SPICEMIX: Integrative single-cell spatial modeling for inferring cell identity

Benjamin Chidester[1,#], Tianming Zhou[1,#], and Jian Ma[1,*]

[1]Computational Biology Department, School of Computer Science,
Carnegie Mellon University, Pittsburgh, PA 15213, USA
[#]These two authors contributed equally.
[*]Correspondence: jianma@cs.cmu.edu

Abstract

Spatial transcriptomics technologies promise to reveal spatial relationships of cell-type composition in complex tissues. However, the development of computational methods that capture the unique properties of single-cell spatial transcriptome data to unveil cell identities remains a challenge. Here, we report SPICEMIX, a new method based on probabilistic, latent variable modeling that enables effective joint analysis of spatial information and gene expression of single cells from spatial transcriptome data. Both simulation and real data evaluations demonstrate that SPICEMIX markedly improves upon the inference of cell types compared with existing approaches. Applications of SPICEMIX to single-cell spatial transcriptome data of the mouse primary visual cortex acquired by seqFISH+ and STARmap show that SPICEMIX can enhance the inference of cell identities and uncover potentially new cell subtypes with important biological processes. SPICEMIX is a generalizable framework for analyzing spatial transcriptome data to provide critical insights into the cell-type composition and spatial organization of cells in complex tissues.

Introduction

The compositions of different cell types in various human tissues remain poorly understood due to the complex interplay among intrinsic, spatial, and temporal factors that collectively contribute to cell identity [1–3]. Single-cell RNA-seq (scRNA-seq) has greatly advanced our understanding of complex cell types in different tissues [4–6], but its utility in disentangling spatial factors in particular is inherently limited by the dissociation of cells from their spatial context. To address this limitation, new spatial transcriptomics technologies based on multiplexed imaging and sequencing [7–17] are able to reveal spatial information of gene expression of dozens to tens of thousands of genes in individual cells *in situ* within the tissue context.

However, the development of computational methods that capture the unique properties of the spatially resolved single-cell transcriptome data to unveil single-cell identities remains a challenge [18]. Zhu et al. [19] previously proposed the use of a hidden Markov random field (HMRF) to model spatial domains after distinguishing spatial and intrinsic genes (based on scRNA-seq). The major drawback of the method of [19] is that it cannot learn contributions of spatial and intrinsic factors to gene expression directly from spatial transcriptome data. In addition, the model relies on the assumptions that spatial subtypes are discrete and exhibit homogeneous spatial patterns, which prohibits it from learning the underlying mixture of diverse factors of cell identity with varied spatial patterns (e.g., distinct layer-like structures or diffuse patterns). Several other methods have been developed to study the relationship of known cell types in local neighborhoods [20], to explore the spatial variance of genes [21–24], and to align scRNA-seq with spatial transcriptome data [25–27]. But no existing method seeks to jointly model spatial patterns of the cells and their expression profiles to reveal cell identity, which is of vital importance to fully utilize spatial transcriptome data.

Here, we report SPICEMIX (Spatial Identification of Cells using Matrix Factorization), a new integrative framework to model spatial transcriptome data. SPICEMIX uses latent variable modeling to express the interplay of spatial and intrinsic factors that comprise cell identity. Crucially, SPICEMIX enhances the non-negative matrix factorization (NMF) [28] of gene expression with a novel integration with the graphical representation of the spatial relationship of cells. Thus, the learned spatial patterns can elucidate the relationship of intrinsic and spatial factors, leading to much more meaningful representations of cell identity. Application to the spatial transcriptome data of the mouse primary visual cortex acquired by seqFISH+ [12] and STARmap [13] demonstrated that the latent representations learned by SPICEMIX can refine the identification of cell types, uncover subtypes missed by other approaches, and reveal important biological processes. SPICEMIX has the potential to provide critical new insights into the cell composition based on spatial transcriptome data.

Results

Overview of SPICEMIX

SPICEMIX models the cell-to-cell relationships of the spatial transcriptome by a new probabilistic graphical model formulation, the NMF-HMRF (Fig. 1). The input of the model consists of gene expression measurements and spatial coordinates of cells from spatial transcriptome data (e.g., seqFISH+ [12] and STARmap [13]). From the spatial coordinates, an undirected graph is constructed to capture pairwise spatial relationships, where each cell is a node in the graph. For each node, a latent state vector explains

the observed gene expression of the cell. More critically, unique to the NMF-HMRF model is the integration of NMF into the HMRF [29] to represent the observations as mixtures of latent factors, modeled by metagenes, where the proportions are the hidden states of the graph. In contrast, in a standard HMRF, hidden states are assumed to be discrete, thus restricting the expressiveness of the model.

In the NMF-HMRF model of SPICEMIX, the potential functions of the graph capture the probabilistic relationships between variables in the model. The potential functions for observations capture the likelihood of the observation given the hidden state of the cell. The potential functions for edges capture the spatial affinity between the metagene proportions of neighboring cells. In a standard HMRF, it is assumed that neighboring nodes will have similar hidden states, resulting in a spatial smoothing effect that is inadequate to describe the heterogeneous spatial patterns of the cells. However, in the formulation of SPICEMIX, we do not assume such a relationship *a priori*, but rather allow the method to learn spatial affinities from the spatial transcriptome data. Crucially, SPICEMIX learns the parameters of the model that best explain the input spatial transcriptome data, while simultaneously learning the underlying metagenes and their proportions that define the identities of the cells. This is achieved by a new optimization algorithm that alternates between maximizing the joint posterior distribution of the parameters in the model and maximizing the posterior distribution of the metagenes in the matrix factorization. The learned parameters, metagenes, and proportions provide biological insights into the latent representation. See Methods for the detailed description of the SPICEMIX model.

Evaluation of SPICEMIX on simulated spatial transcriptome data

We first evaluated SPICEMIX on simulated data that we designed to model the mouse cortex, which has served as a prominent case study for several spatial transcriptomic methods, including seqFISH+ [12] and STARmap [13] (Fig. 2a-b; see Methods for detailed simulation strategy). This region of the brain consists of cell types that exhibit strong, layer-wise patterns of expression as well as cell types that sparsely populate the entire tissue. The goal of the evaluation was to infer the latent metagenes describing gene expression and to reveal the underlying simulated cell types. We compared the inference of SPICEMIX to that of NMF and HMRF, since they are the fundamental underlying models of many relevant computational methods. This comparison also aimed to demonstrate the advantage of the integration of these two models in SPICEMIX, rather than using either alone. We assessed performance by quantitatively comparing the cell types learned from each method with the simulated true cell types, using the adjusted Rand index (ARI). For SPICEMIX and NMF, we applied additional hierarchical clustering to the learned latent representation to group cells into clusters. The number of clusters was determined objectively by maximizing the Calinski-Harabasz (CH) index [30]. The strategy for choosing other hyperparameters for SPICEMIX and NMF is described in Methods. The number of clusters, or discrete states, for HMRF was chosen automatically during operation, given an upper bound, and the smoothing parameter was chosen manually to maximize the ARI, representing its best-case performance. We devised four simulation scenarios for evaluation, which varied the randomness of the data in terms of both the noise variance and the variance of the true hidden states (see Methods).

We found that SPICEMIX consistently produced the best ARI score (0.6-0.8 on average; the maximum value being 1.0) across all scenarios (Fig. 2d). In contrast, NMF achieved an ARI between 0.2-0.4 on average, a reduction by more than 50%. As expected, as the variance of the expression values or hidden states increased, the performance of all methods decreased (Fig. 2d). To ensure that the CH index was not favorably biased towards SPICEMIX, we also evaluated NMF when the number of clusters was instead chosen to maximize the ARI directly rather than according to the CH index (denoted as "NMF*"

in Fig. 2d). The resulting ARI ranged from 0.5-0.65. Thus, the best-case scenario for NMF was significantly worse than the performance of SPICEMIX. In addition, HMRF achieved a far lower ARI, between 0.1-0.2 on average. Looking closer at an example simulated sample reveals that the superior cell type inference of SPICEMIX was due to its successful recovery of both layer-specific and sparse spatial patterns of metagenes (Fig. 2c; metagene 8 shows layer-specific localization whereas metagene 2 has a more diffuse pattern). The precise recovery of these metagenes lead to a much clearer separation of the simulated cell types in the learned latent space of SPICEMIX (Fig. 2f). Notably, this resulted in a clear and accurate delineation of the layer-specific excitatory neurons in the sample (Fig. 2e). We found that, in contrast, the metagenes learned by NMF lacked spatial coherence (Fig. 2c). Consequently, NMF often failed to reveal the excitatory neurons according to their layer-specific enrichment (Fig. 2e). Also, in contrast to both SPICEMIX and NMF, HMRF smoothed over sparse cell types and yet still failed to detect clear layer-wise boundaries (Fig. 2e), despite having optimized the smoothing parameter. Specifically, the spatial patterns of the boundaries between HMRF clusters are not consistent with the ground truth (dashed vertical lines in Fig. 2e), especially in layer L4, where green, yellow, and blue cell types show an interleaving pattern. This same phenomenon was also manifested in our real data application (see later sections and Fig. S4).

Taken together, we showed that the novel integration of matrix factorization and spatial modeling in SPICEMIX yields superior inference of underlying cell identities across a variety of settings, compared to either NMF or HMRF alone. This improvement was seen for cell types with either sparse or layer-specific spatial patterns, both of which are prevalent in real data from complex tissues (e.g., the mouse cortex data used in this work). In addition, our evaluation also confirmed the effectiveness and robustness of our new optimization scheme for fitting the SPICEMIX model to spatial transcriptome data.

SPICEMIX refines cell identity inference from seqFISH+ data

We applied our method to the data acquired by seqFISH+ [12]. Specifically, we sought a robust model of the spatial variation of gene expression using SPICEMIX that would reveal both intrinsic factors of expression as well as spatial patterns, thereby unveiling cell identities more accurately. Here, we used the data of five separate samples of the mouse primary visual cortex, all from the same mouse but from contiguous layers, each from a distinct image or field-of-view (FOV), with single-cell expression of 2,470 genes in 523 cells [12]. We compared the cell identities revealed by SPICEMIX to those of NMF and Eng et al. [12].

The clustering of the learned latent representation of SPICEMIX revealed five excitatory neural subtypes, two inhibitory neural subtypes, and eight glial subtypes (Fig. 3a), supported by scRNA-seq marker genes [31] (Fig. 3b (left)). Although the assignment of major types was consistent between SPICEMIX, NMF, and [12] (Fig. 3b (middle) and Fig. S1), SPICEMIX refined and expanded the identification of cell subtypes (Fig. 3b (middle)). In particular, the identification of layer-specific excitatory neurons by SPICEMIX had a high correspondence with their associated layer (Fig. 3c), whereas several excitatory clusters from the original analysis in [12] were incorrectly dispersed across as many as three layers (see Fig. 3h in [12]). Furthermore, SPICEMIX correctly distinguished eL5b and eL6 neurons, which were mixed together in several clusters in [12] (Fig. 3b (middle)). The expression of marker genes Col6a1 and Ctgf [31] confirmed the identity of these cells (Fig. 3b (left)).

Beyond mere discrete cell type assignments, the metagenes and spatial affinities learned by SPICEMIX provided new insight into the underlying factors of glial cell states. The metagenes of SPICEMIX tend to capture either expression patterns of specific cell types, expressed at high levels, or patterns shared across

cell types, expressed at lower levels (Fig. 3b (right)). Notably, as annotated in Fig. 3b (right), metagene 7 is expressed at a high proportion among oligodendrocytes, distinguishing them from OPCs, while the expression of metagene 8, which is also present in OPCs, distinguished the rare Oligo-2 type from Oligo-1. This separation is confirmed by the expression patterns of the OPC marker gene Cspg4, the differentiating Oligo marker gene Tcf7l2 [32], and the mature Oligo marker gene Mog [33] (Fig. 3b (left)). Furthermore, the expression of the latter two marker genes supports the hypothesis that the Oligo-2 cells of SPICEMIX are likely in an intermediate transition during maturation from OPCs to oligodendrocytes, corresponding to the proportions of metagenes 7 and 8, rather than constituting a discrete cell type. Also, the learned metagene spatial affinities reveal that metagene 7 has a strong affinity for metagenes 3 and 4 (highlighted by black arrows in Fig. 3d (right), which are expressed primarily by the excitatory neurons of deeper tissue layers (eL5a, eL5b, and eL6) (Fig. 3b (right)). Thus, the spatial affinity of this oligodendrocyte-specific metagene 7 led to the separation of the Oligo-1 cells from OPCs, which, in contrast, do not have a strong affinity with any particular excitatory neuron type (Fig. 3d). In contrast, without spatial information to help decompose the highly similar expression profiles of these cell types, both NMF and Eng et al. [12] failed to distinguish these cells from other oligodendrocytes or OPCs (Fig. S1 and Fig. 3b (middle), respectively). Lastly, SPICEMIX revealed an additional separation of a cluster of [12] into SMC and Endo cells, which can be confirmed by the expression of their respective marker genes (i.e., Bgn highly expressed in SMC but not Endo cells, and Flt1 highly expressed in both SMC and Endo cells [31]) (Fig. 3b).

Together, by analyzing the seqFISH+ data with SPICEMIX, we identified cell subtypes of the mouse cortex whose spatial distributions are more consistent with prior experiments. We also delineated rarer subtypes that were not distinguished by other methods. This analysis strongly demonstrates the advantages and unique capabilities of SPICEMIX.

SPICEMIX reveals spatially-enriched cell types and subtypes from STARmap data

Next, we applied SPICEMIX to a single-cell spatial transcriptome dataset of the mouse cortex acquired by STARmap [13]. As in the analysis of the seqFISH+ dataset, the learned latent representation of cell identity of SPICEMIX provided a better characterization of cell subtypes and offered additional insight into their underlying factors. We analyzed a single sample consisting of 930 cells passing quality control, all from a single image or FOV, with expression measurements for 1020 genes. To distinguish cell-type labels between methods, we append an asterisk to the end of the cell labels of Wang et al. [13] when referenced.

We found that SPICEMIX produced more accurate cell labels than [13] and revealed subtypes missed both in [13] and by NMF (Fig. 4, Fig. S2). In comparison to NMF, SPICEMIX uncovered the following additional subtypes: SST inhibitory neuron, Oligo, Astro/Oligo, and two eL6 subtypes (Fig. 4a, b (left); supported by known marker genes [13, 31]). In comparison to the clusters from [13], SPICEMIX refined the assignment of excitatory neurons and further delineated the Oligo type into three subtypes: Oligo-1, Oligo-2, and Astro/Oligo (Fig. 4b (middle)). Specifically, SPICEMIX was able to learn the layer-like structure of excitatory neurons in tissue (Fig. 4c), thereby improving upon the assignments reported in Fig. 5d in [13], which erroneously mixed several neuron subtypes across layer boundaries. We noted that $\geq$15 eL2/3* or eL4* cells of [13] in fact resided not in layers L2-L4 but in layers L5 and L6 (black '$\times$' in the middle panel in Fig. 4c) and $\geq$15 eL5* neurons of [13] resided outside of layer L5 (black dots in the bottom panel in Fig. 4c), which is not consistent with the spatial association of those neurons. The refinement by SPICEMIX is especially notable in the reassignment of 36 cells in excitatory

subtypes eL2/3 and eL4, which yielded a much clearer delineation of their corresponding layers ('×' and '+' in Fig. 4c (middle panel)). We found that the smoothing assumption of HMRF produced overly-smooth boundaries between cell types and failed to identify the layer-wise structure of excitatory neurons (Fig. S4).

Additionally, SPICEMIX corrected the assignment of a large set of eL5 neurons erroneously labeled as astrocytes in [13] (Fig. 4b (middle), c). Astrocytes are known to be dispersed throughout the primary visual cortex, unlike the highly-localized Astro-1* cluster of [13] in the L5 layer (Fig. S5). SPICEMIX learned this spatial pattern without *a priori* knowledge, which it simultaneously used to correct these neurons. Differential expression analysis showed that many of the marker genes of astrocytes (identified by [31]), such as F3 and Sox9, are expressed in Astro-1* at a level significantly lower than those in Astro-2* and comparable to those in eL5 neurons (Fig. S5), confirming this assignment. Furthermore, the Astro-1* cells express marker genes of excitatory neurons, such as Slc17a7, Tcerg1l, Stac, and Parm1, at much higher levels than Astro-2* cells (Fig. S5).

The learned metagenes of SPICEMIX explain the improved cell identity inference and provide insights into the relation of different expression patterns across cell types. Metagenes for layer-specific cell types, such as metagene 3 for eL4 neurons, exhibited precise layer-specificity, enabling SPICEMIX to carefully delineate such subtypes (Fig. 4d; see also Fig. S3a). Sparsely expressed metagenes, such as metagene 8, which led to the identification of PVALB inhibitory neurons, were also successfully recovered by SPICEMIX. Such informative spatial patterns were attainable because of the accurate inference of the spatial affinity between metagenes (Fig. 4e). Excitatory subtypes eL6b and eL6c were also distinguished due to the learned spatial affinity between metagenes. The expression of metagenes 5 and 7 in contrary proportions distinguishes these two subtypes (Fig. 4b (right)), and the detection of these metagenes was aided by their strong affinity to each other (highlighted by a black arrow in Fig. 4e (right)). Our assignments represent a refined, spatially-informed separation for eL6 subtype. Also, metagene 11 shows a moderate affinity with many other metagenes, likely because metagene 11 is enriched in Astro cells. The strong self-affinity of metagenes 2-5, which are associated with excitatory neurons, yielded the clear layer boundaries seen in the labels of SPICEMIX and the correction of the Astro-1* to eL5 neurons mentioned above (Fig. 4c). In contrast, NMF metagenes typically exhibited diffuse and unspecific patterns of metagene expression (Fig. 4d and Fig. S3b). Another important novel observation from SPICEMIX is the expression of metagenes 12 and 13 among Oligo-1 and Astro/Oligo, which shed new light on the relation of these cells (see later section).

These results collectively suggest that SPICEMIX is able to refine cell identity and metagene inference with distinct spatial patterns from STARmap data, further demonstrating its unique advantage.

SPICEMIX identifies continuous myelination stages in oligodendrocytes

As in the results from the seqFISH+ data, the expression of metagenes learned by SPICEMIX from STARmap suggested the existence of continuous factors of cell identity that cannot be described merely by discrete clusters. One such factor, which SPICEMIX alone revealed, provided new insights into the myelination process of oligodendrocytes from the STARmap data. This process was observed among the oligodendrocytes of the Astro/Oligo and Oligo-1 types, which were discovered to lie along a continuum between the two classes in the latent space of SPICEMIX (Fig. 4a). We note that the Astro/Oligo type (colored magenta in Fig. 4) included a distinct group of astrocytes as well, which we removed from consideration. These astrocytes correspond to the thin sliver of magenta-colored cells at the superficial extremity of the tissue, whereas the remainder of Astro/Oligo cells, which we posit are oligodendrocytes,

are located in a deep-tissue layer near Oligo-1 cells (Fig. 4c). Additionally, the expression of astrocyte and oligodendrocyte marker genes among these astrocytes clearly resembles that of the STARmap Astro-2* class, whereas the expression of such genes among the oligodendrocytes of the Astro/Oligo class resembles that of the Oligo* class (Fig. S6a,c,d). We then examined the latent representation of the remaining oligodendrocytes and the cells of Oligo-1 and noted that cells from both clusters expressed metagenes 12 and 13 at high, but strikingly anti-correlated, levels (Fig. 4f).

We hypothesized that the differential expression of these metagenes was capturing the gradually activated myelination process of oligodendrocytes that is reflected by the gradually elevated or reduced expression levels of myelin sheath-related genes. To test this hypothesis, we used linear regression models to fit the relationship between the expression levels of myelin sheath-related genes and the differences in the proportions of metagenes 12 and 13 in individual cells. The eleven genes that we tested were those from the STARmap panel that were attributed to myelin sheath formation, according to the Gene Ontology (GO) database (see Supplementary Results B.1), and were expressed in at least 30% of the cells. We found that the correlations of seven of the eleven genes were significant ($p < 0.05$, after a two-step FDR correction for multiple testing) (Fig. 4f and Fig. S6b), supporting our hypothesis. Furthermore, we identified one gene, Atp1a2, which was confirmed recently by scRNA-seq [34, 35].

This result suggests that the latent representation of SpiceMix is uniquely able to elucidate relevant factors of cell identity, such as the myelination process among oligodendrocytes, further demonstrating the capability of SpiceMix to uncover cellular heterogeneity from spatial transcriptome data.

Discussion

In this work, we developed SpiceMix, an unsupervised method for modeling the diverse factors that collectively contribute to cell identity based on single-cell spatial transcriptome data. The novel integration of NMF and HMRF in SpiceMix combines the expressive power of NMF for modeling gene expression with the HMRF for modeling spatial relationships, advancing current state-of-the-art modeling for spatial transcriptomics. We evaluated the performance of SpiceMix on simulated data that approximates the mouse cortex spatial transcriptome, showing a clear advantage over NMF and HMRF. Applications of SpiceMix to single-cell spatial transcriptome data of the mouse primary visual cortex from seqFISH+ and STARmap demonstrated its effectiveness to produce reliable and informative latent representations of cell identity, unveiling more accurate cell type identification than prior approaches and uncovering important biological process underlying cell states.

As future work, SpiceMix could be further enhanced by incorporating additional modalities such as scRNA-seq data. Other recent computational methods have been used to study scRNA-seq and spatial transcriptomic data jointly [25, 26, 36], but they do not attempt to comprehensively model spatial cell-to-cell relationships. With proper normalization and preprocessing, data from scRNA-seq could be incorporated into the SpiceMix framework. This additional data may improve the inference of the latent variables and parameters of the model, which could further improve the modeling of cellular heterogeneity. In addition, further enhancements could be made to the probabilistic model of SpiceMix including additional priors, such as sparsity, to tailor toward particular application contexts.

As the area of spatial transcriptomics continues to thrive and data become more widely available, SpiceMix will be a uniquely useful tool for enabling new discoveries. In particular, the refined cell identity with SpiceMix has the potential to improve future studies of cell-cell interactions [37]. We considered STARmap and seqFISH+ methods in our study, because they provide gene counts for single

cells and they have some of the largest gene panels of such methods, however, there is still much potential to analyze data of other methods with SPICEMIX and also in a more integrated manner for cross-platform analysis. Nor is SPICEMIX limited to transcriptomic data, but its methodology may also be well-suited for other recent multi-omic spatial data, e.g., DBiT-seq [38]. Overall, SPICEMIX is a powerful framework that can serve as an essential tool for the analysis of diverse types of spatial transcriptome and multi-omic data, with the distinct advantage that it can unravel the complex mixing of latent intrinsic and spatial factors of heterogeneous cell identity in complex tissues.

Methods

Probabilistic formulation of NMF-HMRF in SPICEMIX

Gene expression as matrix factorization

We consider the log-transformed normalized expression of individual cells $Y = [y_1, \ldots, y_N] \in \mathbb{R}_+^{G \times N}$, where constants G and N denote the number of genes and cells, respectively, to be the product of K underlying factors (i.e., *metagenes*), $M = [m_1, \ldots, m_K]$, $m_k \in \mathbb{S}_{G-1}$, and weights, $X = [x_1, \ldots, x_N] \in \mathbb{R}_+^{K \times N}$, i.e.,

$$Y = MX + E. \tag{1}$$

This follows the non-negative matrix factorization (NMF) formulation of gene expression of prior work [39]. The term $E = [e_1, \ldots, e_N] \in \mathbb{R}^{G \times N}$ captures the unexplained variation or noise, which we model as i.i.d. Gaussian, i.e., $e_i \sim \boldsymbol{\mathcal{N}}(0, \sigma_y^2 I)$. To resolve the scaling ambiguity between M and X, we constrain the columns of M to sum to one, so as to lie in the $(G-1)$-dimensional simplex, $\mathbb{S}_{G-1}$. For notation consistency, capital letters are used to denote matrices and lowercase letters denote their column vectors.

Graphical model formulation

Our formulation for the NMF-HMRF in SPICEMIX enhances standard NMF by modeling the spatial correlations among samples (i.e., cells in this context) via the HMRF [29]. This novel integration aids inference of the latent M and X by enforcing spatial consistency. The spatial relationship between cells in tissue is represented as a graph $\boldsymbol{\mathcal{G}} = (\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}})$ of nodes $\boldsymbol{\mathcal{V}}$ and edges $\boldsymbol{\mathcal{E}}$, where each cell is a node and edges are determined from the spatial locations. Any graph construction method for determining edges, such as distance thresholding or Delaunay triangulation, can be used. For each node i in the graph, the measured gene expression vector, y_i, is the set of observed variables and the weights, x_i, describing the mixture of metagenes, are the hidden variables, or state. The observations are related to the hidden variables via the potential function ϕ, which captures the NMF formulation. The spatial affinity between the metagene proportions of neighboring cells is captured by the potential function φ. Together, these elements constitute the HMRF.

Given an observed dataset, the model can be fit by maximizing the likelihood of the data. By the Hammersley-Clifford theorem [40], the likelihood of the data for the pairwise HMRF can be formulated as the product of pairwise dependencies between nodes,

$$P(Y, X|\Theta) = \frac{1}{Z(\Theta)} \prod_{(i,j)\in\boldsymbol{\mathcal{E}}} \varphi(x_i, x_j) \prod_{i\in\boldsymbol{\mathcal{V}}} \phi(y_i, x_i)\pi(x_i), \tag{2}$$

where $\Theta = \{\Delta, M\}$ is the set of model parameters and metagenes and $Z(\Theta)$ is the normalizing partition function that ensures P is a proper probability distribution. The potential function π is added to capture a prior on the hidden states. We assume that the potential functions have an exponential form, allowing them to be written in the form of the exponential of an energy function, i.e.,

$$\phi(y_i, x_i) = \exp\left(-U_y(y_i, x_i)\right), \quad \varphi\left(x_i, x_j\right) = \exp\left(-U_x(x_i, x_j)\right). \tag{3}$$

The energy functions are given by:

$$U_y(y_i, x_i) = \frac{(y_i - Mx_i)^2}{2\sigma_y^2}, \quad U_x(x_i, x_j) = \frac{x_i^\top}{\|x_i\|_1} \Sigma_x^{-1} \frac{x_j}{\|x_j\|_1}. \tag{4}$$

The energy function U_y is the reconstruction error of the measured expression of cell i according to the estimated x_i and M, and σ_y^2 represents the variation of expression, or noise, of the NMF model. The energy function U_x measures the inner-product between the metagene proportions of neighboring cells i and j, weighted by a learned, pairwise correlation matrix Σ_x^{-1}, which captures the spatial affinity of metagenes. By normalizing the weights x_i of each cell, any scaling effects, such as cell size, are removed, such that the similarity is purely a measure of proportions. Additionally, we formulate the potential function of the hidden states X to also have an exponential form, with a scale parameter of

$$\lambda_x = 1, \quad \pi(x_i) = \exp\left(-\lambda_x \|x_i\|_1\right). \tag{5}$$

We normalize the average of the total normalized expression levels in individual cells to K correspondingly.

Parameter priors

We assume a Gaussian prior with zero mean and σ_Σ^2 variance on the elements of the pairwise matrix Σ_x^{-1}, while enforcing that the matrix be transpose symmetric, i.e.,

$$P\left(\Sigma_x^{-1}\right) = \left(\sqrt{\pi/\lambda_\Sigma}\right)^{-K^2/2} \exp\left(-\lambda_\Sigma \left\|\Sigma_x^{-1}\right\|_F^2\right), \tag{6}$$

where $\lambda_\Sigma = 1/(2\sigma_\Sigma^2)$ and F denotes the Frobenius norm. This prior can be viewed as a regularization that allows us to control the importance of the spatial relationships during inference.

Alternating estimation of hidden states and model parameters

To infer the hidden states and model parameters of the NMF-HMRF model in SpiceMix, we optimize the data likelihood via coordinate ascent, alternating between optimizing hidden states and model parameters. This new optimization scheme is summarized in Algorithm 1. First, to make inference tractable, we approximate the joint probability of the hidden states by the pseudo-likelihood [40], which is the product of conditional probabilities of the hidden state of individual nodes given that of their neighbors,

$$P(X|\Theta) \approx \prod_{i \in \boldsymbol{\mathcal{V}}} P(x_i | x_{\eta(i)}, \Theta), \tag{7}$$

where $\eta(i)$ is the set of neighbors to node i.

Estimation of hidden states

Given parameters Θ of the model, we estimate the factorizations X by maximizing their posterior distribution. The maximum a posteriori (MAP) estimate of X is given by:

$$\hat{X} = \operatorname*{argmax}_{X \in \mathbb{R}_+^{K \times N}} P(X|Y, \Theta) = \operatorname*{argmax}_{X \in \mathbb{R}_+^{K \times N}} P(Y, X|\Theta) = \operatorname*{argmax}_{X \in \mathbb{R}_+^{K \times N}} \{\log P(Y, X|\Theta)\} \tag{8}$$

$$= \operatorname*{argmax}_{X \in \mathbb{R}_+^{K \times N}} \left\{ \sum_{i \in \boldsymbol{\mathcal{V}}} \left[-U_y(y_i, x_i) + \log \pi(x_i)\right] - \sum_{(i,j) \in \boldsymbol{\mathcal{E}}} U_x(x_i, x_j) \right\}. \tag{9}$$

This is a quadratic program and can be solved efficiently via the iterated conditional model (ICM) [41] using the software package Gurobi [42] (see Supplementary Methods A.1 for more details).

Algorithm 1 NMF-HMRF model-fitting and hidden state estimation.

Derive an initial estimate $M^{(0)}$ using K-means clustering assuming no spatial relationships.
for $1 < t \leq T_0$ **do**
  Derive an estimate $X^{(t)}$ by minimizing $R(X) = ||Y - M^{(t-1)}X||_2^2$.
  Derive an estimate $M^{(t)}$ by minimizing $R(M) = ||Y - MX^{(t)}||_2^2$.
end for
Set $M^{(0)} = M^{(T_0)}$, $X^{(0)} = X^{(T_0)}$.
Derive an initial estimate $\sigma_y^{(0)} = \sqrt{R(X^{(0)})/(G \times N)}$.
Initialize $(\Sigma_x^{-1})^{(0)}$ to a zero matrix.
for $1 < t \leq T$ **do**
  Derive an estimate $X^{(t)}$ given $\Theta^{(t-1)}$ by maximizing $P(X|Y, \Theta = \Theta^{(t-1)})$.
  Derive an estimate $\Theta^{(t)}$ given $X^{(t)}$ by maximizing $P(\Theta|Y, X = X^{(t)})$.
end for

Estimation of model parameters

Given an estimate of the hidden states X, we can likewise solve for the unknown model parameters Θ by maximizing their posterior distribution. The MAP estimate of the parameters Θ is given by:

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\, P(\Theta|Y, X) = \underset{\Theta}{\operatorname{argmax}}\, P(Y, X|\Theta)P(\Theta) = \underset{\Theta}{\operatorname{argmax}} \{\log P(Y, X|\Theta) + \log P(\Theta)\} \quad (10)$$

$$= \underset{\Theta}{\operatorname{argmax}} \left\{ \sum_{i \in \boldsymbol{\mathcal{V}}} [-U_y(y_i, x_i) + \log \pi(x_i)] - \sum_{(i,j) \in \boldsymbol{\mathcal{E}}} U_x(x_i, x_j) - \log Z(\Theta) + \log P(\Theta) \right\} \quad (11)$$

$$\approx \underset{\Theta}{\operatorname{argmax}} \left\{ \sum_{i \in \boldsymbol{\mathcal{V}}} [-U_y(y_i, x_i) + \log \pi(x_i) - \log Z_i(\Theta)] - \sum_{(i,j) \in \boldsymbol{\mathcal{E}}} U_x(x_i, x_j) + \log P(\Theta) \right\}. \quad (12)$$

Eqn. 12 is an approximation by the mean-field assumption [40], which is used, in addition to the pseudo-likelihood assumption, to make the inference of model parameters tractable. We note that we can estimate metagenes, spatial affinity, and the noise level independently. The MAP estimate of the metagenes M is a quadratic program, which is efficient to solve. The MAP estimate of Σ_x^{-1} is convex and is solved by the optimizer Adam [43]. Due to the complexity of the partition function $Z_i(\Theta)$ of the likelihood, which includes integration over X, it is approximated by Taylor's expansion. Since it is a function of Θ, this computation must be performed at each optimization iteration. See Supplementary Methods A.2 for details of the optimization method.

Initialization

To produce initialize estimates of the model parameters and hidden states, we do the following. First, we use a common strategy for initializing NMF, which is to cluster the data using K-means clustering, with K equal to the number of metagenes, and use the means of the clusters as an estimate of the metagenes. We then alternate for T_0 iterations between solving the NMF objective for X and M. This produces, in only a few quick iterations, an appropriate initial estimate for the algorithm, which will be subsequently refined. We observed that if T_0 is too large, it can cause the algorithm to prematurely reach a local minimum before spatial relationships are considered. However, this value can be easily tuned by experimentation, and in our analysis, we found that just 5 iterations were necessary.

Empirical running time

SPICEMIX takes 0.5-2 hours to run on a spatial transcriptome dataset with 2,000 genes and 1,000 cells on a machine with eight 3.6 GHz CPUs and one GeForce 1080 Ti GPU. The GPU is used for the first 5 iterations, or around that number, only, when the spatial affinity matrix Σ_x^{-1} is changed significantly. Later on, most time is spent solving quadratic programmings. Since the algorithm uses a few iterations of NMF to provide an initial estimate, which is a reasonable starting point, it is expected to find a good initial estimate of metagenes and latent states efficiently.

Generation of simulated data

We generated simulated spatial transcriptomic data following expression and spatial patterns similar to cells in the mouse primary visual cortex. Cells in the mouse cortex are classified into three primary categories: inhibitory neurons, excitatory neurons, and non-neurons or glial cells [31, 44]. Excitatory neurons in the cortex exhibit concentrated, layer-wise specificity, whereas inhibitory neurons spread sparsely throughout. Non-neuronal cells can exhibit either layer-specific or diffuse patterns. We simulated data as acquired by imaging a slice of the tissue, consisting of four distinct vertical layers and eight cell types: four excitatory, two inhibitory, and two glial (Fig. 2a). Each layer was densely populated by one layer-specific excitatory neuron type. The two inhibitory neuron types were scattered sparsely throughout several layers. One non-neuronal type was restricted to the first layer and the other was scattered sparsely throughout several layers. The identity of a cell belonging to a specific type was defined by a specific mixture of metagenes associated with major-type, subtype, or layers, along with three noise metagenes (see Fig. 2b for the average proportions for each cell type). This pattern followed the observed trends of real mouse cortex data. For excitatory neurons, the layer-specific metagene defined the subtype. The two glial types had different major-type metagenes and no subtype metagenes. Given the class-specific metagene proportions, which we denote by the K-dimensional vector b_c for cell type c, the proportions for an individual cell are given by

$$v_i = \frac{\tilde{v}_i}{\sum_k \tilde{v}_{i,k}} \qquad\qquad \tilde{v}_i = b_c + \eta_i,$$

where $\eta_i \sim \mathcal{N}(0, \sigma_x \Sigma_c)$ is a K-dimensional Gaussian random variable that controls the cell-to-cell variation of metagene proportion. The diagonal matrix Σ_c defines the variance for each metagene for cell type c. The parameter σ_x is a scaling constant that controls the overall variance of metagene proportions for a given simulation. To simulate cell-specific variation of the number of total gene counts, we then scaled v_i by a random scaling factor s_i drawn from the Gamma distribution, with a shape parameter of K and a scale parameter of 1. This yields the final hidden state

$$x_i = s_i v_i$$

for each cell i. For each simulated image, or tissue sample, 500 cells were created with randomly generated locations, in such a way so as to maintain a minimum distance between any two cells, so that the density of cells across the sample was roughly constant.

The procedure to generate metagenes is as follows. We randomly generated the value $M_{g,k}$ for the expression of gene g for metagene k, but linked a significant percentage of the genes of subtype metagenes, such that the expression of those genes was the same across linked metagenes. For the two inhibitory neuron subtype-specific metagenes, 75% of the genes were linked. The four layer-specific

metagenes had 90% of their genes linked together. We also linked 75% of the genes of the excitatory and inhibitory major-type metagenes, since they are both neural types. This pattern follows what was observed of metagenes learned from real spatial transcriptomic data in this work. We generated the value for each gene for each metagene from the Gamma distribution with a scale parameter of 1. The shape for noise metagenes was 8, and the parameter for all other metagenes was 4. This achieved a desired level of sparsity among gene expression values. After the initial values for each gene were drawn, we normalized each metagene to sum to one. In this way, any scaling of the total number of counts per cell is captured entirely by the hidden variables x_i.

Given the randomized metagenes and hidden states, the observed expression $Y_{g,i}$ for cell i for gene g is a linear combination of the metagenes, with weights x_i, and with added Gaussian noise, $y_i = Mx_i + e_i$, where $e_i \sim \mathcal{N}(0, \sigma_y^2 I_G)$ and I_G is the identity matrix of dimension G. In our experiments, we set $G = 100$ genes.

To test the robustness of SPICEMIX, we designed four different simulation scenarios following the above description, in which the values of σ_y and σ_x were varied (see Fig. 2d). For each scenario, we generated 20 replicates and reported the 0th, 25th, 50th, 75th, and 100th percentiles (excluding outliers) of the adjusted Rand index (ARI) across replicates. For details on hyperparamter selection for the results on this simulated data, see Supplementary Methods A.3.

Data processing for seqFISH+ and STARmap data

We applied SPICEMIX on seqFISH+ and STARmap datasets that profiled the mouse primary visual cortex where different cell types exhibit distinct spatial distributions. We applied appropriate preprocessing to remove technical biases in the input profiles. See Supplementary Methods A.4 for details. Steps of data processing include: constructing the neighbor graph of cells, selection of hyperparameters for SPICEMIX, NMF, and HMRF, random seed selection, the choice of the number of metagenes, and the choice of the number of clusters for hierarchical clustering. See Supplementary Methods A.5 for details of these steps, as well as Supplementary Methods A.6 and A.7 for specific details of the number of clusters and metagenes for analysis of seqFISH+ and STARmap, respectively. See Fig. S7 and Fig. S3c for the CH index values leading to the choice of the number of clusters for seqFISH+ and STARmap, respectively. For the explanation of our method for constructing the cell-type affinity matrix, see Supplementary Methods A.8.

Code Availability

The source code of SpiceMix can be accessed at: https://github.com/ma-compbio/SpiceMix.

Acknowledgements

This work was supported in part by the National Institutes of Health Common Fund 4D Nucleome Program grant UM1HG011593 (J.M.), National Institutes of Health grant R01HG007352 (J.M.), and National Science Foundation grant 1717205 (J.M.). J.M. is additionally supported by a Guggenheim Fellowship from the John Simon Guggenheim Memorial Foundation.

Author Contributions

Conceptualization, B.C. and J.M.; Methodology, B.C., T.Z., and J.M.; Software, T.Z. and B.C.; Investigation, B.C., T.Z., and J.M.; Writing – Original Draft, B.C., T.Z., and J.M.; Writing – Review & Editing, B.C., T.Z., and J.M.; Funding Acquisition, J.M.

Competing Interests

The authors declare no competing interests.

References

[1] Arendt D, Musser JM, Baker CV, Bergman A, Cepko C, Erwin DH, et al. The origin and evolution of cell types. Nature Reviews Genetics. 2016;17(12):744.

[2] Chen X, Teichmann SA, Meyer KB. From Tissues to Cell Types and Back: Single-Cell Gene Expression Analysis of Tissue Architecture. Annual Review of Biomedical Data Science. 2018;1:29–51.

[3] Consortium H, et al. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. Nature. 2019;574(7777):187.

[4] Trapnell C. Defining cell types and states with single-cell genomics. Genome Research. 2015;25(10):1491–1498.

[5] Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-cell genomics. Nature Biotechnology. 2016;34(11):1145.

[6] Stuart T, Satija R. Integrative single-cell analysis. Nature Reviews Genetics. 2019;20(5):257–272.

[7] Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, et al. Highly multiplexed subcellular RNA sequencing in situ. Science. 2014;343(6177):1360–1363.

[8] Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. Science. 2015;348(6233):aaa6090.

[9] Shah S, Lubeck E, Zhou W, Cai L. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. Neuron. 2016;92(2):342–357.

[10] Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016;353(6294):78–82.

[11] Moffitt JR, Bambah-Mukku D, Eichhorn SW, Vaughn E, Shekhar K, Perez JD, et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. Science. 2018;362(6416):eaau5324.

[12] Eng CHL, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. Nature. 2019;568:235–239.

[13] Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science. 2018;341(6400):eaat5691.

[14] Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, et al. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science. 2019;363(6434):1463–1467.

[15] Vickovic S, Eraslan G, Salmén F, Klughammer J, Stenbeck L, Schapiro D, et al. High-definition spatial transcriptomics for in situ tissue profiling. Nature Methods. 2019;16(10):987–990.

[16] Zhuang X. Spatially resolved single-cell genomics and transcriptomics by imaging. Nature Methods. 2021;18(1):18–22.

[17] Larsson L, Frisén J, Lundeberg J. Spatially resolved transcriptomics adds a new dimension to genomics. Nature Methods. 2021;18(1):15–18.

[18] Lein E, Borm LE, Linnarsson S. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. Science. 2017;358(6359):64–69.

[19] Zhu Q, Shah S, Dries R, Cai L, Yuan GC. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. Nature Biotechnology. 2018;36(12):1183.

[20] Schapiro D, Jackson HW, Raghuraman S, Fischer JR, Zanotelli VR, Schulz D, et al. histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. Nature Methods. 2017;14(9):873.

[21] Svensson V, Teichmann SA, Stegle O. SpatialDE: identification of spatially variable genes. Nature Methods. 2018;15(5):343.

[22] Arnol D, Schapiro D, Bodenmiller B, Saez-Rodriguez J, Stegle O. Modeling Cell-Cell Interactions from Spatial Molecular Data with Spatial Variance Component Analysis. Cell Reports. 2019;29(1):202–211.

[23] Nitzan M, Karaiskos N, Friedman N, Rajewsky N. Gene expression cartography. Nature. 2019;576(7785):132–137.

[24] Sun S, Zhu J, Zhou X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nature Methods. 2020;17(2):193–200.

[25] Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck III WM, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177(7):1888–1902.e21.

[26] Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-Cell Multiomic Integration Compares and Contrasts Features of Brain Cell Identity. Cell. 2019;177(7):1873–1887.e17.

[27] Elosua M, Nieto P, Mereu E, Gut I, Heyn H. SPOTlight: Seeded NMF regression to Deconvolute Spatial Transcriptomics Spots with Single-Cell Transcriptomes. bioRxiv. 2020.

[28] Lee DD, Seung HS. Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems; 2001. p. 556–562.

[29] Zhang Y, Brady M, Smith S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Transactions on Medical Imaging. 2001;20(1):45–57.

[30] Caliński T, Harabasz J. A dendrite method for cluster analysis. Commun Stat Simul Comput. 1974 Jan;3(1):1–27.

[31] Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nature Neuroscience. 2016;19(2):335.

[32] Zhao C, Deng Y, Liu L, Yu K, Zhang L, Wang H, et al. Dual regulatory switch through interactions of Tcf7l2/Tcf4 with stage-specific partners propels oligodendroglial maturation. Nature Communications. 2016;7(1):1–15.

[33] Linington C, Bradl M, Lassmann H, Brunner C, Vass K. Augmentation of demyelination in rat acute allergic encephalomyelitis by circulating mouse monoclonal antibodies directed against a myelin/oligodendrocyte glycoprotein. The American Journal of Pathology. 1988;130(3):443.

[34] Marques S, van Bruggen D, Vanichkina DP, Floriddia EM, Munguba H, Väremo L, et al. Transcriptional convergence of oligodendrocyte lineage progenitors during development. Developmental Cell. 2018;46(4):504–517.

[35] Beiter RM, Fernández-Castañeda A, Rivet-Noor C, Merchak A, Bai R, Slogar E, et al. Evidence for oligodendrocyte progenitor cell heterogeneity in the adult mouse brain. bioRxiv. 2020.

[36] Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nature Methods. 2019:1–8.

[37] Armingol E, Officer A, Harismendy O, Lewis NE. Deciphering cell–cell interactions and communication from gene expression. Nature Reviews Genetics. 2020:1–18.

[38] Liu Y, Yang M, Deng Y, Su G, Enninful A, Guo CC, et al. High-Spatial-Resolution Multi-Omic

Sequencing via Deterministic Barcoding in Tissue. Cell. 2020.
[39] Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the National Academy of Sciences. 2004;101(12):4164–4169.
[40] Murphy K. Machine learning: a probabilistic perspective. MIT Press; 2012.
[41] Besag J. On the statistical analysis of dirty pictures. Journal of the Royal Statistical Society: Series B (Methodological). 1986;48(3):259–279.
[42] Gurobi Optimization L. Gurobi Optimizer Reference Manual; 2020. Available from: http://www.gurobi.com.
[43] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.
[44] Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature. 2007;445(7124):168–176.

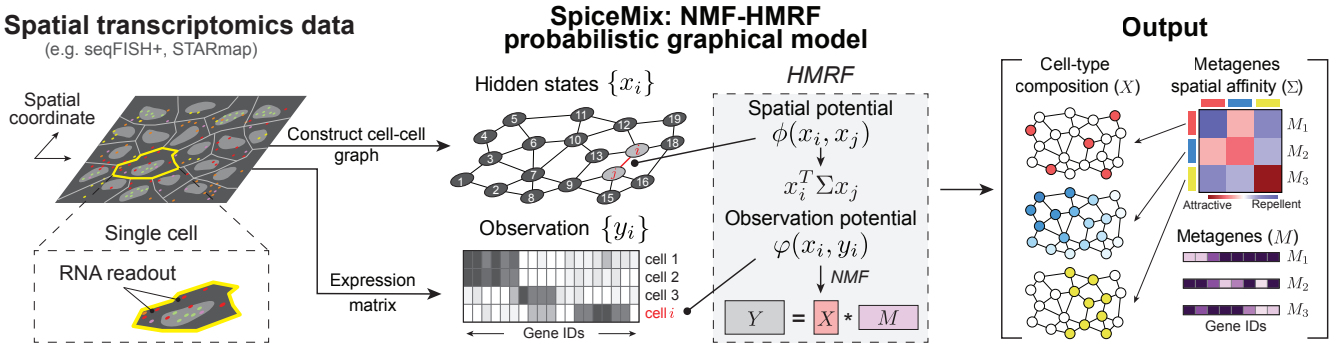

Figure 1: Overview of SPICEMIX. Gene expression measurements and a neighbor graph are extracted from *in situ* single-cell spatial trancriptome data and fed into the SPICEMIX framework. SPICEMIX decomposes the expression y_i in cell i into a mixture of metagenes weighted by the hidden state x_i. Spatial interaction between neighboring cells i and j is modeled by an inner product of their hidden states, weighted by inferred spatial affinities between metagenes Σ. Collectively, the mixture weights for individual cells X, the metagene spatial affinity Σ, and K metagenes M, all inferred by SPICEMIX, provide unique insight into the latent intrinsic and spatial factors of cell identity.

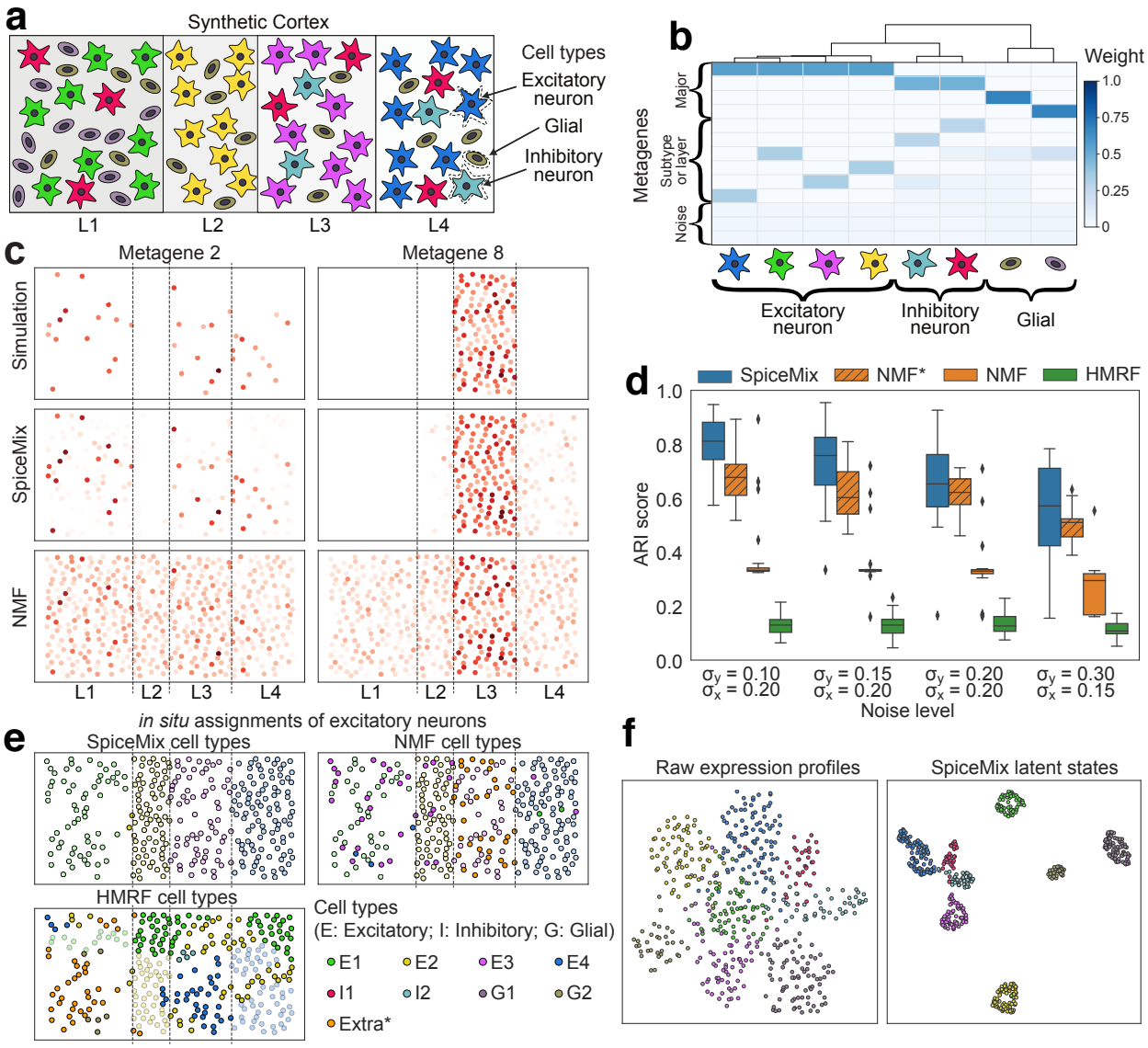

Figure 2: Overview of the simulated spatial transcriptome data of the mouse cortex, and performance comparison between SpiceMix, NMF, and HMRF. **a.** Illustration of three major cell types distributed in four layers. In this depiction, excitatory and inhibitory neurons are star-shaped and glial cells are ovals. Subtypes are distinguished by their colors. **b.** Dendrogram showing the similarity of the expression profiles of the eight subtypes (top), their metagene profiles (middle), and their colors and shapes used in panel **a**. The top four rows correspond to metagenes that determine major type, the next six rows correspond to metagenes that determine subtypes or are layer-specific, and the bottom three rows correspond to noise metagenes. **c.** Simulated expression of metagenes 2 and 8, from a single sample, in their spatial context (top) and the estimated expression of those metagenes by SpiceMix (middle) and NMF (bottom). Visualizations in **e** and **f** are of the same sample. **d.** Box plots of the adjusted Rand index (ARI) that measures the quality of the matching between the identified cell types for each method and the true simulated cell types. The optimal number of cell types for NMF was determined by the Calinski-Harabasz index ('NMF' in the legend) or by maximizing the ARI score ('NMF*' in the legend). Results are reported across four simulation scenarios with varying noise levels. **e.** Assignments of excitatory neurons for each method in their spatial context. Colors were assigned to cells by the closest matching simulated type. Cells assigned to the incorrect cell types have bright colors. Cells assigned to the correct cell types have faint colors. Cells in orange belong to a cell type that does not match any simulated cell type. **f.** UMAP plots of raw expression values of cells (left) and the learned latent states of SpiceMix (right). Colors match those of the spatial maps.

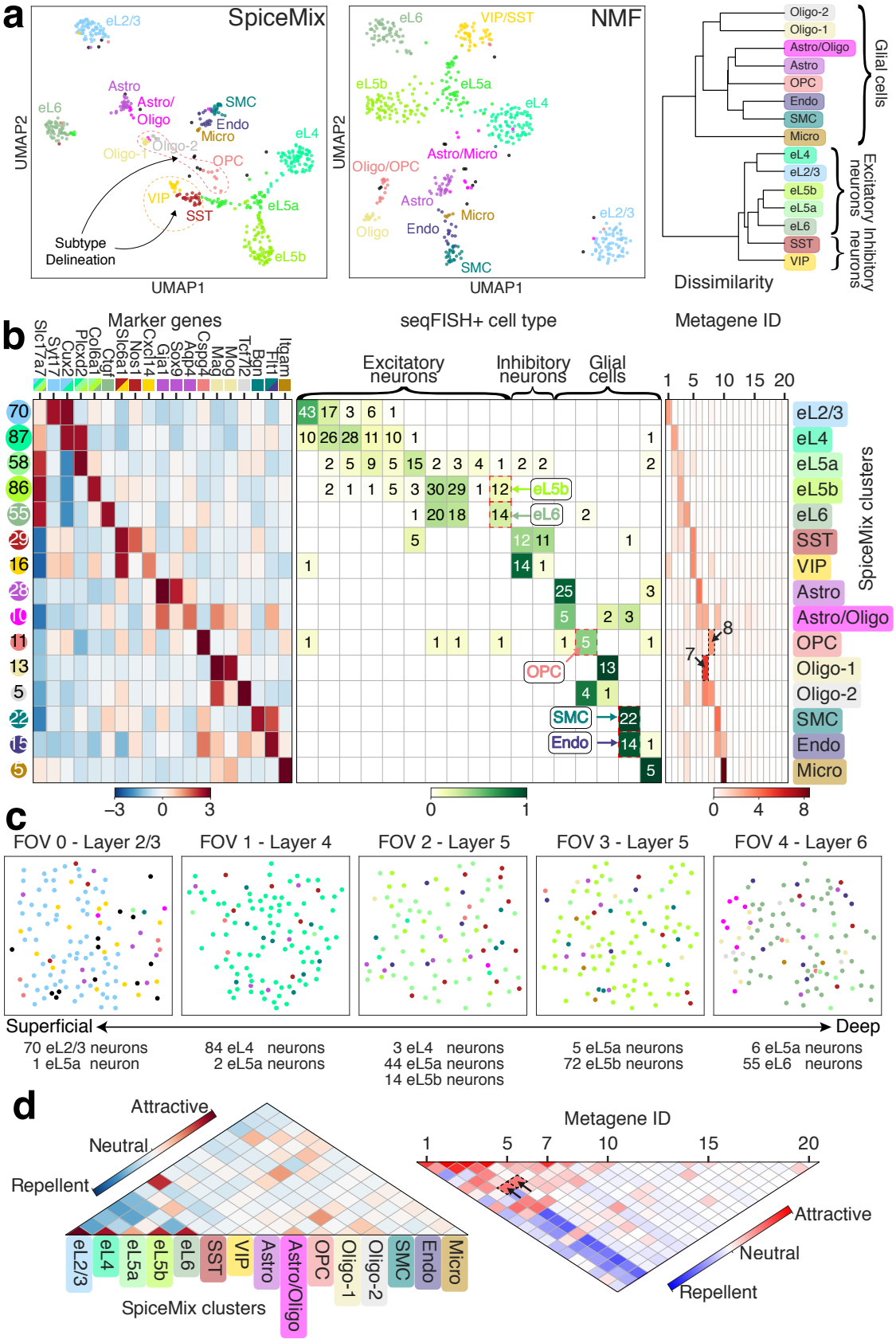
a
SpiceMix
NMF
UMAP1
UMAP2
Subtype Delineation
Dissimilarity
Glial cells
Excitatory neurons
Inhibitory neurons
b
Marker genes
seqFISH+ cell type
Metagene ID
Excitatory neurons
Inhibitory neurons
Glial cells
SpiceMix clusters
c
FOV 0 - Layer 2/3
FOV 1 - Layer 4
FOV 2 - Layer 5
FOV 3 - Layer 5
FOV 4 - Layer 6
Superficial
Deep
70 eL2/3 neurons
1 eL5a neuron
84 eL4 neurons
2 eL5a neurons
3 eL4 neurons
44 eL5a neurons
14 eL5b neurons
5 eL5a neurons
72 eL5b neurons
6 eL5a neurons
55 eL6 neurons
d
Attractive
Neutral
Repellent
SpiceMix clusters
Metagene ID

Figure 3 ***(preceding page)*****:** Application of SpiceMix to the seqFISH+ data from the mouse primary visual cortex [12]. Note that colors throughout the figure of cells and labels correspond to the cell-type assignments of SpiceMix. **a.** UMAP plots of the latent states of SpiceMix (left) and NMF (middle), as well as the dendrogram of the arithmetic average of the expression for each cell type of SpiceMix (right). It is highlighted in **a** (left) that SpiceMix further delineated inhibitory neurons into VIPs (yellow) and SSTs (brown) enclosed by the orange dashed cycle, and delineated Oligos and OPCs into separate subtypes: Astro/Oligo (magenta), Oligo-1 (light yellow), Oligo-2 (silver), and OPC (red), enclosed with the red dashed cycle. The colors in the NMF UMAP plot match NMF cell types to the most similar SpiceMix cell type. **b.** (Left) Average expression of known marker genes within SpiceMix cell types, along with the number of cells belonging to each type (colored circles). The colored boxes following the name of each marker gene correspond to their known associated cell type. (Middle) Agreement of SpiceMix cell-type assignments with those of the original analysis in [12]. (Right) Average expression of inferred metagenes within SpiceMix cell types. **c.** SpiceMix cell-type assignments for all cells in each of the 5 FOVs. Samples from superficial layers are on the left and samples from deep tissue layers are on the right. **d.** (Left) The inferred pairwise spatial affinity of cell types. (Right) The inferred pairwise spatial affinity of metagenes, or Σ_x^{-1}.

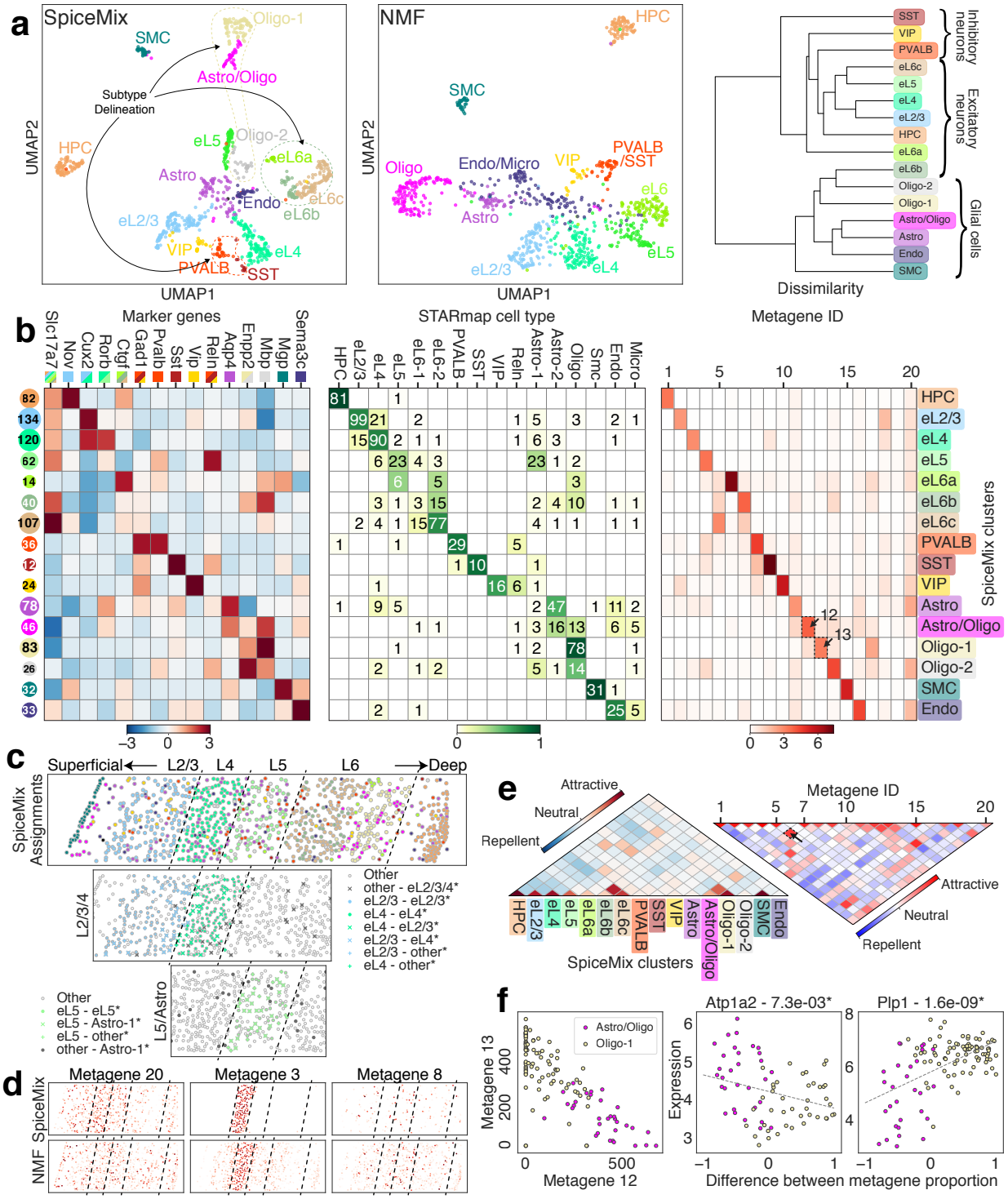
a
SpiceMix
NMF
UMAP1
UMAP2
Subtype Delineation
Dissimilarity
Inhibitory neurons
Excitatory neurons
Glial cells
b
Marker genes
STARmap cell type
Metagene ID
SpiceMix clusters
c
Superficial
L2/3
L4
L5
L6
Deep
SpiceMix Assignments
L2/3/4
L5/Astro
d
Metagene 20
Metagene 3
Metagene 8
SpiceMix
NMF
e
Attractive
Neutral
Repellent
Metagene ID
SpiceMix clusters
f
Astro/Oligo
Oligo-1
Metagene 13
Metagene 12
Atp1a2 - 7.3e-03*
Plp1 - 1.6e-09*
Expression
Difference between metagene proportion

Figure 4 *(preceding page)*: Application of SpiceMix to the STARmap data from mouse primary visual cortex [13]. Colors throughout the figure of cells and labels correspond to the cell-type assignments of SpiceMix. **a.** UMAP plots of the latent states of SpiceMix (left) and NMF (middle), as well as the dendrogram of the arithmetic average of the expression for each cell type of SpiceMix (right). It is highlighted in **a** (left) that SpiceMix separated PVALBs (orange) and SSTs (brown) enclosed by the orange dashed cycle, delineated eL6 neurons into three subtypes enclosed in the dark green cycle, and delineated Oligos and OPCs into three separate subtypes: Oligo-1 (light yellow), Oligo-2 (silver), and Astro/Oligo (magenta), enclosed with the red dashed cycle. The colors in the NMF UMAP plot match NMF cell types to the most similar SpiceMix cell type. **b.** (Left) Average expression of known marker genes within SpiceMix cell types, along with the number of cells belonging to each type (colored circles). The colored boxes following the name of each marker gene correspond to their known associated cell types. (Middle) Agreement of SpiceMix cell-type assignments with those of the original analysis in [13]. (Right) Average expression of inferred metagenes within SpiceMix cell types. **c.** SpiceMix cell-type assignments for all cells in the sample are shown at the top. Lower panels show comparisons of SpiceMix clusters and those of the original analysis [13] for specific cell types. The labels in the legend are the SpiceMix cell type, followed by a dash, followed by the cell type of [13], denoted by an asterisk. The regions in the lower panels are cropped from the top panel and share x-coordinates with the top panel. The layers are separated by black dashed lines and are annotated above the top panel. **d.** Example spatial maps of expression of metagenes from SpiceMix (top) and NMF (bottom). These metagenes are labeled by their indices in SpiceMix, and the lower panels show the counterparts in NMF. **e.** (Left) The inferred pairwise spatial affinity of cell types. (Right) The inferred pairwise spatial affinity of metagenes, or Σ_x^{-1}. **f.** (Left) The expression of metagene 13 plotted against the expression of metagene 12 for oligodendrocytes of the SpiceMix Oligo-1 and Astro/Oligo types. (Right) The expression of important marker genes for myelin-sheath formation in oligodendrocytes plotted against the relative expression of metagenes 12 and 13 of the same cells. The title of each plot consists of the gene symbol and the corrected *p*-value of having a nonzero slope, respectively. An asterisk after the *p*-value means that the result is significant under the threshold of 0.05 (see Supplementary Methods B.1 for details).