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 probabilistic model that enables effective joint analysis of spatial information and gene expression of single cells based on spatial transcriptome data. Both simulation and real data evaluations demonstrate that SPICEMIX consistently improves upon the inference of the intrinsic cell types compared with existing approaches. As a proof-of-principle, we use SPICEMIX to analyze single-cell spatial transcriptome data of the mouse primary visual cortex acquired by seqFISH+ and STARmap. We find that SPICEMIX can improve cell identity assignments and uncover potentially new cell subtypes. SPICEMIX is a generalizable framework for analyzing spatial transcriptome data that may 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 comprise the identity of a cell (Arendt et al., 2016; Chen et al., 2018; Consortium et al., 2019). Single-cell RNA-seq (scRNA-seq) has greatly advanced our understanding of complex cell types in different tissues (Trapnell, 2015; Wagner et al., 2016; Stuart and Satija, 2019), 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 (Lee et al., 2014; Chen et al., 2015; Shah et al., 2016; Ståhl et al., 2016; Moffitt et al., 2018; Wang et al., 2018; Eng et al., 2019; Rodriques et al., 2019; Vickovic et al., 2019) have revealed spatial information of gene expression of dozens to tens of thousands of genes in individual cells *in situ* within the tissue context.

Although spatially resolved single-cell transcriptome data are being made available by these exciting emerging technologies, the development of computational methods that capture the unique properties of such data to unveil single-cell identities remains a challenge (Lein et al., 2017). Methods have been proposed to align spatial transcriptomic data with scRNA-seq data to learn latent variable models of single-cell transcription (Stuart et al., 2019; Welch et al., 2019; Elosua et al., 2020), but these methods do not explicitly model cell-to-cell relationships. Graphical models are a natural paradigm for modeling spatial data, and in this vein, Zhu et al. (2018) recently proposed the use of a hidden Markov random field (HMRF) to learn spatial subtypes of cells. A crucial limitation of this method, however, is its reliance upon prior knowledge of marker genes from scRNA-seq data to impute labels of known underlying cell types, rather than learning intrinsic and spatial types simultaneously from spatial transcriptome data directly. Additionally, the standard HMRF assumption of discrete hidden states makes it unsuited for learning the underlying mixture of diverse spatial and intrinsic factors of a cell. Several other notable methods have been developed to study the relationship of known cell types in local neighborhoods (Schapiro et al., 2017) and to explore the spatial variance of genes (Svensson et al., 2018; Arnol et al., 2019; Nitzan et al., 2019; Sun et al., 2020). However, none of the existing methods seek to jointly model cell-to-cell relations and the interplay of factors of cell identity, which is of fundamental importance to fully utilize spatial transcriptome data.

Here, we develop a new integrative framework, called Spatial Identification of Cells using Matrix Factorization (SpiceMix), to model spatial transcriptome data. SpiceMix uses latent variable modeling to express the interplay of various spatial and intrinsic factors that comprise cell identity. It builds upon non-negative matrix factorization (NMF) (Lee and Seung, 2001), which has become a popular paradigm for latent variable modeling of gene expression (Brunet et al., 2004; Stein-O'Brien et al., 2019; Kotliar et al., 2019; Rodriques et al., 2019; Elosua et al., 2020). Crucially, SpiceMix markedly enhances the standard NMF formulation with a graphical representation of the spatial relationship of cells to explicitly capture spatial factors. SpiceMix also uses an HMRF as the graphical model, however, the model is significantly enhanced by integrating the NMF formulation of gene expression into each cell in the graph. Thus, the learned spatial patterns can elucidate the relationship of confounding intrinsic and spatial factors, leading to much more meaningful representations of cell identity. Importantly, SpiceMix does not rely on key genes from scRNA-seq data *a priori*, but rather learns spatial and intrinsic factors jointly from spatial transcriptome information directly. Both simulation and real data applications demonstrate the advantage of SpiceMix over other methods. As a proof-of-principle, we further use SpiceMix to analyze spatial transcriptomic data of the mouse cortex acquired by seqFISH+ (Eng et al., 2019) and STARmap (Wang et al., 2018), two methods that can provide single-cell expression of thousands of genes. We find that the latent representation learned by SpiceMix can refine the identification of spatially-enriched cell-types, reveal subtypes missed by other approaches, and describe important bi-

ological processes. SPICEMIX has the potential to provide critical new insights into the cell composition based on the unique spatial transcriptiome 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 transcriptomics (e.g., seqFISH+ (Eng et al., 2019) and STARmap (Wang et al., 2018)). 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 explains the observed gene expression of the cell. More critically, unique to the NMF-HMRF model is the integration of NMF into the HMRF to describe the observations as linear mixtures of latent factors, where the mixing weights 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.

The potential functions of the graph capture the probabilistic relationships between variables in the model. The potential functions on observations capture the likelihood of the observation given the hidden state of the cell. The potential functions on edges capture the spatial affinity between the latent factors of neighboring cells. In a standard HMRF, it is assumed that neighboring nodes will have similar hidden states, resulting in a spatial smoothing effect. 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 data. Crucially, SPICEMIX learns the parameters of the model that best explain the input spatial transcriptome data, while simultaneously learning the underlying latent factors and mixtures that define the identities of the cells. This is achieved by an optimization algorithm that alternates between maximizing the joint posterior distribution of the parameters of the model and the latent linear representation. From the learned parameters, factors, and mixing weights, subsequent analysis can provide insights into the biological meaning of this latent representation. See Methods for the detailed description of the SPICEMIX model.

Evaluation of SPICEMIX on simulated spatial transcriptomic data

We first evaluated SPICEMIX on simulated data that we modeled after the mouse cortex, which has been the tissue of interest in many recent studies of spatial transcriptomics, including seqFISH+ (Eng et al., 2019) and STARmap (Wang et al., 2018) (Fig. 2a-b; see Methods for detailed simulation strategy). We compared the performance of SPICEMIX with that of NMF and HMRF across eight experimental settings, which varied the signal-to-noise ratio (SNR) between either 18dB or 24dB and the number of replicates, or field-of-views (FOVs), between 1-4. We found that SPICEMIX was able to reliably recover spatially-enriched metagenes, leading to consistently improved estimation of the hidden states of cells compared to NMF and also more accurate cell labeling than NMF and HMRF. We varied the spatial regularization term of SPICEMIX, λ_Σ, which controls the influence of spatial information, between $2|\boldsymbol{\mathcal{E}}| \times 10^{-4}$ or $2|\boldsymbol{\mathcal{E}}| \times 10^{-2}$. We found that in every experimental setting, SPICEMIX markedly reduced the average relative estimation error of the hidden states of cells, compared to NMF, with reductions between 35-66%, and that a regularization term giving spatial information more influence yielded the lowest error (Fig. 2c (top)) (see Supplementary Results A.9 for details). Generally, the estimation error of some metagenes was much higher for NMF (Fig. 2c (bottom)). Qualitatively, this improvement can be clearly observed by the recovery of metagenes with significant spatial enrichment by SPICEMIX (Fig. 2d). In the representative example experiment shown, with SNR of 18 dB and using 3 FOVs, NMF

cannot identify the spatial patterning of metagenes 10 and 11, which distinguish the excitatory neurons of the associated layers. We found that, in contrast to both SpiceMix and NMF, HMRF always results in an overly-smooth result that misses sparse cell types and still misses layer-wise patterns, which is also manifested in our real data application (see later sections and Fig. S5).

Together, based on the simulation evaluation, we showed that SpiceMix can harvest spatial interaction between cells and genes and capture finer cell types, metagenes, and the spatial affinity between metagenes with improved performance compared with other approaches.

SpiceMix refines excitatory neural and glial subtype identification in seqFISH+ data

We sought to demonstrate the effectiveness of SpiceMix on different spatial transcriptome data. We first applied our method to the data acquired by seqFISH+ (Eng et al., 2019), a multiplexed imaging-based spatial transcriptomic approach. Specifically, we sought to use SpiceMix to produce a more informative latent representation of cells, based on metagenes and their spatial affinity, that refines discrete, intrinsic cell clusters and reveals continuous states of cell identity. The data from Eng et al. (2019) consist of measured single-cell expression of 10,000 genes of a single mouse in 7 FOVs: 5 FOVs from cortex and 2 FOVs from the subventricular zone (SVZ). Considering the distinct cell compositions in the cortex and SVZ, we selected and analyzed the 5 cortex FOVs only. We compared the cell identities revealed by SpiceMix to those of NMF and Eng et al. (2019).

To identify discrete, intrinsic cell types from the learned latent representation of SpiceMix, we applied hierarchical clustering, which revealed five excitatory neural subtypes, two inhibitory neural subtypes, and eight glial subtypes (Fig. 3a). These associations were established by the correlation of each cluster with established marker genes, as determined by prior scRNA-seq experiments (Tasic et al., 2016) (Fig. 3b (left)). The assignment of major types was consistent between SpiceMix, NMF, and Eng et al. (2019) (Fig. 3b (middle) and Fig. S2). More importantly, SpiceMix refined and expanded the identification of subtypes (Fig. 3b (middle)). Specifically, by leveraging spatial affinities, SpiceMix refined the assignment of neuronal subtypes. Excitatory neurons reside in specific layers of the cortex, ranging from deep (L6) to superficial (L2), and since each FOV of the data is of a subsequent layer of the cortex, each excitatory subtype should be exclusive to a single FOV. The excitatory subtypes of SpiceMix reflected this trend (Fig. 3c), whereas several excitatory clusters from the original analysis were equally distributed across as many as three FOVs (see Fig. 3h in Eng et al. (2019)). In particular, there is a notable improvement by SpiceMix among eL5b and eL6 neurons. These cells were grouped together in several clusters in (Eng et al., 2019) (Fig. 3b (middle)), but SpiceMix correctly separates them, which is confirmed by the expression of marker genes Col6a1 and Ctgf (Fig. 3b (left)), respectively, discovered by scRNA-seq (Tasic et al., 2016). SpiceMix also identified VIP and SST subtypes of inhibitory neurons, although, due to the small number of such cells and the absence of important putative marker genes in the set of 10,000 measured genes, it is yet to be validated if the analysis of Eng et al. (2019) or SpiceMix produced a more accurate assignment. In contrast, NMF grouped the VIP and SST cells into one type (Fig. 3a).

In addition, SpiceMix provided unique insight into the latent representation of glial cells and revealed several subtypes that were missed by other methods. The metagenes learned by SpiceMix tend to capture either expression patterns of intrinsic cell type expressed at high levels or patterns shared across major types and subtypes expressed at lower levels (Fig. 3b (right)). Notably, as annotated in Fig. 3b (right), metagene 7 is expressed at a high proportion among Oligo cells, separating them from OPCs, while metagene 8 is shared between the OPC and Oligo-2 types. This allowed SpiceMix to produce the separation of OPCs and Oligo-2, which was missed in the analysis of Eng et al. (2019). This separation is confirmed by the expression of the OPC marker gene Cspg4 and the lack of expression of

differentiating Oligo marker gene Tcf7l2 (Zhao et al., 2016) and mature Oligo marker gene Mog (Linington et al., 1988). Furthermore, Oligo-2 might be cells that are in an intermediate transition state during maturation from OPCs to Oligos, rather than belonging to a single, discrete cell type, and this is supported by their high expression level of Tcf7l2, a marker gene of an intermediate differentiation state (Zhao et al., 2016). In contrast, both NMF and Eng et al. (2019) failed to delineate these cells from other oligodendrocytes (Fig. 3b (middle) and Fig. S2), because of the similarity of the expression profiles of OPCs and oligodendrocytes. The accuracy of the Oligo assignments is further confirmed by their localization to a deep layer (Parnavelas et al., 1983), which is also captured by the high affinity of Oligo-1 and Oligo-2 to themselves and eL6 neurons (Fig. 3d), which reside in a deep layer. Lastly, SpiceMix revealed an additional separation of a cluster of Eng et al. (2019) into SMC and Endo cells, which can be confirmed by expression of their respective marker genes (Fig. 3b).

Taken together, by analyzing the seqFISH+ data with SpiceMix, we identified subtypes of the cortex whose spatial distributions are consistent with prior experiments. We also delineated rarer subtypes missed by other methods, including the OPC class from oligodendrocytes and VIP and SST inhibitory subtypes, and separated SMCs, Endos, and Micros as individual glial types. These analysis strongly demonstrate the advantages and unique capabilities of SpiceMix.

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

Next, we sought similarly to refine our understanding of the mouse cortex by analyzing another very recent spatial single-cell transcriptomic dataset acquired by STARmap (Wang et al., 2018). As in the analysis of the seqFISH+ dataset, the learned latent representation of cell identity by SpiceMix revealed a better characterization of cell types and subtypes and offered additional insight into continuous factors of cell state. We analyzed a single replicate consisting of 930 cells passing quality control, all within a single 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. (2018) when referenced.

We found that SpiceMix produced more precise cell type assignments than Wang et al. (2018) and revealed subtypes that were missed both in Wang et al. (2018) and by NMF. SpiceMix discovered six layer-wise excitatory neural types, three inhibitory neural types, two astrocyte types, three Oligo types, and single SMC, HPC, and Endo types (Fig. 4a). In comparison to NMF, SpiceMix uncovered an SST inhibitory neuronal type, an Oligo type, an Astro/Oligo type, and two eL6 subtypes. The accuracy of these assignments was validated by the expression of known marker genes (Tasic et al., 2016; Wang et al., 2018) (Fig. 4b (left)). In comparison to the analysis of Wang et al. (2018), SpiceMix refined the assignment of excitatory neurons and revealed Oligo subtypes that were missed (Fig. 4b (middle)). Specifically, the assignments of SpiceMix produced a clear, layer-like structure of excitatory neurons, unlike the assignments in the original analysis (see Fig. 5d in Wang et al. (2018)), which erroneously mixed several neuron subtypes across layer boundaries (Fig. 4c). We noted that $\geq$15 cells in layers 5 and 6 are classified into eL2/3* or eL4* type (black crosses in the middle plot) and $\geq$15 cells outside layer 5 are classified as eL5* neurons (black dots in the bottom plot in Fig. 4c). This refinement is especially notable in the assignment of 36 cells in excitatory subtypes eL2/3 and eL4 ('$\times$' and '$+$' in Fig. 4c (middle)). These 36 cells may belong to the L4-Arf5 subtype of excitatory neurons defined in Tasic et al. (2016), which is located at the boundary of layers 2/3 and 4. We found that NMF also corrected some of the eL2/3 and eL4 assignments (Fig. S4 (middle)), but it was not able to produce as clear of a boundary as SpiceMix, due to its inability to learn layer-specific metagenes (Fig. S3a). We also applied the HMRF of Zhu et al. (2018) and found that, in contrast, it produced overly-smooth boundaries between cell types and failed to identify the layer-wise structure of excitatory neurons (Fig. S5). This again confirms the roles of both the NMF formulation in improving cell type identification and of the spatial modeling of

SPICEMIX in further refining spatial associations.

Additionally, SPICEMIX identified a large set of eL5 neurons that were incorrectly labeled as astrocytes in Wang et al. (2018) (Fig. 4b (middle)). Astrocytes are known to diffusely populate the primary visual cortex of the mouse (Parnavelas et al., 1983), yet the Astro-1* cluster of Wang et al. (2018) showed high spatial enrichment in the L5 layer (Fig. S6c). SPICEMIX learned the high spatially affinity of eL5 metagenes to be near other eL5 cells and used this affinity to reassign these Astro-1* cells as eL5 neurons (Fig. 4b,c,e). Differential expression analysis showed that many of the marker genes of astrocytes identified by Tasic et al. (2016), 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. S6a,b). 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. S6). We found that NMF also produced a cell cluster eL5 that is a mixture of Astro-1* and eL5* (Fig. S4 (middle)), which confirms that this was not a mere artifact of the learned spatial affinity. We reason that these cells were classified as non-neuronal types in the original STARmap analysis because they express the Mgp gene at a high level.

Notably, the learned latent representation of SPICEMIX provides insights into the relation of different expression patterns across types, especially subtypes. Similar to the seqFISH+ analysis, each cell type generally expressed one metagene, often unique to that cell type and at a high level, that may capture intrinsic processes for each type, along with other metagenes expressed at a lower level that capture subtypes and layers (Fig. 4b (right)). SPICEMIX discovered layer-specific metagenes (metagenes 1-5), which were expressed highly in HPCs and excitatory neurons, but were also expressed at lower levels in several other cell types that were co-located in the respective layer, including inhibitory neurons and some oligodendrocytes. Another important observation is the expression of metagenes 12 and 13 among Oligo-1 and Astro/Oligo, but at contrary levels (Fig. 4f), which shed new light on the relation of these cells (see below). The estimation of metagenes and their proportions was refined by the learned spatial affinity (Fig. 4e), which allowed SPICEMIX to learn more coherent patterns of spatial enrichment and spatial sparsity, compared to NMF, which exhibited diffuse, unspecific patterns of metagene expression (Fig. 4d and Fig. S3a).

These results collectively suggest that, by incorporating spatial interaction between cells and metagenes, SPICEMIX is able to refine excitatory neural types with better spatial pattern, and identify finer glial subtypes, further demonstrating its unique advantage compared with other approaches.

SPICEMIX identifies continuous myelination stages in oligodendrocytes

We next asked whether the learned metagenes of SPICEMIX could uncover the factors of cellular identity that cannot be captured by discrete cluster assignments. One such factor, which was revealed only by SPICEMIX, shed new light on 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). To preface, we note that the Astro/Oligo type (colored pink in Fig. 4) included a distinct group of astrocytes as well, which we removed from consideration. These cells correspond to the thin sliver of pink-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. S7a). 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 anti-correlated, levels (Fig. 4f). We hypothesized that the

differential expression of these metagenes was capturing the myelination process of oligodendrocytes. To test this hypothesis, we used linear regression models to fit the relationship between the expression levels of myelin sheath-related genes and the proportions of metagenes 12 and 13 in individual cells. The eleven genes which 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.2), and were expressed in at least 30% of the cells. We found that correlations of seven of the eleven genes were significant ($p<0.05$) (Fig. 4f and Fig. S7), supporting our hypothesis. Furthermore, we identified one gene, Atp1a2, which was confirmed recently by scRNA-seq (Marques et al., 2018; Beiter et al., 2020).

These results show 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 profound 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 constitute a cell's identity based on single-cell spatial transcriptome data. 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 brain transcriptome, showing a clear advantage over NMF. Importantly, SpiceMix revealed cell metagenes that exhibited strong spatial patterns missed by NMF. Additionally, the ability of SpiceMix to represent cells as mixtures of latent factors and to learn spatial affinities yielded major advantages over HMRF.

As a proof-of-principle, we applied SpiceMix to real single-cell spatial transcriptomic data of the mouse cortex from seqFISH+ and STARmap. We demonstrated the effectiveness of the method to produce reliable and meaningful latent representations of *in situ* single-cell gene expression. These learned representations revealed strong spatial enrichment patterns that confirm the layer-wise structure of the mouse cortex, but were less apparent in the analysis of other methods. This in turn led to more accurate cell type identification than prior approaches and also revealed missed subtypes. Specifically, SpiceMix more accurately identified layer-specific excitatory neuronal subtypes, revealed sparse inhibitory neuronal subtypes, and unveiled the maturation process of oligodendrocytes. This maturation process was uncovered by latent factors learned by SpiceMix from both STARmap and seqFISH+ data and led to the identification of several oligodendrocyte subtypes, which were missed by other methods. This discovery highlights the advantage of SpiceMix over existing methods, particularly those that assume discrete cell-type assignments, such as the standard HMRF (Zhu et al., 2018), which are unable to express such continuous processes of cell identity.

As future work, SpiceMix could be further enhanced by incorporating scRNA-seq data from additional experiments. Other recent computational methods have been used to study scRNA-seq and spatial transcriptomic data jointly (Stuart et al., 2019; Welch et al., 2019; Korsunsky et al., 2019), but they do not attempt to comprehensively model spatial cell-to-cell relationships. With proper normalization and pre-processing, data from scRNA-seq could be incorporated into the SpiceMix model 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. An optimization strategy based on variational inference could be developed to reduce runtime, as larger datasets become available.

As the area of spatial transcriptomics continues to thrive and data become more widely available,

SPICEMIX will be a useful tool for enabling new discoveries. In particular, refined cell identity with SPICEMIX has the potential to improve future studies of cell-cell interactions (Armingol et al., 2020). 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 (Liu et al., 2020). Overall, SPICEMIX is a powerful framework that can serve as an essential tool for the analysis of diverse types of spatial transcriptomic 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 formulation of the standard non-negative matrix factorization (NMF) formulation of prior work to model gene expression (Brunet et al., 2004). 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 SpiceMix enhances standard NMF by incorporating the spatial correlations among samples (i.e., cells in this context). This 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, such as distance thresholding or Delaunay triangulation, can be used. We utilize the hidden Markov random field (HMRF) (Zhang et al., 2001). For all the nodes in the graph, the measured gene expression, y_i, are the observed variables and the latent factorization, x_i, are the hidden variables. The observations are related to the hidden variables via the potential function ϕ, which captures the NMF formulation. The spatial affinity between the latent factorization 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 (Murphy, 2012), 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 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 normalized factorization 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 factorization 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, which is summarized in Algorithm 1. First, to make inference tractable, we approximate the joint probability of the hidden states by the pseudo-likelihood (Murphy, 2012), 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:

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

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

This is a quadratic program and can be solved efficiently via the iterated conditional model (ICM) (Besag, 1986) using the software package Gurobi (Gurobi Optimization, 2020) (see Supplementary Methods A.7 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 (Murphy, 2012), 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 (Kingma and Ba, 2014). 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.8 for details of the optimization method.

Initialization

To produce initialize estimates of the model parameters and hidden states, we do the following. First, we 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. This is a common strategy for initializing NMF. We then alternate for T_0 iterations between solving the NMF objective for X and M. This produces, in only a few quick iteration, 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 only in around the first 5 iterations, where the spatial affinity matrix Σ_x^{-1} is changed significantly. Later on, most time is spent on 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 (Tasic et al., 2016; Lein et al., 2007). This region in the brain consists of several layers, and certain cell types are enriched in specific layers. 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 created simulated data to emulate such patterns by sectioning the image, or replicate, of the data into four layers, with relative widths of 0.3, 0.2, 0.25, and 0.25, and assigning cells to spatial subtypes accordingly (Fig. 2a). Every excitatory subtype is enriched in exactly one layer. For each cell type, the composition of metagenes, several of which are layer-specific, defined the major type and minor subtype of cells (Fig. 2b), which follow the observed trends from mouse cortex data. The gene expression of a cell within a subtype is a mixture of major-type, subtype, and spatial metagenes. For each FOV, 700 cells were created with randomly generated locations so as to maintain a minimum distance between any two cells, so that the density of cells across the FOV was roughly constant. For each experiment, we generated between one to four replicates, or FOVs, to characterize the improvement in performance using additional data. We assigned the membership, or proportion, of metagenes for each cell to be roughly 60% major type, 30% minor type, and 10% a mixture of a common extraneous metagene or a layer-specific metagene shared by another cell type. We used this ratio as the mean vector of a multi-dimensional exponential distribution, constrained to the simplex, from which we sampled the hidden state proportions for individual cells. After the cell memberships are determined, they are scaled by a random scaling factor from the Gamma distribution, with parameter value of 1, to simulate variation in numbers of counts in a way that is independent of cell type. This yields the final hidden state x_i for each cell i.

The procedure to generate each metagene m_k is as follows. We drew an initial value $M_{g,k}$ for the expression of gene g for metagene k at random from the Gamma distribution with a parameter of 2.5, which balances the level of sparsity of the metagene. However, we enforced that the metagenes of similar types share common expression patterns of some genes. We let the metagenes capture the major type of inhibitory and excitatory neurons to share 75% of genes. Of inhibitory neurons and glial cells, we also made 75% of genes to have the same expression values. Of layer-specific metagenes, which also define excitatory neuronal subtypes, we let 90% of genes have the same expression values. This patterning follows what was observed of metagenes from analysis of real spatial transcriptomic data in this work. After the initial values for each gene were drawn, we normalized each metagene to sum to one. In this way, any scaling is captured entirely by the hidden variables x_k.

The observed expression $Y_{g,i}$ of cell i for gene g is a linear combination of intrinsic, spatial subtype, and extrinsic metagenes with added Gaussian noise, $y_i = Mx_i + e_i$, where $e_i \sim \boldsymbol{\mathcal{N}}(0, \sigma_y^2 I_G)$ and I_G is the identity matrix of dimension G. In our experiments, we set $G = 500$ genes. We considered signal-to-noise ratios (SNR) of 18dB and 27dB, where the SNR is calculated as the ratio of the average observed

expression value of a gene for a cell to the noise level,

$$\text{SNR} = 20 \log_{10} \left(E[Y_{g,i}]^2 / \sigma_y^2 \right).$$

Combined with the variation of number of FOVs between 1-4, this provided eight datasets for assessing model performance. For details on hyperparamter selection for the results on this synthetic data, see Supplementary Methods A.1.

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.2 for details. Steps of data processing include: constructing the neighbor graph of cells, selection of hyperparameters for both SPICEMIX and NMF, 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.3 for details of these steps, as well as Supplementary Methods A.4 and A.5 for specific details of the number of clusters and metagenes for analysis of seqFISH+ and STARmap, respectively. For the explanation of our method for constructing the cell-type affinity matrix, see Supplementary Methods A.6.

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

D. Arendt, J. M. Musser, C. V. Baker, A. Bergman, C. Cepko, D. H. Erwin, M. Pavlicev, G. Schlosser, S. Widder, M. D. Laubichler, et al. The origin and evolution of cell types. *Nature Reviews Genetics*, 17(12):744, 2016.

E. Armingol, A. Officer, O. Harismendy, and N. E. Lewis. Deciphering cell–cell interactions and communication from gene expression. *Nature Reviews Genetics*, pages 1–18, 2020.

D. Arnol, D. Schapiro, B. Bodenmiller, J. Saez-Rodriguez, and O. Stegle. Modeling cell-cell interactions from spatial molecular data with spatial variance component analysis. *Cell Reports*, 29(1):202–211, 2019.

R. M. Beiter, A. Fernández-Castañeda, C. Rivet-Noor, A. Merchak, R. Bai, E. Slogar, S. M. Seki, D. A. Rosen, C. C. Overall, and A. Gaultier. Evidence for oligodendrocyte progenitor cell heterogeneity in the adult mouse brain. *bioRxiv*, 2020.

J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(3):259–279, 1986.

J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, 2004.

K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, and X. Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233):aaa6090, 2015.

X. Chen, S. A. Teichmann, and K. B. Meyer. From tissues to cell types and back: Single-cell gene expression analysis of tissue architecture. *Annual Review of Biomedical Data Science*, 1:29–51, 2018.

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

M. Elosua, P. Nieto, E. Mereu, I. Gut, and H. Heyn. SPOTlight: Seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *bioRxiv*, 2020.

C.-H. L. Eng, M. Lawson, Q. Zhu, R. Dries, N. Koulena, Y. Takei, J. Yun, C. Cronin, C. Karp, G.-C. Yuan, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, 568: 235–239, 2019.

L. Gurobi Optimization. Gurobi optimizer reference manual, 2020. URL http://www.gurobi.com.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P.-r. Loh, and S. Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, pages 1–8, 2019.

D. Kotliar, A. Veres, M. A. Nagy, S. Tabrizi, E. Hodis, D. A. Melton, and P. C. Sabeti. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife*, 8, 2019.

D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, pages 556–562, 2001.

J. H. Lee, E. R. Daugharthy, J. Scheiman, R. Kalhor, J. L. Yang, T. C. Ferrante, R. Terry, S. S. Jeanty, C. Li, R. Amamoto, et al. Highly multiplexed subcellular RNA sequencing in situ. *Science*, 343 (6177):1360–1363, 2014.

E. Lein, L. E. Borm, and S. Linnarsson. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science*, 358(6359):64–69, 2017.

E. S. Lein, M. J. Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernard, A. F. Boe, M. S. Boguski, K. S. Brockway, E. J. Byrnes, et al. Genome-wide atlas of gene expression in the adult mouse brain.

Nature, 445(7124):168–176, 2007.

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

Y. Liu, M. Yang, Y. Deng, G. Su, A. Enninful, C. C. Guo, T. Tebaldi, D. Zhang, D. Kim, Z. Bai, E. Norris, A. Pan, J. Li, Y. Xiao, S. Halene, and R. Fan. High-spatial-resolution multi-omic sequencing via deterministic barcoding in tissue. *Cell*, 2020.

S. Marques, D. van Bruggen, D. P. Vanichkina, E. M. Floriddia, H. Munguba, L. Väremo, S. Giacomello, A. M. Falcão, M. Meijer, Å. K. Björklund, et al. Transcriptional convergence of oligodendrocyte lineage progenitors during development. *Developmental Cell*, 46(4):504–517, 2018.

J. R. Moffitt, D. Bambah-Mukku, S. W. Eichhorn, E. Vaughn, K. Shekhar, J. D. Perez, N. D. Rubinstein, J. Hao, A. Regev, C. Dulac, et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, 362(6416):eaau5324, 2018.

K. Murphy. *Machine learning: a probabilistic perspective*. MIT Press, 2012.

M. Nitzan, N. Karaiskos, N. Friedman, and N. Rajewsky. Gene expression cartography. *Nature*, 576 (7785):132–137, 2019.

J. Parnavelas, R. Luder, S. Polland, K. Sullivan, and A. Lieberman. A qualitative and quantitative ultrastructural study of glial cells in the developing visual cortex of the rat. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 301(1103):55–84, 1983.

S. G. Rodriques, R. R. Stickels, A. Goeva, C. A. Martin, E. Murray, C. R. Vanderburg, J. Welch, L. M. Chen, F. Chen, and E. Z. Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, 2019.

D. Schapiro, H. W. Jackson, S. Raghuraman, J. R. Fischer, V. R. Zanotelli, D. Schulz, C. Giesen, R. Catena, Z. Varga, and B. Bodenmiller. histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nature Methods*, 14(9):873, 2017.

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

P. L. Ståhl, F. Salmén, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

G. L. Stein-O'Brien, B. S. Clark, T. Sherman, C. Zibetti, Q. Hu, R. Sealfon, S. Liu, J. Qian, C. Colantuoni, S. Blackshaw, L. A. Goff, and E. J. Fertig. Decomposing cell identity for transfer learning across cellular measurements, platforms, tissues, and species. *Cell Systems*, 8(5):395–411, 2019.

T. Stuart and R. Satija. Integrative single-cell analysis. *Nature Reviews Genetics*, 20(5):257–272, 2019.

T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, 2019.

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

V. Svensson, S. A. Teichmann, and O. Stegle. SpatialDE: identification of spatially variable genes. *Nature Methods*, 15(5):343, 2018.

B. Tasic, V. Menon, T. N. Nguyen, T. K. Kim, T. Jarsky, Z. Yao, B. Levi, L. T. Gray, S. A. Sorensen, T. Dolbeare, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience*, 19(2):335, 2016.

C. Trapnell. Defining cell types and states with single-cell genomics. *Genome Research*, 25(10):1491–1498, 2015.

S. Vickovic, G. Eraslan, F. Salmén, J. Klughammer, L. Stenbeck, D. Schapiro, T. Äijö, R. Bonneau,

L. Bergenstråhle, J. F. Navarro, et al. High-definition spatial transcriptomics for in situ tissue profiling. *Nature Methods*, 16(10):987–990, 2019.

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

X. Wang, W. E. Allen, M. A. Wright, E. L. Sylwestrak, N. Samusik, S. Vesuna, K. Evans, C. Liu, C. Ramakrishnan, J. Liu, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 341(6400):eaat5691, 2018.

J. D. Welch, V. Kozareva, A. Ferreira, C. Vanderburg, C. Martin, and E. Z. Macosko. Single-cell multiomic integration compares and contrasts features of brain cell identity. *Cell*, 177(7):1873–1887.e17, 2019.

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

C. Zhao, Y. Deng, L. Liu, K. Yu, L. Zhang, H. Wang, X. He, J. Wang, C. Lu, L. N. Wu, et al. Dual regulatory switch through interactions of Tcf7l2/Tcf4 with stage-specific partners propels oligodendroglial maturation. *Nature Communications*, 7(1):1–15, 2016.

Q. Zhu, S. Shah, R. Dries, L. Cai, and G.-C. Yuan. Identification of spatially associated subpopulations by combining scrnaseq and sequential fluorescence in situ hybridization data. *Nature Biotechnology*, 36(12):1183, 2018.

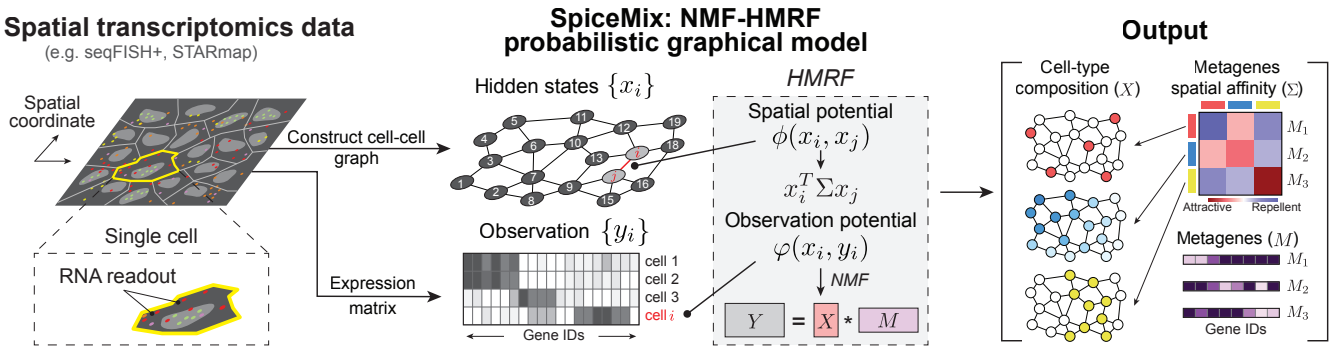

Figure 1: Overview of SPICEMIX. Gene expression measurements and a neighbor graph are extracted from *in situ* single-cell spatial trancriptome data and fed into the SPICEMIX framework. SPICEMIX decomposes the expression y_i in cell i as a mixture of metagenes weighted by its latent state x_i, and models spatial interaction between cells i and j as an inner product of their latent states, weighted by inferred spatial affinities between metagenes Σ. Collectively, SPICEMIX uncovers mixture weights for individual cells X, metagene spatial affinity Σ, and K metagenes M. Hierarchical clustering is applied to the latent space to define categorical cell types in an unsupervised manner.

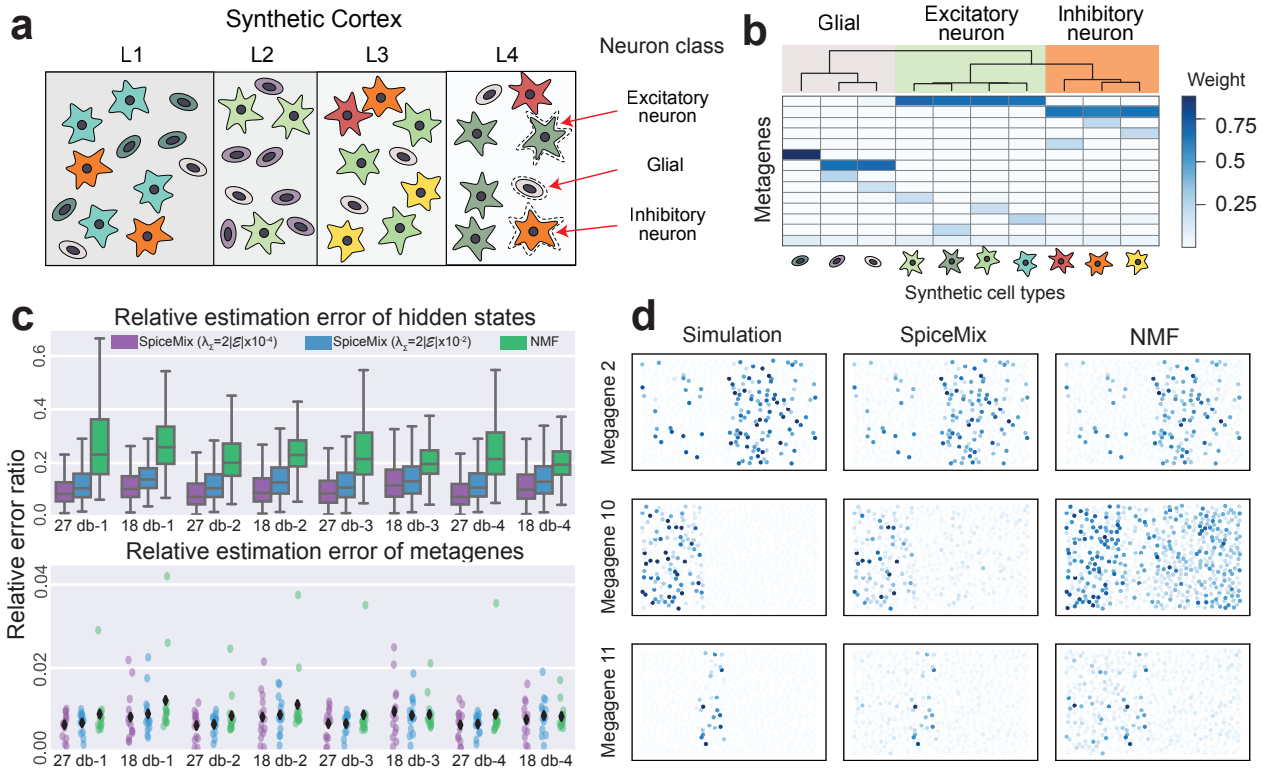

Figure 2: Overview of the simulated spatial transcriptomic data of the mouse cortex, and performance comparison between SoiceMix and NMF. **a.** Illustration of three major cell types distributed in four layers. Excitatory and inhibitory neurons are represented as asterisks, colored in cold and warm colors, respectively. Glial cells are denoted by ovals. Subtypes are distinguished by minor difference in colors. **b.** Dendrogram of 10 subtypes (top) and their metagene profiles (bottom). Metagenes 1 and 2 are pan-excitatory and pan-inhibitory, respectively. Every subtype has its own metagene. **c.** Comparison of estimation error of hidden states (top), shown as box plots, and metagenes (bottom), shown as scatter plots, across eight experiments. Black diamonds on the lower plot show the mean error. The label for each experiment gives the noise level (in dB) followed by the number of fields-of-view used. **d.** Comparison of estimation of hidden states *in situ* for some example metagenes between the simulation (ground truth), SpiceMix, and NMF for the example experiment.

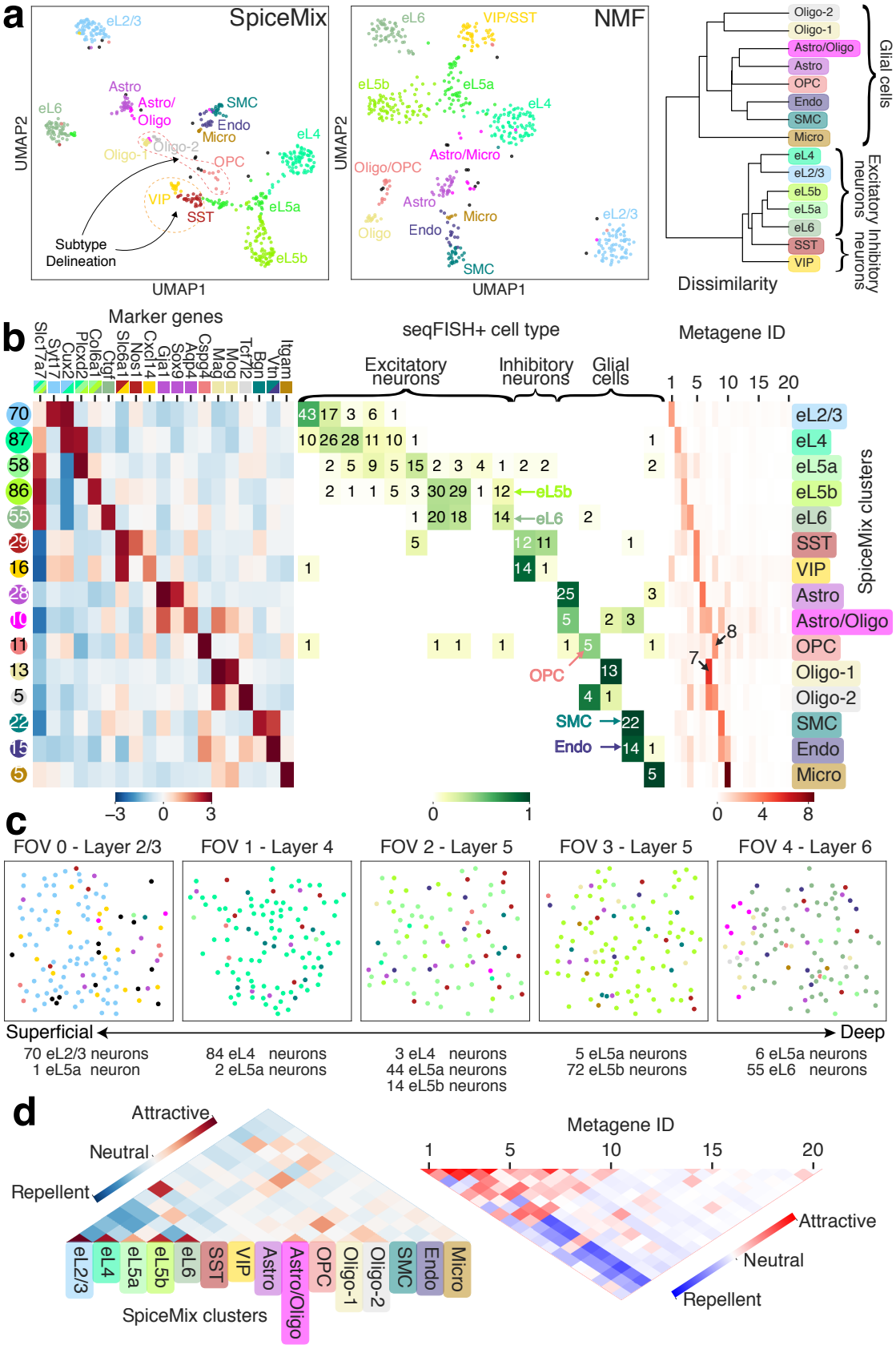
a
SpiceMix
NMF
UMAP1
UMAP2
Subtype Delineation
Dissimilarity
Glial cells
Excitatory neurons
Inhibitory neurons
b
Marker genes
seqFISH+ cell type
Metagene ID
SpiceMix clusters
c
FOV 0 - Layer 2/3
FOV 1 - Layer 4
FOV 2 - Layer 5
FOV 3 - Layer 5
FOV 4 - Layer 6
Superficial
Deep
70 eL2/3 neurons
1 eL5a neuron
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
Metagene ID
SpiceMix clusters

Figure 3 ***(preceding page)*****:** Application of SpiceMix to the seqFISH+ data from mouse brain cells (Eng et al., 2019). Note that colors throughout the figure of cells and labels correspond to cell-type assignments. **a.** UMAP plots of the latent states of SpiceMix (left) and NMF (middle) as well as the Dendrogram of SpiceMix states (right). **b.** (Left) Average expression of known marker genes within SpiceMix cell types, along with the number of cells belonging to each type (colored circles). (Middle) Agreement of SpiceMix cell-type assignments with those of the original analysis in Eng et al. (2019). (Right) Average expression of inferred metagenes within SpiceMix cell types. **c.** SpiceMix cell-type assignments for all cells in each of the 5 FOVs. **d.** (Left) The inferred pairwise spatial affinity of cell types. (Right) The inferred pairwise spatial affinity of metagenes, or Σ_x^{-1}.

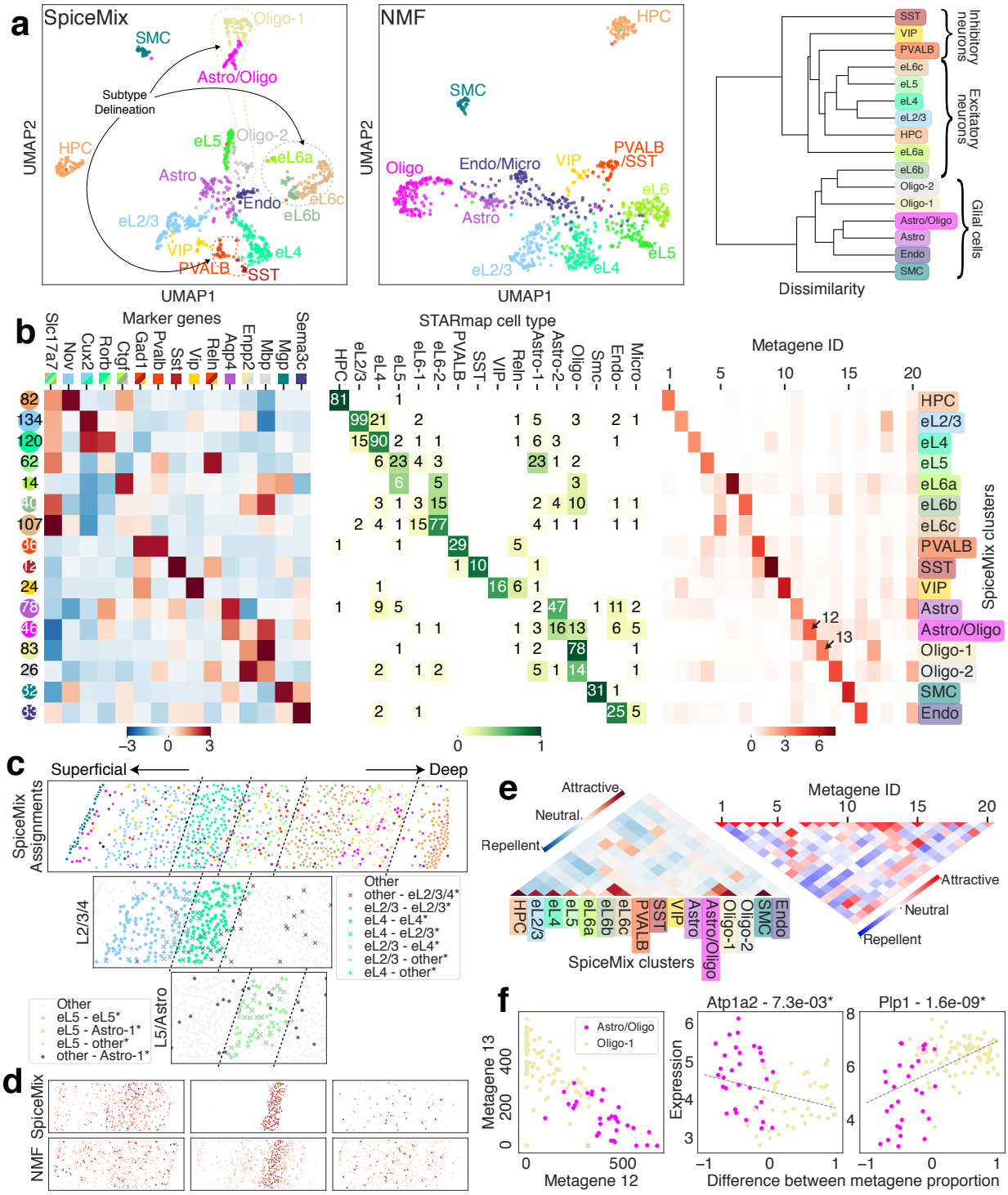
a
SpiceMix
Oligo-1
SMC
Astro/Oligo
Subtype Delineation
HPC
eL5
Oligo-2
eL6a
Astro
Endo
eL6c
eL6b
eL2/3
VIP
eL4
PVALB
SST
UMAP2
UMAP1
NMF
HPC
SMC
Oligo
Endo/Micro
VIP
PVALB /SST
eL6
Astro
eL5
eL2/3
eL4
Inhibitory neurons
Excitatory neurons
Glial cells
Dissimilarity
b
Marker genes
Slc17a7
Nov
Cux2
Rorb
Ctgf
Gad1
Pvalb
Sst
Vip
Reln
Aqp4
Enpp2
Mbp
Mgp
Sema3c
STARmap cell type
Metagene ID
SpiceMix clusters
c
Superficial
Deep
SpiceMix Assignments
L2/3/4
L5/Astro
Other
other - eL2/3/4*
eL2/3 - eL2/3*
eL4 - eL4*
eL4 - eL2/3*
eL2/3 - eL4*
eL2/3 - other*
eL4 - other*
eL5 - eL5*
eL5 - Astro-1*
eL5 - other*
other - Astro-1*
d
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 brain cells (Wang et al., 2018). Colors throughout the figure of cells and labels correspond to cell-type assignments. **a.** UMAP plots of the latent states of SPICEMIX (left) and NMF (middle) as well as the Dendrogram of SPICEMIX states (right). **b.** (Left) Average expression of known marker genes within SPICEMIX cell types, along with the number of cells belonging to each type (colored circles). (Middle) Agreement of SPICEMIX cell-type assignments with those of the original analysis in Wang et al. (2018). (Right) Average expression of inferred metagenes within SPICEMIX cell types. **c.** SPICEMIX cell-type assignments for all cells in the FOV are shown at the top. Lower panels show comparisons of SPICEMIX clusters and those of the original analysis (Wang et al., 2018) for specific cell types. The asterisk denotes cell-type labels from Wang et al. (2018). **d.** Example spatial maps of expression of metagenes from SPICEMIX (top) and NMF (bottom). **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.2 for details).