression counts. Figure 4C displays the UMAP (uniform manifold approximation and projection[36]) embedding of the derived top 50 PCs. It can be seen that there is strong batch effect in this dataset as spots from different tissue samples are clustered based on their sample ID rather than their underlying biological cell types. Although these samples are from the same tissue and sequencing platform, this observed batch effect in the data calls for the need of batch effect correction when analyzing multiple tissue samples to reduce the potential influence from any confounding technical factor. We therefore implement Harmony in MUSTANG to derive the batch corrected top 50 PCs. The UMAP embeddings of the batch-corrected PCs are shown in Figure 4D, where the spots from different samples are now mixed together while preserving potential expression differences. We further construct the KNN graph of spots based on these top PCs and apply Louvain clustering, resulting in eight distinct transcriptional sub-populations. In Figure 4E, the spots from four samples are colored by their transcriptional clusters in the UMAP embedding space. With that, the spot transcriptional adjacency matrix and, consequently, the spot similarity graph, can be constructed. Finally, we fit our joint Bayesian deconvolution model to the concatenated data with $K = 7$ cell types (i.e., six cortical layers plus WM). Based on the collected post burn-in MCMC samples, we derive the posteriors of the joint deconvolution model parameters such as spot-wise cell-type proportions, cell-type cell numbers, and normalized cell-type-specific gene expression. Figure 4F demonstrates the spatial scatter pie chart plot of our four DLPFC tissue samples, in which spots are plotted in their physical coordinates and at each spot there is a circular pie chart representing the inferred proportions of assigned cell types in that spot. The high similarity between the spatial patterns of cell-type proportions in the spatial pie chart plots of all four samples and the ground truth annotations from the original study demonstrates the capability of MUSTANG to simultaneously infer the underlying spot-wise biological cell-type proportions across multiple tissue samples.

As the ground truth cell-type proportions and cell-type cell numbers do not exist for multi-cell resolution ST data, inspired by the guidelines described in the recent benchmarking study of cell-type deconvolution methods for ST data,[37] we quantify the cell-type cell number inference performance of MUSTANG based on the PCC between the predicted spot-wise cell counts of specific cell type (i.e., d_{sk} in Equation 1) and the corresponding marker gene expression profiles. Specifically, we benchmark MUSTANG with BayesTME, which is an ST data deconvolution tool capable of inferring cell-type cell numbers without the need for paired reference expression profiles. As BayesTME is designed for single-sample analysis, we analyze each brain tissue sample separately using BayesTME as the baseline.

To calculate the PCC values, we first gather the list of known layer-specific marker genes from two previous brain studies[38,39] that were also used in the original DLPFC dataset paper.[35] Specifically, we only use those marker genes that are annotated to be related to only one of the DLPFC layers except for the WM layer, for which as we could not find any WM-specific markers in the two references, we select the marker genes that are shared between layer 6 and the WM. The heatmap plot in Figure 5A shows the list of selected layer-specific marker genes. The colors in the plot represent the corresponding reference papers that reported the corresponding marker genes.

Next, we extract the layer-specific gene expression profiles of DLPFC layers based on the ''pseudo-bulking'' approach noted in the original study of the DLPFC dataset,[35] in which the UMI counts for each gene within each layer across 12 spatial replicates are summed up to generate layer-enriched expression profiles. The layer-specific gene expression profiles of DLPFC layers have shown previously in Maynard et al.[35] to capture biological properties inherent in DLPFC layers. The pseudo-bulk data are available as ''sce_layer data'' for download through the `fetch_data` function in `spatialLIBD` R package. Following the instructions for cell-type deconvolution benchmarking described in Li et al.,[37] for each DLPFC layer we calculate the PCC between the expression profile of each layer in the extracted pseudo-bulk data and the inferred normalized expression profile of all cell types (i.e., φ_{kg} in Equation 1) from MUSTANG, choose the best-paired inferred cell type with the highest PCC and match it to that layer. After assignment, this chosen cell type would be ignored in the future steps. Then, we repeat the aforementioned steps on the next layer until all layers are iterated. For now, each layer should be paired with the best suitable cell type without duplication.

Finally, to complete the quantitative comparison between different ST analysis methods, for each DLPFC layer we

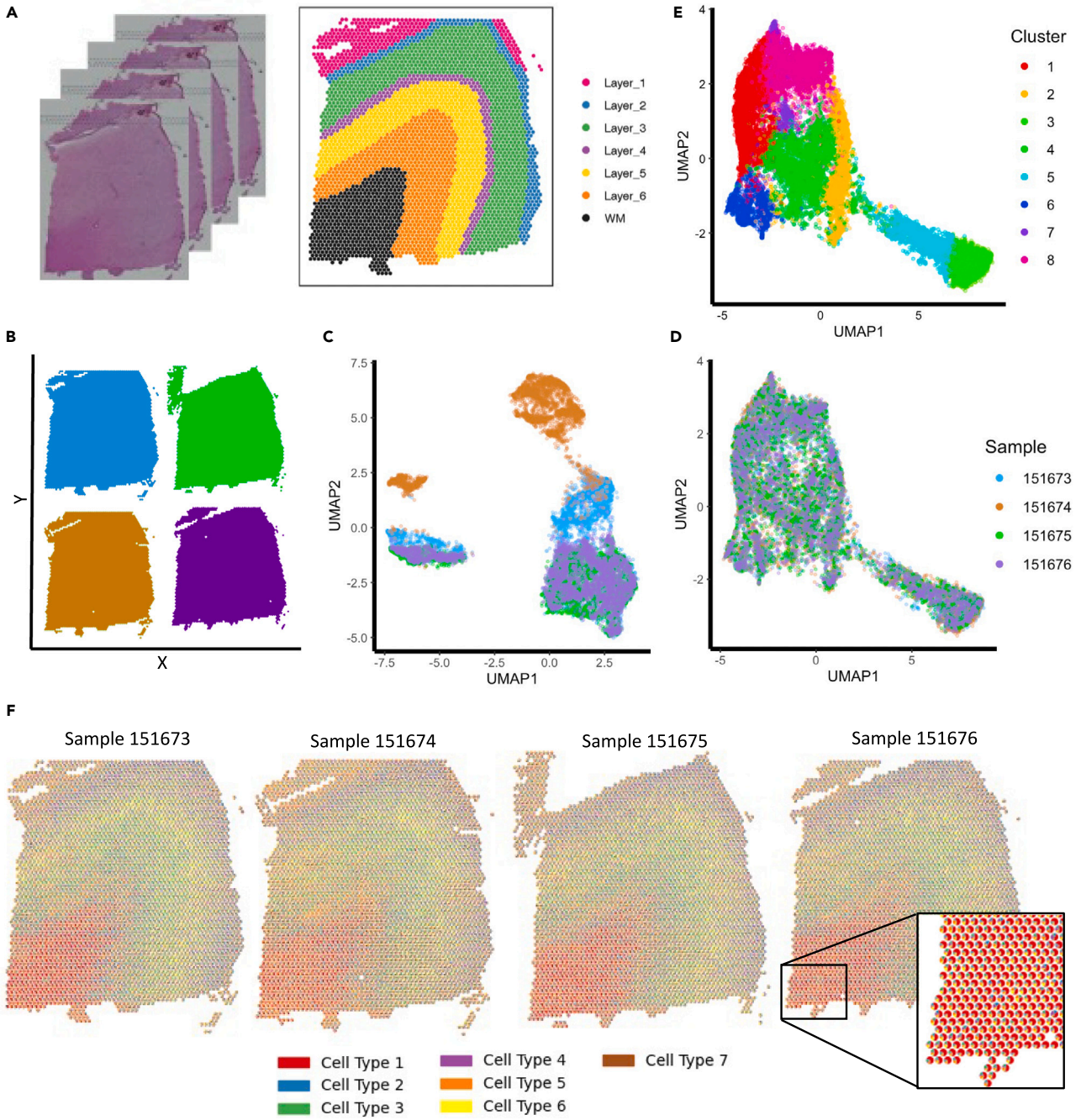

Figure 4. Analysis of four human brain DLPFC tissue samples with MUSTANG

(A) H&E staining images of four tissue samples (right) and the reference annotations of spots for the sample 151673 (left).

(B) Overlaying tissue samples on a grid space to construct spot spatial adjacency matrix.

(C and D) (C) UMAP embedding visualization of spots by top 50 PCs before and (D) after batch correction.

(E) Visualization of clustering based on batch corrected top 50 PCs. The spots are colored based on their transcriptional cluster label inferred from Louvain clustering.

(F) Spot-based spatial pie charts of MUSTANG-inferred cell-type proportions across all four DLPFC tissue samples matching the reference annotations from the original study.

calculate the PCC value between the corresponding marker gene expression of that layer in Figure 5A and the inferred cell number corresponding to the best-paired cell type. We calculate PCC values for each of the four tissue samples separately after jointly analyzing them with MUSTANG. We repeat the same procedure for analyzing tissue samples separately using BayesTME and calculate the corresponding PCC values. The boxplots in Figure 5B show the PCC values for each method on each sample

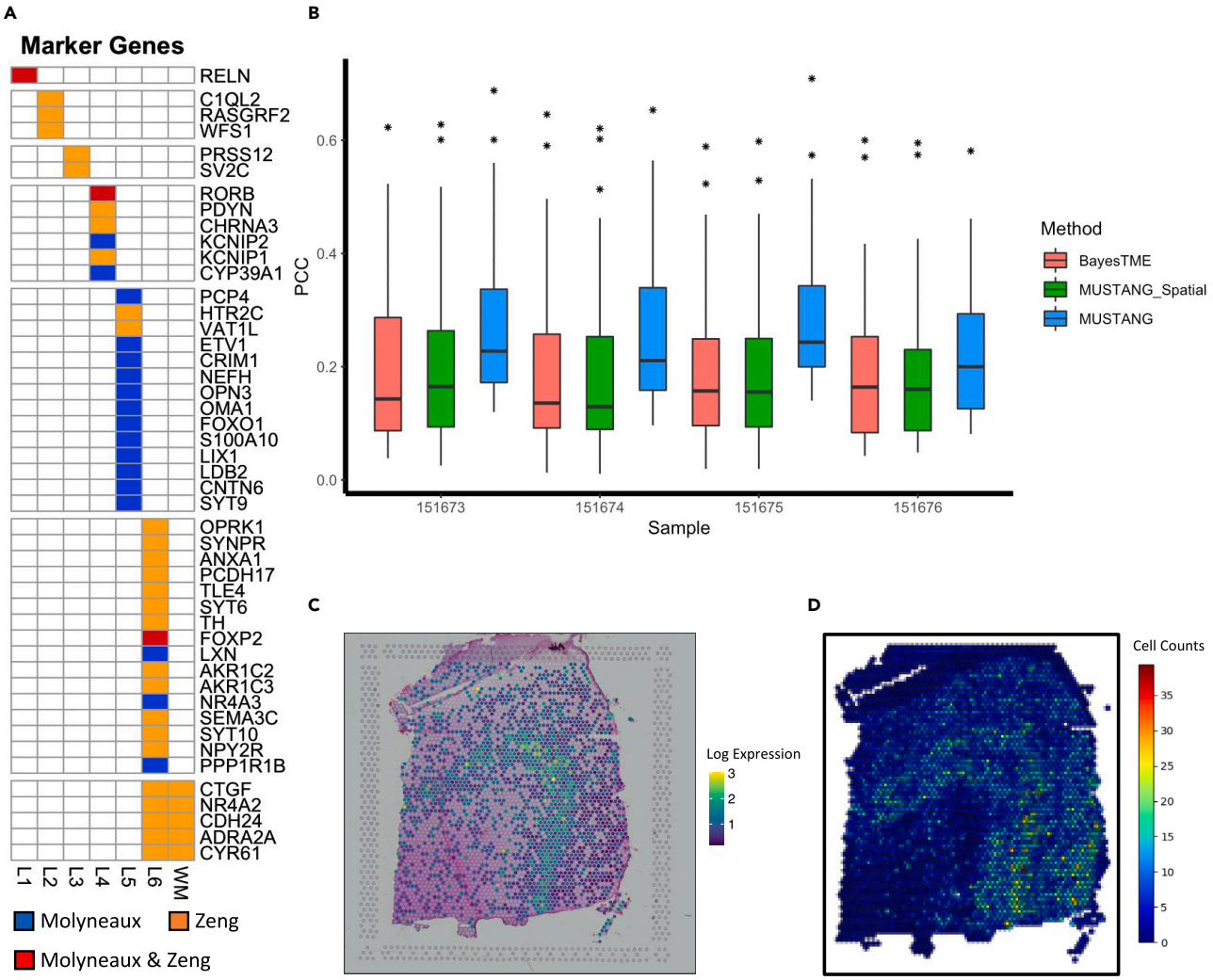

Figure 5. Quantitative performance benchmarking on four DLPFC tissue samples

(A) List of layer-specific gene markers from two brain tissue studies.[38,39]

(B) Boxplots showing the calculated PCC values for three different reference-free cell-type deconvolution methods: MUSTANG, MUSTANG_Spatial, and BayesTME. Higher PCC values indicate better deconvolution performance identifying annotated cell types.

(C and D) (C) Spot-level log2 expression visualization of the L5 layer marker gene PCP4 correlates with the spatial pattern of (D) MUSTANG-inferred cell numbers for the L5 layer best paired cell type for the sample 151674 (PCC = 0.42).

separately. As depicted in the figure, on all four tissue samples, jointly analyzing them with MUSTANG leads to higher average PCC values compared with separately deconvolving them using BayesTME. This superior performance of MUSTANG illustrates the benefit of simultaneously analyzing tissue samples with an approach that allows for effective cross-sample information sharing. As an example of the spatial expression pattern of the marker genes and inferred cell-type cell numbers, we have visualized the log2 expression of the *L5* layer marker gene *PCP4* as well as the MUSTANG-inferred cell numbers for the L5 layer best paired cell type for sample 151674 in Figures 5C and 5D, respectively. The derived PCC value for this gene is 0.42. Here, we would like to emphasize that, due to the nature of quantitative analysis we did in this section while STdeconvolve deconvolution model does not explicitly model cell type cell numbers (i.e., d_{sk} in our deconvolution model), it is not possible to benchmark STdencovolve with other comparison methods for the presented performance evaluation results. It worth mentioning that adjusting for this parameter during the deconvolution of aggregated ST signals in multi-cellular spot resolution ST datasets is crucial to assure biological considerations such as monotonic relationship between the number of cells and aggregated read measurement in each spot. As currently, to the best of our knowledge, only MUSTANG and BayesTME adjust for this source of variation, we have only included results of these methods in Figure 5B and excluded STdeconvolve from this quantitative analysis.

To better understand the corresponding contributions of different components in MUSTANG to its superior performance for multi-sample ST data analysis, we further conducted an ablation study that analyzes the tissue samples with a simplified version of MUSTANG without using the spot transcriptional

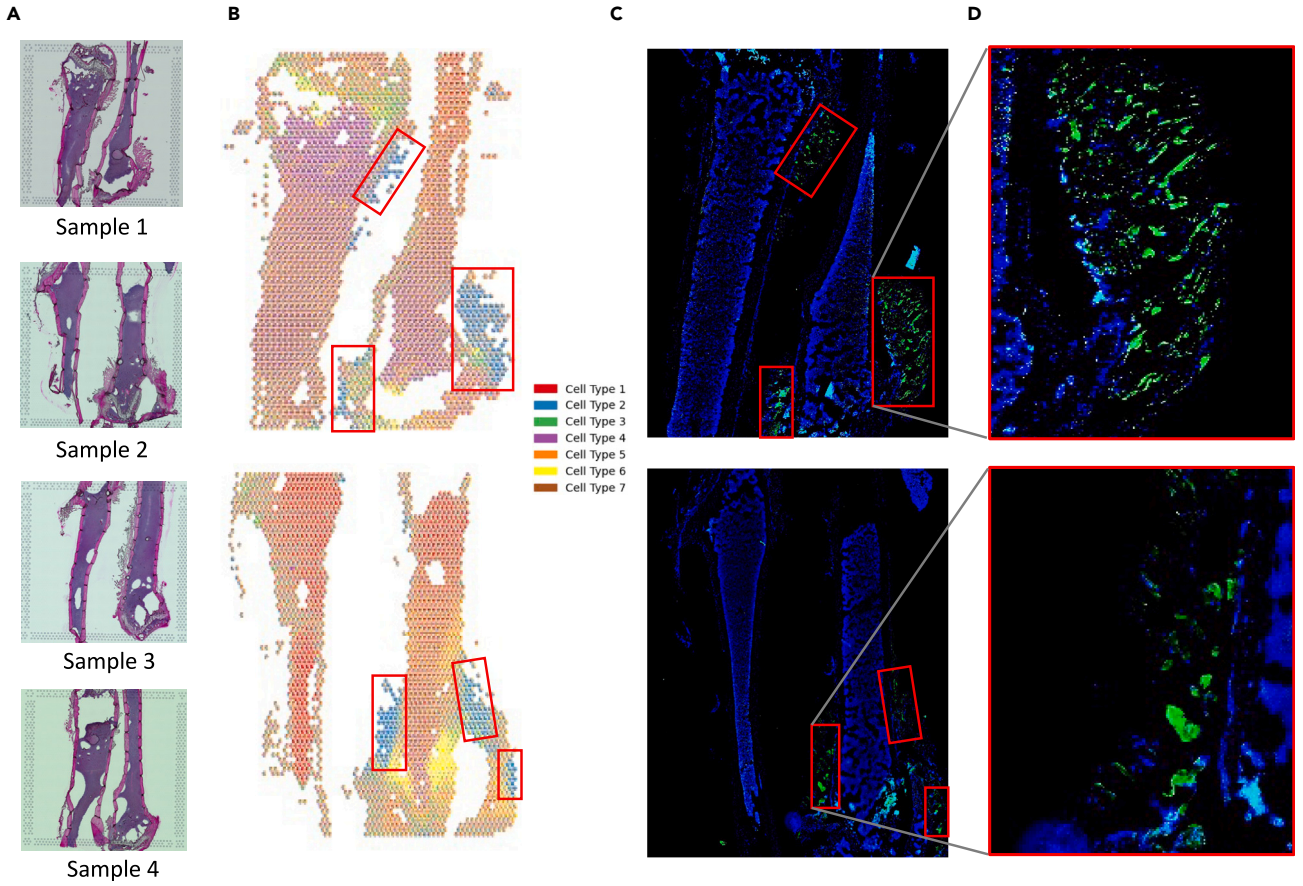

Figure 6. Analysis of four mouse bone marrow tissue samples with MUSTANG

(A) H&E staining images of the four samples profiles with the Visium platform.

(B) Spot-based spatial pie charts of MUSTANG-inferred cell-type proportions for (top) sample 1 and (below) sample 2.

(C) Matching IF staining images of (top) sample 1 and (below) sample 2.

(D) Closer look at the IF staining image regions with high density of green dots, indicating the presence of tumor cells.

adjacency matrix across samples. This means that we deconvolve tissue samples without cross-sample transcriptional information sharing. We call this simpler version of MUSTANG, "MUSTANG_Spatial" because, after removing transcriptional edges from spot similarity graph, it gets reduced to using only the spot spatial coordinates. As shown in Figure 5B, the PCC values in all four samples get significantly lower in the obtained results by MUSTANG_Spatial in comparison with those by the complete MUSTANG workflow. Clearly, removing transcriptional information sharing from MUSTANG leads to, on average, similar PCC values of the results using BayesTME, which deconvolves tissue samples separately. This is expected as BayesTME, similar to MUSTANG_Spatial, only allows within-sample information sharing across physically neighboring spots by performing spatial smoothing on cell-type assignment probabilities. This ablation study clarifies the significance of intra-sample transcriptional similarity guidance on boosting the performance of MUSTANG.

Mouse bone marrow data

The TME plays a critical role in tumor development, progression, and therapeutic response.[40] Recently, several studies have reported that the spatial organization of the TME is the key determinant of the disease behavior and treatment outcomes.[41,42] Thus, a comprehensive understanding of the spatial architecture and expression patterns of the TME holds great promise for the development of novel therapeutic treatment strategies. Taking advantage of the TME ST data helps unveil the underlying complex spatial organization and intricate interplay between tumor cells and their microenvironment.

For the final application of MUSTANG analyzing ST data of tissue samples, we study and characterize mouse bone marrow tissue TME. To obtain the ST data, we have profiled the bone tissue of 6- to 8-week mouse after bone lesions generation via the 10X Visium platform to profile four bone marrow tissue sections. The H&E staining images of the four bone tissue sections are shown in Figure 6A. The multi-sample ST data generation details can be found in supplemental information,section A.5.

To identify and characterize the spatial organization of tumor cells within the bone marrow tissue TME, we jointly analyze the ST data from the four bone tissue sections with MUSTANG. We follow the same MUSTANG workflow steps described in detail in the MUSTANG model overview section to infer the deconvolved components of the bone tissue samples. We pick the number of cell types K based on the results of applying

unsupervised cell-type number inference algorithms implemented in BayesTME[12] and STdeconvolve[15] to each of the individual four bone tissue samples leading to eight different inferred numbers of cell types. We then select K to be 7 as it is the most frequently inferred value of total cell-type numbers out of the eight derived values by BayesTME and STdeconvolve (four occurrences; details in the supplemental information).

After simultaneously analyzing the four bone tissue samples using MUSTANG, we plot the spatial scatter pie chart visualization of the inferred deconvolved cell-type proportions. The spatial pie chart plots for samples 1 and 2 are visualized in Figure 6B. To validate the identification of tumor cell types in the bone marrow TME by MUSTANG, we generate matched IF staining images for each bone tissue section separately. Specifically, the bone sections were stained with antibodies to depict the potential tumor cell-enriched tissue section parts (the detailed protocol for generation of IF staining images can be found in the supplemental information). The generated IF staining images for bone tissue samples 1 and 2 are shown in Figure 6C. The green dots in the IF staining images highlight the tumor cell-enriched parts (Figure 6D). Matching the green dots regions in IF staining images with the spatial pie chart plots of tissue samples from MUSTANG revealed the presence of high MUSTANG-inferred proportions of cell type 2 (colored blue in Figure 6B). We plot red boxes to highlight the regions of IF staining images of bone tissue samples with high enrichment of green dots (i.e., tumor cells) and overlay the boxes on the spatial pie charts. The spots in the matching red boxes of the spatial pie charts are composed of high inferred cell-type number 2 proportions with MUSTANG. This demonstrates the capability of MUSTANG to identify tumor cell-type cells in the bone marrow TME.

DISCUSSION

We have developed MUSTANG, a multi-sample ST data analysis workflow that jointly analyzes multiple tissue samples by leveraging transcriptional information sharing across samples as well as spatial dependency in gene expression patterns within samples. By our proposed workflow, including spot similarity graph construction and batch effect correction removing unwanted nuisance factors obscuring the inherent biological signal in ST data, the joint Bayesian decovolution model in MUSTANG extends the previous developments for reference-free single-sample ST data analysis[12] to joint multi-sample ST data analysis, allowing for the robust simultaneous spatial characterization of cell sub-populations across spots in all tissue samples. We have introduced a new spot-based knowledge graph, spot similarity graph, that captures sufficient and comprehensive similarity information between spots to be used in our joint Bayesian deconvolution model to improve the multi-sample analysis performance beyond existing methods analyzing single ST samples separately. By providing extensive results on a simulated and three real-world multi-sample ST data, we have demonstrated the superior performance of MUSTANG in terms of cell-type deconvolution and spatial characterization of complex tissue environments. Future work concerns further improving the capability of MUSTANG to decipher tissue structures by performing joint cell-cell interaction analysis between cells of different sub-populations across multi-sample tissue samples.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Xiaoning Qian (xqian@ece.tamu.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

All data used in the manuscript are publicly available and are referenced in the article. Specifically, the sagittal mouse brain ST data are accessible on the 10X Genomics website at https://support.10xgenomics.com/spatial-gene-expression/datasets. The human brain ST data samples are available using the fetch_data() function from spatialLIBD R package. The code for the software and tutorials for reproducing the results is available at https://github.com/namini94/MUSTANG. Long-term archive of code repository is made available via Zenodo at https://doi.org/10.5281/zenodo.10818888.[43]

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Gibbs sampling inference

Here, we provide the detailed posterior Gibbs sampling procedure for the joint Bayesian deconvolution model described in the MUSTANG model overview section.

Sampling Y_{sgk}. Since we are modeling the raw ST counts Y_{sg} as

$$Y_{sg} = \sum_k Y_{sgk} \sim \text{Pois}\left(\sum_k \beta_k d_{sk} \varphi_{kg}\right), \quad \text{(Equation 4)}$$

and leveraging the relationship between the Poisson and multinomial distribution, the Y_{sgk} parameters can be sampled from a multinomial distribution. If we define the auxiliary variables $\pi_{sk} = \frac{\beta_k d_{sk} \varphi_{kg}}{\sum_k \beta_k d_{sk} \varphi_{kg}}$, then

$$(Y_{sg1}, \ldots, Y_{sgK} | -) \sim \text{Mult}(Y_{sg}; \pi_{s1}, \ldots, \pi_{sK}). \quad \text{(Equation 5)}$$

Sampling β_k. To infer the cell-type-dependent expected total counts parameter β_k, we write its posterior as

$$\begin{aligned} P(\beta_k | Y_{sgk}, d_{sk}, \varphi_{kg}) &= \prod_s \prod_g P(Y_{sgk} | \beta_k, d_{sk}, \varphi_{kg}) P(\beta_k) \\ &\propto \prod_s P(Y_{sk\cdot} | \beta_k, d_{sk}, \varphi_{kg}) P(\beta_k) \end{aligned} \quad \text{(Equation 6)}$$

where $Y_{sk\cdot}$ is $(Y_{sk1}, \ldots, Y_{skG})$. Then, we can write the likelihood of reads $Y_{sk\cdot}$ as

$$\begin{aligned} P(Y_{sk\cdot} | \beta_k, d_{sk}, \varphi_{kg}) &= \prod_g \frac{\exp(-\beta_k d_{sk} \varphi_{kg}) (\beta_k d_{sk} \varphi_{kg})^{Y_{skg}}}{Y_{sgk}!} \\ &= \frac{\exp\left(-\beta_k d_{sk} \sum_g \varphi_{kg}\right) (\beta_k d_{sk})^{\sum_g Y_{skg}} \prod_g \varphi_{kg}^{Y_{skg}}}{\prod_g Y_{skg}!} \\ &= \frac{\exp(-\beta_k d_{sk}) (\beta_k d_{sk})^{Y_{sk}} \prod_g \varphi_{kg}^{Y_{skg}}}{\prod_g Y_{skg}!} \end{aligned} \quad \text{(Equation 7)}$$

where in the last equation we take advantage of facts that $\sum_g \varphi_{kg} = 1$ and $\sum_g Y_{skg} = Y_{sk}$. Now, based on Equation 7, we can simplify the posterior of cell-type-dependent parameter β_k in Equation 6 as

$$\begin{aligned} P(\beta_k | Y_{sgk}, d_{sk}, \varphi_{kg}) &\propto \prod_s P(Y_{sk\cdot} | \beta_k, d_{sk}, \varphi_{kg}) P(\beta_k) \\ &= \exp\left(-\beta_k \sum_s d_{sk}\right) \beta_k^{\sum_s Y_{sk}} \\ &\quad \left(\prod_s \frac{\prod_g \varphi_{kg}^{Y_{skg}}}{\prod_g Y_{skg}!} d_{sk}^{Y_{sk}}\right) P(\beta_k) \\ &= \exp\left(-\beta_k \sum_s d_{sk}\right) \beta_k^{\sum_s Y_{sk}} \end{aligned}$$

$$\left(\prod_s \frac{\prod_g \varphi_{kg}^{Y_{skg}}}{\prod_g Y_{skg}!} d_{sk}^{Y_{sk}}\right)$$

$$\left(\frac{f^e}{\Gamma(e)} \beta_k^{e-1} \exp(-f\beta_k)\right)$$

$$\propto \exp\left(-\beta_k\left(\sum_s d_{sk} + f\right)\right) \beta_k^{\sum_s Y_{sk}+e-1}. \quad \text{(Equation 8)}$$

Note that, in Equation 8, we leverage the Gamma prior distribution (i.e., Gamma(e, f)) we imposed on β_k as described in the main text. Thus, based on Equation 8 we can update the β_k as

$$(\beta_k|-) \sim \text{Gamma}\left(\sum_s Y_{sk} + e, \sum_s d_{sk} + f\right). \quad \text{(Equation 9)}$$

Sampling φ_k. As described in the main text, we impose Dirichlet prior distribution over the normalized cell-type-dependent gene expression profile parameter $\varphi_k = (\varphi_{k1}, \ldots, \varphi_{kG})$ (i.e., $\varphi_k \sim \text{Dir}(\alpha_k)$) and $\sum_g \varphi_{kg} = 1$. We have

$$\begin{aligned} Y_{sk1}, \ldots, Y_{skG} &\sim \text{Mult}\left(\sum_g Y_{skg}; \frac{\beta_k d_{sk}\varphi_{k1}}{\sum_g \beta_k d_{sk}\varphi_{kg}}, \ldots, \frac{\beta_k d_{sk}\varphi_{kG}}{\sum_g \beta_k d_{sk}\varphi_{kg}}\right) \\ &= \text{Mult}\left(\sum_g Y_{skg}; \varphi_{k1}, \ldots, \varphi_{kG}\right) \end{aligned} \quad \text{(Equation 10)}$$

Thus, the normalized gene expression profiles can be updated using the Dirichlet-multinomial conjugacy as

$$(\varphi_k|-) \sim \text{Dir}\left(\alpha_k + \sum_s Y_{sk1}, \ldots, \alpha_k + \sum_s Y_{skG}\right) \quad \text{(Equation 11)}$$

Sampling D_s and d_{sk}. By modeling the cell number distribution as a hidden Markov model and exploiting the forward-filtering backward-sampling algorithm introduced in Zhang et al.,[12] we can update d_{sk} in an efficient approach. Specifically, in the forward-filtering algorithm we calculate the "alpha" values of our hidden latent stats, which includes the cell-type cell numbers (i.e., x_k), which we define as

$$\alpha(x_k) = P(d_{sk}, Y_{s1:k}) \quad \text{(Equation 12)}$$

and in the backward-sampling, based on the derivations in Zhang et al.,[12] the cell-type cell number values are updated based on

$$P(d_{sk}|x_{k+1}, Y_{1:T}) \propto \alpha(x_k) P(x_{k+1}|x_k). \quad \text{(Equation 13)}$$

Additional results with semi-synthetic data

In this section, we present additional results and data demonstrations to comprehensively report the results in the semi-synthetic ST data experiment. Figure S1A illustrates the simulated semi-synthetic multi-cell per spot samples generated from the STARmap mouse visual cortex data with window sizes of 700, 650, and 600 pixels. Furthermore, in Table S1, the number of spots, amounts of cells per spot, and number of genes are reported for each of the simulated four samples with varying grid sizes. Figure S1B visualizes the MUSTANG-estimated proportions of some of the major clusters in simulated spots when jointly analyzing the four samples with MUSTANG. As indicated, the inferred spatial patterns match the ground-truth proportions for all four samples in each cluster. PCC, SSIM, RMSE, and JSD values for the cell-type composition of the spots simulated from STARmap mouse visual cortex data for all clusters from MUSTANG, BayesTME, STdeconvolve, Cell2location, and RCTD are visualized in the boxplots in Figure S1C, with center lines as median and green triangle as mean. For PCC and SSIM values, higher is better, and for RMSE and JSD metrics, lower is better.

Reference anatomical regions of mouse brain tissue

The annotations for the major anatomical regions of the sagittal mouse brain are extracted from the Allen Brain Atlas[32,33] and illustrated in Figure S2. When we analyze the mouse brain ST data with MUSTANG, we consider the number of cell types K to be 11 covering the olfactory bulb, cortex, striatum, pallidum, hippocampus, thalamus, hypothalamus, midbrain, pons, medulla, and cerebellum regions based on the reference mouse brain annotations.

Inferring total number of cell types (K) and spatial smoothness (λ) hyperparameters

Here, we describe how one can select the two adjustable hyperparameters in MUSTANG's Bayesian deconvolution model: K, the total number of cell types, and λ, the spatial smoothness parameter. We specifically illustrate the hyperparameter tuning process on the mouse bone marrow ST data but one can repeat the procedure for any arbitrary multi-sample ST dataset to derive the ideal values for the hyperparameters.

First, we describe the results of applying unsupervised cell-type number inference algorithms implemented in BayesTME[12] and STdeconvolve[15] to each of the individual four mouse bone marrow tissue samples. Based on the instructions in Miller et al.,[15] to find optimal number of cell types in bone tissue samples with STdeconvolve, we fit a number of different LDA models with different K values and then, based on the inferred number of "rare" predicted cell types and perplexity values, we pick the number of cell types. Specifically, we change K from 2 to 15 for each bone tissue sample and plot the perplexity and number of "rare" predicted cell types versus the K values. Figure S3 shows the STdeconvolve inferred perplexity and number of "rare" cell types versus different K values for four bone marrow samples 1 to 4, respectively. As described in STdeconvolve workflow,[15] we pick the number of cell types to be the value from that perplexity stabilizes and has the lowest number of rare sub-predicted cell types to avoid over-clustering. This leads to inferring 6, 7, 6, and 7 numbers of cell types for samples 1–4, respectively (Table S2).

Then, we use BayesTME to infer total number of cell types (K) and the degree of spatial smoothness (λ). Specifically, BayesTME does this by performing 5-fold cross-validation for each $K = (2, \ldots, 12)$ values with 5% of spots held out in each fold. Then, in each fold, a Poisson-based discrete deconvolution model is fitted over a discrete grid of λ smoothness values $(10^0, 10^1, \ldots, 10^5)$ and average log likelihood for the held out spots are calculated. Finally, the K with highest averaged likelihood is picked to be the total number of cell types and the value of λ with average cross-validation log likelihood closest to the overall average will be selected as the ideal λ for the sample under study.[12] Figure S4 shows the calculated average cross-validation log likelihood versus the number of cell types for each of four bone tissue samples. Based on these figures, the inferred total numbers of cell types for samples 1 to 4 are 8, 7, 7, and 8, respectively (Table S2).

Table S2 summarizes the inferred total number of cell types from STdeconvolve and BayesTME. We then select the K to be 7 in our multi-sample analysis with MUSTANG as it is the most frequently inferred value of total cell-type numbers out of the eight derived values.

Furthermore, as illustrated in Figure S4, $\lambda = 1,000$ has the closest average cross-validation log likelihood to the overall average (the bold black graph corresponding to λ_{mean}) for all four samples. We then pick λ to be 1,000 in our multi-sample Bayesian deconvolution analysis with MUSTANG as it is the most frequently inferred value of spatial smoothness degree.

Mouse bone marrow ST data generation details

Here, we explain the mouse bone marrow TME ST data generation details and protocols. To generate the data, we have profiled the bone tissue of 6- to 8-week mice after bone lesion generation by intra-iliac injection. For spatial analysis, ST data are obtained via the 10X Visium platform to profile four bone marrow tissue sections. Specifically, thin (10 μm) mouse bone marrow sections were mounted directly onto separate designated capture areas on the 10X Visium spatial gene expression slides and data preprocessing was done per the manufacturer's protocols. In brief, after H&E staining, each section was imaged using color brightfield by Cytation 5. The sections were then processed following the 10X Visium gene expression protocols until the cDNA libraries were constructed, which were later sequenced using the Novaseq 6000 system with 150 bp paired-end reads, aiming at 300 million raw reads per section. The H&E staining images of the four bone tissue sections are shown in Figure 6A. The Visium Spatial Gene Expression Solution from 10X Genomics allows for the analysis of mRNA using high-throughput sequencing and subsequently maps a transcript's expression pattern in tissue sections

using high-resolution microscope imaging. This provides gene expression data at 5,000 capture spots in each Visium slide within the context of tissue architecture, tissue microenvironments, and cell groups. SpaceRanger was used to process Visium spatial RNA-seq output and bright-field and fluorescence microscope images to detect tissue, align reads, and generate feature-spot matrices. SpaceRanger built-in function `mkfastq` was used to wrap Illumina's `bcl2fastq` to correctly demultiplex Visium-prepared sequencing runs and to convert barcode and read data to FASTQ files. SpaceRanger function `count` was used to take a microscope slide image and FASTQ files from SpaceRanger `mkfastq` and perform alignment, tissue detection, fiducial detection, and barcode/UMI counting. In our study, raw sequence reads were mapped to mice reference genome (mm10) to obtain the gene expression profile at each spot.

IF staining images generation protocol

Here, we describe the protocols for IF staining of thick sections and bone clearing. In brief, femur bone sections were cleaned, pretreated with 1 mg/mL sodium borohydride solution, and then blocked before whole-mount staining. Then, the bone sections were stained with antibodies. IF staining was performed in 1 mL staining buffer for 3 days at 4°C with constant rotation and followed by a whole day of PBS washing. The stained samples were then dehydrated by a series of methanol solutions before being completely cleared by BABB solution. The bone sections were later sealed in imaging glass cassettes with BABB solution. The images were taken using an Olympus FV1200 MPE confocal microscope.

Additional results with mouse bone marrow data

Here, we present the additional results of jointly analyzing four bone tissue samples as well as the IF staining images for the profiled tissue samples, which highlights the tumor cells. Specifically, here, we focus on mouse bone marrow tissue samples 3 and 4 as the results of the other two samples are discussed in detail in the mouse bone marrow ST data analysis section. Figure S5A shows the spatial pie chart plots generated by MUSTANG for samples three and four and same as what we described in the mouse bone marrow ST data analysis section, the IF staining images are generated and used to validate MUSTANG results by identifying tumor cells in bone marrow TME. Figure S5B shows the matched IF staining images for bone marrow tissue samples 3 and 4. As the figures suggest, the green dots that highlight the tumor cells regions can be matched with the tissue areas in samples 3 and 4 that have high proportions of cells of cell type 2, illustrating the capability of MUSTANG to characterize tumor cells in mouse bone marrow TME.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2024.100986.

ACKNOWLEDGMENTS

The authors acknowledge support from National Science Foundation grants CCF-1553281 and IIS-2212419. In addition, the presented work is supported in part by the US Department of Energy, Office of Science, Biological and Environmental Research (BER) program under B&R number KP1601017.

AUTHOR CONTRIBUTIONS

Methodology, S.N., J.S., S.T.C.W., and X.Q.; data curation, Y.C., X.Z., Z.X., and L.W.; data analysis, S.N.; writing – review & editing, S.N., J.S., and X.Q.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 18, 2023
Revised: January 25, 2024
Accepted: April 10, 2024
Published: May 2, 2024

REFERENCES

1. Hwang, B., Lee, J.H., and Bang, D. (2018). Single-cell rna sequencing technologies and bioinformatics pipelines. Exp. Mol. Med. *50*, 1–14.
2. Niyakan, S., Hajiramezanali, E., Boluki, S., Zamani Dadaneh, S., and Qian, X. (2021). Simcd: Simultaneous clustering and differential expression analysis for single-cell transcriptomic data. Preprint at arXiv. https://doi.org/10.48550/arXiv.2104.01512.
3. Niyakan, S., Yoon, B.-J., Qian, X., and Luo, X. (2024). Biologically Interpretable VAE with Supervision for Transcriptomics Data Under Ordinal Perturbations. Preprint at bioRxiv. https://doi.org/10.1101/2024.03.28.587231.
4. Zhao, E., Stone, M.R., Ren, X., Guenthoer, J., Smythe, K.S., Pulliam, T., Williams, S.R., Uytingco, C.R., Taylor, S.E.B., Nghiem, P., et al. (2021). Spatial transcriptomics at subspot resolution with bayesspace. Nat. Biotechnol. *39*, 1375–1384.
5. Miller, B.F., Bambah-Mukku, D., Dulac, C., Zhuang, X., and Fan, J. (2021). Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomic data with nonuniform cellular densities. Genome Res. *31*, 1843–1855.
6. Moses, L., and Pachter, L. (2022). Museum of spatial transcriptomics. Nat. Methods *19*, 534–546.
7. Marx, V. (2021). Method of the year: spatially resolved transcriptomics. Nat. Methods *18*, 9–14.
8. Walker, B.L., Cang, Z., Ren, H., Bourgain-Chang, E., and Nie, Q. (2022). Deciphering tissue structure and function using spatial transcriptomics. Commun. Biol. *5*, 220.
9. 10x Genomics (2022). 10x genomics: Visium spatial gene expression. https://www.10xgenomics.com/products/spatial-gene-expression.
10. Rodriques, S.G., Stickels, R.R., Goeva, A., Martin, C.A., Murray, E., Vanderburg, C.R., Welch, J., Chen, L.M., Chen, F., and Macosko, E.Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science *363*, 1463–1467.
11. Vickovic, S., Eraslan, G., Salmén, F., Klughammer, J., Stenbeck, L., Schapiro, D., Äijö, T., Bonneau, R., Bergenstråhle, L., Navarro, J.F., et al. (2019). High-definition spatial transcriptomics for in situ tissue profiling. Nat. Methods *16*, 987–990.
12. Zhang, H., Hunter, M.V., Chou, J., Quinn, J.F., Zhou, M., White, R.M., and Tansey, W. (2023). Bayestme: An end-to-end method for multiscale spatial transcriptional profiling of the tissue microenvironment. Cell Syst. *14*, 605–619.e7.
13. Tu, J.J., Li, H.S., Yan, H., and Zhang, X.F. (2023). Endecon: cell type deconvolution of spatially resolved transcriptomics data via ensemble learning. Bioinformatics *39*, btac825.
14. Ma, Y., and Zhou, X. (2022). Spatially informed cell type deconvolution for spatial transcriptomics. Nat. Biotechnol. *40*, 1349–1359.
15. Miller, B.F., Huang, F., Atta, L., Sahoo, A., and Fan, J. (2022). Reference-free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data. Nat. Commun. *13*, 2339.
16. Kleshchevnikov, V., Shmatko, A., Dann, E., Aivazidis, A., King, H.W., Li, T., Elmentaite, R., Lomakin, A., Kedlian, V., Gayoso, A., et al. (2022). Cell2location maps fine-grained cell types in spatial transcriptomics. Nat. Biotechnol. *40*, 661–671.
17. Lopez, R., Li, B., Keren-Shaul, H., Boyeau, P., Kedmi, M., Pilzer, D., Jelinski, A., Yofe, I., David, E., Wagner, A., et al. (2022). Destvi identifies continuums of cell types in spatial transcriptomics data. Nat. Biotechnol. *40*, 1360–1369.
18. Cable, D.M., Murray, E., Zou, L.S., Goeva, A., Macosko, E.Z., Chen, F., and Irizarry, R.A. (2022). Robust decomposition of cell type mixtures in spatial transcriptomics. Nat. Biotechnol. *40*, 517–526.
19. Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I., and Heyn, H. (2021). Spotlight: seeded nmf regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. Nucleic Acids Res. *49*, e50.

20. Charytonowicz, D., Brody, R., and Sebra, R. (2023). Interpretable and context-free deconvolution of multi-scale whole transcriptomic data with unicell deconvolve. Nat. Commun. *14*, 1350.

21. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J.E., Mestdagh, P., and De Preter, K. (2020). Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nat. Commun. *11*, 5650.

22. Andersson, A., Larsson, L., Stenbeck, L., Salmén, F., Ehinger, A., Wu, S.Z., Al-Eryani, G., Roden, D., Swarbrick, A., Lundeberg, J., et al. (2021). Spatial deconvolution of her2-positive breast cancer delineates tumor-associated cell type interactions. Nat. Commun. *12*, 6012.

23. Mantri, M., Scuderi, G.J., Abedini-Nassab, R., Wang, M.F.Z., McKellar, D., Shi, H., Grodner, B., Butcher, J.T., and De Vlaminck, I. (2021). Spatiotemporal single-cell rna sequencing of developing chicken hearts identifies interplay between cellular differentiation and morphogenesis. Nat. Commun. *12*, 1771.

24. Maynard, K.R., Collado-Torres, L., Weber, L.M., Uytingco, C., Barry, B.K., Williams, S.R., Catallini, J.L., Tran, M.N., Besich, Z., Tippani, M., et al. (2021). Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nat. Neurosci. *24*, 425–436.

25. Allen, C., Chang, Y., Ma, Q., and Chung, D. (2022). Maple: A hybrid framework for multi-sample spatial transcriptomics data. Preprint at bioRxiv. https://doi.org/10.1101/2022.02.28.482296.

26. Joglekar, A., Prjibelski, A., Mahfouz, A., Collier, P., Lin, S., Schlusche, A.K., Marrocco, J., Williams, S.R., Haase, B., Hayes, A., et al. (2021). A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. Nat. Commun. *12*, 463.

27. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.R., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with harmony. Nat. Methods *16*, 1289–1296.

28. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. *2008*, P10008.

29. Tansey, W., Athey, A., Reinhart, A., and Scott, J.G. (2017). Multiscale spatial density smoothing: an application to large-scale radiological survey and anomaly detection. J. Am. Stat. Assoc. *112*, 1047–1063.

30. Xu, Z., Schmidt, D.F., Makalic, E., Qian, G., and Hopper, J.L. (2016). Bayesian grouped horseshoe regression with application to additive models. In AI 2016: Advances in Artificial Intelligence: 29th Australasian Joint Conference Proceedings, *29*, pp. 229–240.

31. Wang, Y.-X., Sharpnack, J., J. Smola, A., and Tibshirani, R.J. (2016). Trend filtering on graphs. J. Mach. Learn. Res. *17*, 1–41.

32. Daigle, T.L., Madisen, L., Hage, T.A., Valley, M.T., Knoblich, U., Larsen, R.S., Takeno, M.M., Huang, L., Gu, H., Larsen, R., et al. (2018). A suite of transgenic driver and reporter mouse lines with enhanced brain-cell-type targeting and functionality. Cell *174*, 465–480.e22.

33. Lein, E.S., Hawrylycz, M.J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J., et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. Nature *445*, 168–176.

34. Li, B., Zhang, W., Guo, C., Xu, H., Li, L., Fang, M., Hu, Y., Zhang, X., Yao, X., Tang, M., et al. (2022). Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. Nat. Methods *19*, 662–670.

35. Maynard, K.R., Collado-Torres, L., Weber, L.M., Uytingco, C., Barry, B.K., Williams, S.R., Catallini, J.L., 2nd, Tran, M.N., Besich, Z., Tippani, M., et al. (2021). Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nat. Neurosci. *24*, 425–436.

36. McInnes, L., Healy, J., and Melville, J. (2020). Umap: Uniform manifold approximation and projection for dimension reduction. Preprint at arXiv. https://doi.org/10.48550/arXiv.1802.03426.

37. Li, H., Zhou, J., Li, Z., Chen, S., Liao, X., Zhang, B., Zhang, R., Wang, Y., Sun, S., and Gao, X. (2023). A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics. Nat. Commun. *14*, 1548.

38. Molyneaux, B.J., Arlotta, P., Menezes, J.R.L., and Macklis, J.D. (2007). Neuronal subtype specification in the cerebral cortex. Nat. Rev. Neurosci. *8*, 427–437.

39. Zeng, H., Shen, E.H., Hohmann, J.G., Oh, S.W., Bernard, A., Royall, J.J., Glattfelder, K.J., Sunkin, S.M., Morris, J.A., Guillozet-Bongaarts, A.L., et al. (2012). Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. Cell *149*, 483–496.

40. Kalbasi, A., and Ribas, A. (2020). Tumour-intrinsic resistance to immune checkpoint blockade. Nat. Rev. Immunol. *20*, 25–39.

41. Fu, T., Dai, L.J., Wu, S.Y., Xiao, Y., Ma, D., Jiang, Y.Z., and Shao, Z.M. (2021). Spatial architecture of the immune microenvironment orchestrates tumor immunity and therapeutic response. J. Hematol. Oncol. *14*, 98.

42. Blise, K.E., Sivagnanam, S., Banik, G.L., Coussens, L.M., and Goecks, J. (2022). Single-cell spatial architectures associated with clinical outcome in head and neck squamous cell carcinoma. npj Precis. Oncol. *6*, 10.

43. Niyakan, S., and Sheng, J. (2024). Mustang: Multi-sample spatial transcriptomics data analysis. Zenodo. https://doi.org/10.5281/zenodo.10818888.