frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 25 April 2024
DOI 10.3389/fgene.2024.1356709

OPEN ACCESS

EDITED BY
Himel Mallick,
Cornell University, United States

REVIEWED BY
Suvo Chatterjee,
Indiana University, United States
Boyu Ren,
McLean Hospital, United States
Xiangyu Luo,
Renmin University of China, China
Piyali Basak,
Merck, United States

*CORRESPONDENCE
Qiwei Li,
qiwei.li@utdallas.edu

RECEIVED 16 December 2023
ACCEPTED 08 April 2024
PUBLISHED 25 April 2024

CITATION
Yang J, Jiang X, Jin KW, Shin S and Li Q (2024), Bayesian hidden mark interaction model for detecting spatially variable genes in imaging-based spatially resolved transcriptomics data. *Front. Genet.* 15:1356709.
doi: 10.3389/fgene.2024.1356709

Bayesian hidden mark interaction model for detecting spatially variable genes in imaging-based spatially resolved transcriptomics data

Jie Yang[1], Xi Jiang[2], Kevin Wang Jin[3], Sunyoung Shin[4] and Qiwei Li[1]*

[1]Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, TX, United States, [2]Department of Statistics and Data Science, Southern Methodist University, Dallas, TX, United States, [3]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, United States, [4]Department of Mathematics, Pohang University of Science and Technology, Pohang, Republic of Korea

Recent technology breakthroughs in spatially resolved transcriptomics (SRT) have enabled the comprehensive molecular characterization of cells whilst preserving their spatial and gene expression contexts. One of the fundamental questions in analyzing SRT data is the identification of spatially variable genes whose expressions display spatially correlated patterns. Existing approaches are built upon either the Gaussian process-based model, which relies on *ad hoc* kernels, or the energy-based Ising model, which requires gene expression to be measured on a lattice grid. To overcome these potential limitations, we developed a generalized energy-based framework to model gene expression measured from imaging-based SRT platforms, accommodating the irregular spatial distribution of measured cells. Our Bayesian model applies a zero-inflated negative binomial mixture model to dichotomize the raw count data, reducing noise. Additionally, we incorporate a geostatistical mark interaction model with a generalized energy function, where the interaction parameter is used to identify the spatial pattern. Auxiliary variable MCMC algorithms were employed to sample from the posterior distribution with an intractable normalizing constant. We demonstrated the strength of our method on both simulated and real data. Our simulation study showed that our method captured various spatial patterns with high accuracy; moreover, analysis of a seqFISH dataset and a STARmap dataset established that our proposed method is able to identify genes with novel and strong spatial patterns.

KEYWORDS

zero-inflated negative binomial mixture model, bayesian mark interaction model, spatial transcriptomics, energy function, double metropolis-hastings algorithm

1 Introduction

Recent advancements in spatially resolved transcriptomics (SRT) technology have fundamentally transformed our capacity to study cellular behavior at a molecular level, while preserving their spatial and gene expression contexts. This technological leap has opened new avenues for exploring complex biological systems at unprecedented levels of detail and accuracy. Efremova et al. (2020) and Liao et al. (2021) found that the positional

context of gene expression is important to understanding tissue functionality and pathology changes, which highlights the pivotal role of SRT techniques. Broadly, SRT technologies are categorized into sequencing-based and imaging-based methods based on differences in RNA profiling: sequencing-based and imaging-based. Spatial transcriptomics, one of the next-generation sequencing (NGS) technologies, resolves gene expression profiles at a resolution of 100 μm. Spatial transcriptomics implemented by the 10x Visium platform achieved 55 μm resolution, allowing for a detailed study of spatial organization. On the other hand, imaging-based technologies have revolutionized the field of transcriptomics by achieving single-cell resolution, with prominent examples such as sequential fluorescence *in situ* hybridization seqFISH (Ståhl et al., 2016), seqFISH+ (Eng et al., 2019), and multiplexed error-robust FISH (MERFISH) (Moffitt et al., 2018). Datasets profiled by SRT technologies have inspired the exploration of the spatial organization of gene expression within tissues. Cohorts with single-cell resolution motivate more biological analysis, such as cell-cell communication analysis *via* CellChat (Jin et al., 2021), characterization of ligand-receptor interactions between different cell types (Efremova et al., 2020) and so on. Hence, spatial information provided by imaing-based SRT data makes it more feasible to identify and quantify gene expression in specific regions of a tissue.

One of the most interesting questions arising along the development of SRT techniques is the identification of spatially variable genes (SVGs) whose expressions display spatially correlated patterns. Studies have found that SVGs demarcate clear spatial substructure, and are relevant to disease progression (Svensson et al., 2018; Hu et al., 2021). Various methods across different fields have been developed to identify SVGs, each capitalizing on distinct strengths. Trendsceek (Edsgärd et al., 2018) is built upon marked point processes to rank and evaluate the spatial pattern of each gene; however, it yields unsatisfactory performance (Sun et al., 2020) and is inhibited from scaling to large-scale data due to the expensive computational cost (Sun et al., 2020; Dries et al., 2021). SpatialDE (Svensson et al., 2018), SPARK (Sun et al., 2020), and BOOST-GP (Li et al., 2021) capture spatial correlation patterns by utilizing the Gaussian process. Specifically, SpatialDE models normalized gene expression levels *via* a multivariate Gaussian model with a spatial covariance function characterizing linear and periodic spatial patterns. SPARK models raw counts using a generalized linear spatial model with different periodic and Gaussian kernels. BOOST-GP models raw counts with a Bayesian zero-inflated negative binomial (ZINB) model with a squared exponential kernel covariance matrix. However, the performance of these kernel-based methods relies heavily on the resemblance between the underlying spatial expression patterns and the predefined kernel functions (Jiang et al., 2022). BinSpect (Dries et al., 2021), a non-model based method, identifies SVGs through statistical enrichment analysis of spatial network neighbors with binarized gene expression states. SpaGCN (Hu et al., 2021) defines SVGs as genes as those exhibiting differential expression among spatial domains and employs a deep learning model to identify these domains. BOOST-MI (Jiang et al., 2022) utilizes an energy-based modified Ising model to identify SVGs exclusively for sequencing-based SRT data, with the limitation that the spatial position of measured spots needs to be on the regular lattice grid. Compared to kernel-based models, energy-based interaction characterization enables the detection of broader types of gene spatial expression patterns.

As mentioned, gene expressions resolved by imaging-based SRT have single-cell resolution, which potentially unearths more biological insights. We aimed to develop a model that can identify SVGs with higher accuracy to be used on data from imaging-based SRT platforms, and uncover more biological mechanisms. Drawing inspiration from the success of energy-based models over kernel-based approaches (Jiang et al., 2022), we propose a novel joint Bayesian framework model, BOOST-HMI. This model utilizes a recently proposed energy function for mark interaction (Li et al., 2019). In particular, we adopt a ZINB mixture model to handle the unique data characteristics of SRT, including excess zeros and unknown mean-variance structures. Additionally, our method introduces a latent binary gene expression indicator to distinguish high and low expression states at the cellular level, thereby enhancing the model's robustness against noise. Unlike BOOST-MI, our proposed BOOST-HMI is not constrained by the spatial distribution requirements of sequencing-based SRT data, making it versatile for imaging-based datasets where cells are randomly distributed. Furthermore, BOOST-HMI directly models raw counts within a joint Bayesian framework, addressing uncertainties associated with dichotomization. Our comprehensive simulation studies, covering various scenarios, demonstrate the superior accuracy of BOOST-HMI in detecting SVGs. We also applied our model to two real datasets: a mouse hippocampus seqFISH dataset and a mouse visual cortex STARmap dataset, where it successfully detected more spatial patterns and layer-specific SVGs, potentially unveiling novel biological insights.

The rest of the paper is organized as follows: section 2 introduces our ZINB mixture model for identifying SVGs from SRT count data and discusses the extension of the Bayesian mark interaction model to SRT data. In section 3, we describe the Markov chain Monte Carlo (MCMC) algorithms for posterior sampling and the resulting posterior inference. Finally, section 4 presents our method's performance on simulated and real SRT datasets, compared with five other methodologies.

2 Methods

In this section, we introduce a ZINB mixture model for directly modeling the imaging-based SRT count data, and a hidden mark interaction model to quantify the spatial dependency of latent binary gene expression levels. The schematic diagram of BOOST-HMI is shown in Figure 1, and the graphical and hierarchical representations are presented in Supplementary Figure S5 and Supplementary Table S1, respectively, in the Supplementary Material.

Before introducing the models, we summarize the SRT data notations as follows. We denote the gene expression raw counts as a n-by-p matrix $\boldsymbol{Y}$ with each entry $y_{ij} \in \mathbb{N}$ denoting the number of read counts for gene j at cell i. Every column $\boldsymbol{y}_{\cdot j}$ in $\boldsymbol{Y}$ denotes the expression counts across all measured cells for gene j, while each row $\boldsymbol{y}_{i\cdot}$ denotes the counts of all genes on cell i where $i = 1, \ldots, n, j = 1, \ldots, p$. As to geospatial profile, let a n-by-2 matrix T be the matrix for the spatial location of cells, where each row $\boldsymbol{t}_{i\cdot} = (t_{i1}, t_{i2}) \in \mathbb{R}^2$ records the coordinates of cell i in the 2D Cartesian plane.

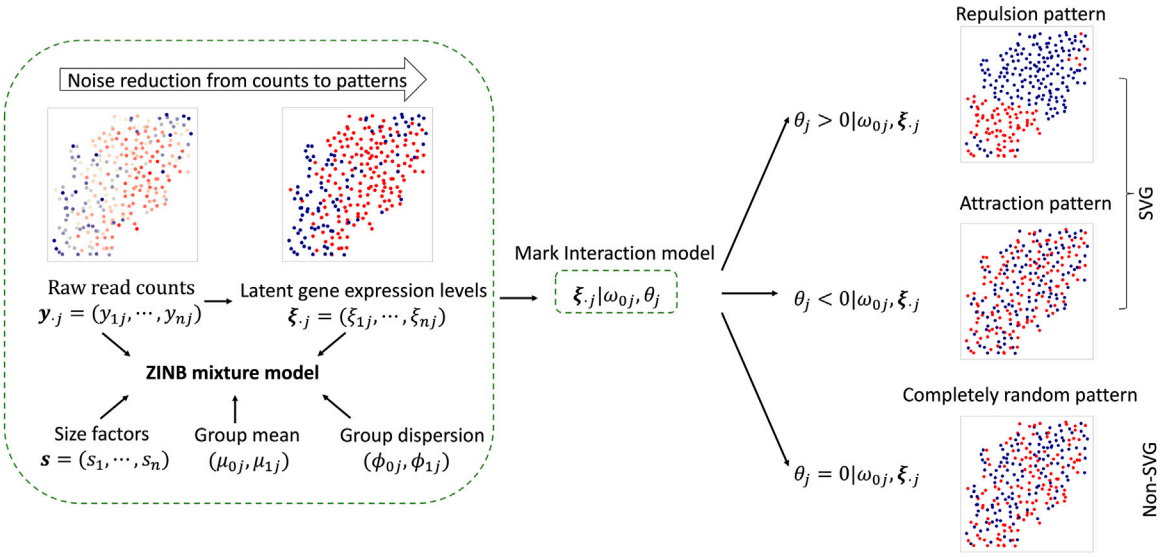

FIGURE 1
The schematic diagram of the proposed BOOST-HMI model.

2.1 A ZINB mixture model for modeling gene expression count data

For the majority of SRT techniques, gene expression measurements obtained are in the form of counts. In the context of for imaging-based SRT platforms, gene expressions are collected as the count of barcoded mRNA corresponding to a particular transcript within a single cell (Zhao et al., 2022). Due to the characteristics of these measurements, observed count data often suffers from over-dispersion and zero-inflation. The negative binomial distribution can effectively account for the mean-variance relationship in the raw counts. Moveover, the gene expression count matrix $\boldsymbol{Y}$ is characterized by an inflated number of zeros, resulting from imaging sensitivity and hybridization efficiency (Zhao et al., 2022); therefore, we generalized the negative binomial (NB) model to the ZINB model to account for both the over-dispersion and the high sparsity level, i.e.,

$$y_{ij}|\pi_i, \nu_{ij}, \phi_j \sim \pi_i I\left(y_{ij} = 0\right) + (1-\pi_i)\mathrm{NB}\left(s_i\nu_{ij}, \phi_j\right), \text{ or }, y_{ij}|\pi_i, \nu_{ij}, \phi_j \sim \mathrm{ZINB}\left(\pi_i, s_i\nu_{ij}, \phi_j\right), \tag{1}$$

where parameter $\pi_i \in [0, 1]$ represents the false zero proportion measured on cell i. NB(ν, ϕ) denotes a negative binomial distribution with mean ν and dispersion parameter ϕ. Consequently, the variance is $\nu + \nu^2/\phi$. $1/\phi$ controls the overdispersion scaled by the square of mean ν^2. The probability mass function is $\frac{\Gamma(y+\phi)}{y!\Gamma(\phi)}\left(\frac{\phi}{\nu+\phi}\right)^{\phi}\left(\frac{\nu}{\nu+\phi}\right)^{y}$. Given our particular circumstances, the NB mean is decomposed into two multiplicative effects, the size factor s_i and the expression level ν_{ij}. The collection of $\boldsymbol{s} = (s_1, \ldots, s_n)^{\top}$ reflects nuisance effects across cells. We follow SPARK Sun et al. (2020), setting s_i proportional to the summation of the total number of read counts across all genes for cell i, and combine it with a constraint of $\prod_i s_i = 1$, which gives $s_i = \sum_j y_{ij} / \prod_i \sum_j y_{ij}$. By setting the constraint for s_i's, we avoid the identifiability problem between s_i's and ν_{ij}'s.

To denoise the relative expression levels, we aim to partition ν_{ij} into two groups by introducing the ZINB mixture model. Dichotomization has been widely applied as a step in the analysis of SRT data. BinSpect (Dries et al., 2021) and BOOST-MI (Jiang et al., 2022) discretize the normalized expression levels for each gene into two groups for more robust SVG detection results. Here, we introduce a latent binary gene expression level indicator vector $\boldsymbol{\xi}_{\cdot j}$ to denote the dichotomized expression profiles of each gene j. If $\xi_{ij} = 1$, gene j is highly expressed at cell i, and if $\xi_{ij} = 0$, gene j has low expression at cell i. A mixture model is suggested to allow different ZINB model parametrizations for high and low expression levels for gene j, in which we assume the raw expression count y_{ij} is generated one of two independent ZINB distributions with different means given the underlying binary indicator ξ_{ij},

$$y_{ij}|\xi_{ij}, \pi_i, \mu_{0j}, \phi_{0j}, \mu_{1j}, \phi_{1j} \sim \left(1-\xi_{ij}\right)\mathrm{ZINB}\left(\pi_i, s_i\mu_{0j}, \phi_{0j}\right) + \xi_{ij}\mathrm{ZINB}\left(\pi_i, s_i\mu_{1j}, \phi_{1j}\right), \tag{2}$$

where μ_{1j} and μ_{0j}, denote the group mean of read count for highly and lowly expressed genes, respectively. To guarantee that the mean expression level for a highly expressed gene is higher than a lowly expressed gene, we set a constraint for NB distribution mean across two expression levels: $\mu_{1j} > \mu_{0j}$. ϕ_{1j} and ϕ_{0j} represent the dispersion parameters of the NB model for the highly and lowly expressed gene, respectively.

To complete the model, we specify the following prior distributions: μ_{0j}, μ_{1j} ~Ga (a_μ, b_μ), s.t. $\mu_{1j} > \mu_{0j} > 0$ and ϕ_{0j}, ϕ_{1j} ~Ga (a_ϕ, b_ϕ). For prior distribution setting, small values such as $a_\mu = a_\phi = 0.01$ and $b_\mu = b_\phi = 0.01$ are recommended to impose minimal information (Gelman, 2006). To create an environment conducive to model fitting, we introduce a latent variable η_{ij} to indicate whether a zero count y_{ij} is from the zero or NB component in Eq. 1, and impose a Bernoulli prior η_{ij} ~Bern(π_i), which can be further relaxed by formulating a Beta(a_π, b_π) hyperprior on π_i, leading to a Beta-Bernoulli prior for η_{ij} with expectation $a_\pi/(a_\pi + b_\pi)$. For the

Bernoulli prior, we recommend the noninformative setting with $\pi_i = 0.5$. Similarly, in Beta-Bernoulli prior, we recommend $a_\pi = b_\pi = 1$.

2.2 A brief review of the Bayesian mark interaction model

Marked point interaction models are statistical models for spatial point pattern analysis with applications across diverse fields such as geostatistics, ecology, material physics, and so on (Edsgärd et al., 2018; Li et al., 2019). These models are designed to study the interactions among points with numerical or categorical marks in a planar region. Marked point models are receiving greater and greater focus in biology: for instance, Trendsceek applies the marked point process to identify SVGs (Edsgärd et al., 2018). The Bayesian mark interaction model, proposed by Li et al. (2019), is a full Bayesian model that characterizes spatial correlations among cell types from tumor pathology images.

Let $(t_{i1}, t_{i2}) \in \mathbb{R}^2$, $i = 1, \ldots, n$ be the x- and y-coordinates of point i. Let $G = (V, E)$ denote an interaction network with a finite set of points V and a set of direct interactions E. In the introduced Bayesian mark interaction model, we assume points have categorical marks. Here, we denote $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\top$ as the categorical marks of n points on the plane. $\xi_i \in (1, \ldots, Q)$, $Q \geq 2$ are the marks of point i.

The Bayesian mark interaction model formulates the pattern of marks $\boldsymbol{\xi}$ *via* the energy function, which is first introduced in statistical mechanics. The energy function has terms to account for both first- and second-order properties of the marked point data. Specifically, to model the interaction energy between two points, an exponential decay function with respect to the distance between the two points is used. Moreover, the Bayesian mark interaction model neglects interaction terms of point pairs from E when the corresponding distance is beyond a threshold c. Consequently, the model focuses on a sparse network $G' = (V, E')$, where E' includes edges joining pairs of points i and i' with distance $d_{ii'} < c$. The setting of the distance threshold is added to avoid the high computation cost incurred when summing over n data points and $(n-1)n/2$ interacting pairs of points with large n. Then, the potential energy of G' is measured by two addictive terms,

$$V(\boldsymbol{\xi}|\boldsymbol{\omega}, \boldsymbol{\Theta}, \lambda) = \sum_q \left(\omega_q \sum_i I(\xi_i = q) \right) + \sum_q \sum_{q'} \left(\theta_{qq'} \sum_{(i \sim i') \in E'} \exp(-\lambda d_{ii'}) I(\xi_i = q, \xi_{i'} = q') \right), \tag{3}$$

where $q, q' \in \{1, \ldots, Q\}$ are the categories of marks. $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_Q)^\top$ and $\boldsymbol{\Theta} = [\theta_{qq'}]_{Q \times Q}$ are defined as first- and second-order intensities. $(i \sim i')$ denotes the collection of interacting pairs of cells in G'. $d_{ii'} = \sqrt{(t_{i1} - t_{i'1})^2 + (t_{i2} - t_{i'2})^2}$ denotes the Euclidean distance between point i and i'. λ is the decay parameter of the distance between two points in the exponential decay function, where a larger λ makes energy diminish quickly with respect to the increase in point pair distance.

By restricting the interaction effect within radius c_d, the Bayesian mark interaction model defines a local energy. According to the fundamental Hammersley-Clifford theorem (Clifford, 1990), a probability measure with a Markov property exists if we have a locally defined energy, called a Gibbs measure. This measure gives the probability of observing categorical marks associated with their locations in a particular state. We can write the joint probability on marks $\boldsymbol{\xi}$ as,

$$\pi(\boldsymbol{\xi}|\boldsymbol{\omega}, \boldsymbol{\Theta}, \lambda) = \frac{\exp(-V(\boldsymbol{\xi}|\boldsymbol{\omega}, \boldsymbol{\Theta}, \lambda))}{\sum_{\boldsymbol{\xi}' \in \Xi} \exp(-V(\boldsymbol{\xi}'|\boldsymbol{\omega}, \boldsymbol{\Theta}, \lambda))} \tag{4}$$

which is proportional to the exponential of the negative energy of marks $\boldsymbol{\xi}$ calculated by Eq. 3. The denominator is a normalizing constant that needs to sum over the entire space Ξ of marks combination consisting of Q^n states, which is intractable even for a small size model.

The joint probability (Eq. 4) can be considered as the full data likelihood. To interpret the parameters clearly, we write the probability of observing point i having mark category q conditional on its neighborhood configuration,

$$\pi(\xi_i = q|\boldsymbol{\xi}_{-i}, \boldsymbol{\omega}, \boldsymbol{\Theta}, \lambda) \propto \exp\left(-\omega_q - \sum_{q'} \left(\theta_{qq'} \sum_{\{i': (i \sim i') \in E'\}} \exp(-\lambda d_{ii'}) I(\xi_{i'} = q') \right) \right), \tag{5}$$

where $\boldsymbol{\xi}_{-i}$ denotes the collection of all marks, except the ith one. Eq. 5 shows that the probability of point i with mark q depends on parameter ω_q, $\theta_{qq'}$, and the decay parameter λ. Parameters in Eq. 5 are interpreted below. Suppose there is no interaction between any two points in the space, i.e., $\theta_{qq'} = 0$; then, the conditional probability of point i with $\xi_i = q$ is $\pi(\xi_i = q|\cdot) \propto \exp(-\omega_q)$. Therefore, the model parameter ω_q is related to the abundance of points with mark q. Fixing $\boldsymbol{\omega}$ as equal values, we obtain the conditional probability of point i with $\xi_i = q$ is $\pi(\xi_i = q|\cdot) \propto \exp\left(-\sum_{q'}[\sum_{\{i': (i \sim i') \in E'\}} \exp(-\lambda d_{ii'}) I(\xi_{i'} = q')]\right)$. The second-order intensity $\theta_{qq'}$ quantifies the dependency of mark q with the nearby points, with mark q' scaling by the distance decay function. A detailed parameter interpretation is provided by Li et al. (2019).

2.3 A hidden mark interaction model for identifying SVGs

In Section 2.1, we describe how our ZINB mixture model is used to convert read counts $\boldsymbol{y}_{\cdot j}$ for each gene j into their corresponding hidden binary states $\boldsymbol{\xi}_{\cdot j}$. This dichotomization process allows us to represent gene expression in a binary format. We then treat the spatial distribution of cells as a two-dimensional point process, with the binary gene expressions $\boldsymbol{\xi}_{\cdot j}$ serving as the markers of these points. This setup enables us to effectively use the mark interaction model to assess the spatial correlations among these markers. In the context of SVG detection *via* the energy-based approach such as outlined by Jiang et al. (2022), the core concept involves quantifying the interactions between spots or cells of high and low expression levels for a given gene j, designated by $q = 1$ and 0, respectively. To streamline the energy function presented in Eq. 5, we only focus on the second-order intensity, θ_{12}, hereafter referred to as θ_j, while omitting self-interaction terms, θ_{11} and θ_{22}. In other words, interactions between the neighboring points with the same marks are excluded. This adjustment notably simplifies model complexity, rendering the negative energy function used in BOOST-MI a special

case of the proposed BOOST-HMI, assuming $\lambda = 0$ and c_d is chosen to match the distance between adjacent spots or cells. For simplicity, as outlined in Section 2.2, we treat the decay parameter λ as a predefined hyperparameter λ_0. Within this framework, the energy function can be expressed as follows:

$$V\left(\boldsymbol{\xi}_{\cdot j}|\omega_{0j},\omega_{1j},\theta_j\right) = \omega_{0j}\sum_i I\left(\xi_{ij}=0\right) + \omega_{1j}\sum_i I\left(\xi_{ij}=1\right) + \sum_{i\sim i'} I\left(d_{ii'}<c_d\right)\theta_j \exp\left(-\lambda_0 d_{ii'}\right) I\left(\xi_{ij}\neq\xi_{i'j}\right) \tag{6}$$

To interpret the model parameters, we provide the conditional probability of observing a high-expression level $\xi_{ij} = 1$ of gene j at cell i, given the expression levels of other cells for gene j:

$$\pi\left(\xi_{ij}=1|\boldsymbol{\xi}_{-i,j},\boldsymbol{\omega}_j,\theta_j\right) \propto \exp\left(-\omega_{1j} - \theta_j\sum_{i\sim i'} I\left(d_{ii'}<c_d\right)\exp\left(-\lambda_0 d_{ii'}\right) I\left(\xi_{i'j}=0\right)\right) \tag{7}$$

As introduced in Section 2.2, the model parameters $\boldsymbol{\omega}_j = (\omega_{0j},\omega_{1j})^\top$ represent the first-order property, and θ_j reflects the second-order property of the spatial distribution of marks. Specifically, model parameters ω_{0j} and ω_{1j} are related to the abundance of cells with low and high expression levels of gene j, respectively. From Eq. 7, when $\theta_j = 0$, it follows that $\pi(\xi_{ij}=1|\cdot) \propto \exp(-\omega_{1j})$ and $\pi(\xi_{ij}=0|\cdot) \propto \exp(-\omega_{0j})$. This means the conditional distribution of ξ_{ij} for any given cell is independent of the states of all other cells, thereby generating a completely random expression pattern indicative of a non-SVG. If the interaction parameter θ_j between highly and lowly expressed cells is positive, then Eq. 7 becomes a decreasing function with respect to the number of lowly expressed neighbors $\sum_{i\sim i'} I(\xi_{i'j}=0)$. This implies that a cell i is more likely to be in the highly expressed group when there are fewer low expressed cells in the surrounding area. In other words, a positive θ_j suggests that the gene expression level at cell i tends to be the same with the majority of its neighboring cells, leading to a repulsion pattern. Conversely, a negative θ_j indicates an attraction pattern, where cells of differing expression levels are more likely to be adjacent. Both the attraction and repulsion patterns are characteristic of SVGs. It is important to recognize that the energy functions used in BOOST-HMI and BOOST-MI differ in their signs. As a result, θ_j assumes opposite meanings between these two models. Parameter λ_0 controls the change of interaction strength between a pair of points with respect to their distance. A larger λ_0 causes the interaction between 2 cells to diminish faster, resulting in a smaller interactive neighborhood for each cell. As a hyperparameter, λ_0 needs to be set appropriately to reflect the interaction neighborhood for cells.

An identifiability problem arises when adding a nonzero constant to ω_{0j} and ω_{1j}, as it causes the joint probability $\pi(\boldsymbol{\xi}_{\cdot j}|\cdot)$ to remain invariant. Therefore, we constrain $\omega_{1j} = 1$ and establish prior distributions for ω_{0j} and θ_j to complete the parameter model settings for the hidden mark interaction model: $\omega_{0j} \sim \mathrm{N}(\mu_\omega,\tau_\omega^2)$ and $\theta_j \sim \mathrm{N}(\mu_\theta,\tau_\theta^2)$. The recommended hyperparameter setting is discussed in Section 4.1.

3 Model fitting

In this section, we introduce the MCMC algorithm for model fitting and posterior inference. Our model space consists of ($\boldsymbol{M}$, $\boldsymbol{\Phi}$, $\boldsymbol{H}$, $\boldsymbol{\Xi}$, $\boldsymbol{\omega}_0$, $\boldsymbol{\theta}$) with the underlying grouped gene expression levels $\boldsymbol{M} = [\mu_{kj}]_{2\times p}$, the dispersion parameters $\boldsymbol{\Phi} = [\phi_{kj}]_{2\times p}$, the extra zero indicators $\boldsymbol{H} = [\eta_{ij}]_{n\times p}$, the binary expression level indicators $\boldsymbol{\Xi} = [\xi_{ij}]_{n\times p}$, the first-order intensity parameter $\boldsymbol{\omega}_0 = (\omega_{01},\ldots,\omega_{0p})^\top$ and the interaction parameter $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_p)^\top$ in the mark interaction model. Each gene is examined independently by BOOST-HMI. We give the full posterior distribution for gene j as,

$$\pi\left(\boldsymbol{\mu}_{\cdot j},\boldsymbol{\phi}_j,\boldsymbol{\eta}_{\cdot j},\boldsymbol{\xi}_{\cdot j},\omega_{0j},\theta_j|\boldsymbol{y}_{\cdot j}\right) \propto \left[\prod_i\left(y_{ij}|\xi_{ij},\eta_{ij},\boldsymbol{\mu}_{\cdot j},\boldsymbol{\phi}_j\right)\right] \times \pi\left(\boldsymbol{\xi}_{\cdot j}|\omega_{0j},\omega_{1j}=1,\theta_j\right) \times \pi\left(\mu_{0j}\right)\times\pi\left(\mu_{1j}\right)\times\pi\left(\phi_{0j}\right)\times\pi\left(\phi_{1j}\right) \times \left[\prod_{i=1}^n \pi\left(\eta_{ij}\right)\right]\times\pi\left(\omega_{0j}\right)\times\pi\left(\theta_j\right). \tag{8}$$

Our primary aim was to infer ω_{0j}, θ_j and $\boldsymbol{\xi}_{\cdot j}$, which define the Gibbs probability measure based on the local energy function. We provide estimation and inference on first-order intensity ω_{0j}, which represents the abundance of lowly expressed levels of gene j, and the second-order intensity θ_j, which captures the spatial correlation between two expression levels. The estimated latent gene expression level indicator provides a robust estimation of the spatial organization of marks.

3.1 MCMC algorithms

We estimate μ_{0j}, μ_{1j}, ϕ_{0j} and ϕ_{1j} using the random walk Metropolis-Hastings (RWMH) algorithm. $\boldsymbol{\eta}_{\cdot j}$ and $\boldsymbol{\xi}_{\cdot j}$ are estimated with a Gibbs sampler. The Gibbs probability measure for the distribution of latent gene expression indicator $\boldsymbol{\xi}_{\cdot j}$ in Eq. 7 omits an intractable normalizing constant $C(\omega_{0j},\omega_{1j},\theta_j) = \sum_{\xi'_{\cdot j}} \exp(-H(\boldsymbol{\xi}'_{\cdot j}|\omega_{0j},\omega_{1j},\theta_j))$, which makes the Metropolis-Hastings algorithm infeasible. For instance, to model a gene expression profile with $n = 257$ cells, we need to traverse $2^{257} \approx 2.3\times10^{77}$ different arrangements of $\boldsymbol{\xi}$ for every gene, which is a heavy computational burden. To overcome this issue, we use the double Metropolis-Hastings (DMH) algorithm proposed by Liang et al. Liang (2010) to estimate ω_{0j} and θ_j by canceling the intractable normalizing constant. The DMH is an efficient auxiliary variable MCMC algorithm. In contrast to other auxiliary MCMC algorithms, it does not require drawing the auxiliary variables from a perfect sampler, which usually increases computational cost (Møller et al., 2006). The full details of MCMC algorithms is described in Section S1 of the Supplementary Material.

3.2 Posterior inference

Posterior inference of parameters μ_{0j}, μ_{1j}, ϕ_{0j}, ϕ_{1j}, ω_{0j}, and θ_j is obtained by averaging the MCMC posterior samples after burn-in. We are interested in identifying the SVGs by summarizing the interaction parameter $\boldsymbol{\theta}$. As stated in Section 2.3, investigating whether θ_j is positive or negative is of great importance to inferring the spatial expression pattern of gene j. To test if gene j demonstrates an attraction pattern, we applied hypothesis testing $\mathcal{M}_0$: $\theta_j \geq 0$ versus $\mathcal{M}_1$: $\theta_j < 0$. To test repulsion pattern, the

hypothesis testing is $\mathcal{M}_0$: $\theta_j \le 0$ versus $\mathcal{M}_1$: $\theta_j > 0$. If there is strong evidence to reject the null hypothesis $\mathcal{M}_0$, we conclude that gene j is an SVG. The Bayes factor (BF) is computed to infer whether θ_j is positive or negative with statistical significance from the MCMC algorithm results. The Bayes factor measures the favorability of $\mathcal{M}_1$ as

$$\mathrm{BF}_j = \frac{\pi\left(\boldsymbol{y}_{\cdot j}|\mathcal{M}_1\right)}{\pi\left(\boldsymbol{y}_{\cdot j}|\mathcal{M}_0\right)} = \frac{\pi\left(\mathcal{M}_1|\boldsymbol{y}_{\cdot j}\right)}{\pi\left(\mathcal{M}_0|\boldsymbol{y}_{\cdot j}\right)}\frac{\pi(\mathcal{M}_0)}{\pi(\mathcal{M}_1)} \approx \begin{cases} \dfrac{\sum_u I\left(\theta_j^{(u)} < 0|\cdot\right)}{\sum_u I\left(\theta_j^{(u)} \ge 0|\cdot\right)}, & \text{for attraction,} \\ \dfrac{\sum_u I\left(\theta_j^{(u)} > 0|\cdot\right)}{\sum_u I\left(\theta_j^{(u)} \le 0|\cdot\right)}, & \text{for repulsion,} \end{cases} \tag{9}$$

where u indexes the iteration and U is the total number of iterations after burn-in. The larger the BF_j, the more likely gene j is an SVG with an attraction pattern. The smaller the BF_j, the more likely gene j is an SVG with a repulsion pattern.

Another important parameter in our model is the latent gene expression level indicator $\boldsymbol{\xi}_{\cdot j}$. We summarize the posterior distribution of $\boldsymbol{\xi}_{\cdot j}$ *via maximum-a-posteriori* (MAP) estimates, which is the mode of the posterior distribution. A more comprehensive summary of $\boldsymbol{\xi}_{\cdot j}$'s is based on their marginal posterior probabilities of inclusion (PPI), where $\mathrm{PPI}_{ij} = \sum_{u=1}^{U} \xi_{ij}^{(u)}/U$. Then, the latent expression indicator indicates a high expression spot when PPIs are greater than a threshold c_p:

$$\boldsymbol{\xi}_{\cdot j}^{\mathrm{PPI}} = \left(I\left(\mathrm{PPI}_{1j} \ge c_p\right), \ldots, I\left(\mathrm{PPI}_{nj} \ge c_p\right)\right)^{\top}.$$

4 Results

4.1 Simulation study

We generated simulated data to evaluate the ability of BOOST-HMI to identify SVGs and provided a comparison with five competing methods: SpatialDE (Svensson et al., 2018), SPARK (Sun et al., 2020), SPARK-X (Zhu et al., 2021), BinSpect (Dries et al., 2021), and BOOST-GP (Li et al., 2021).

Spatial locations of simulated data were from the geospatial profile of the mouse hippocampus dataset field 43 (Shah et al., 2016) with $n = 257$ cells, which we present in Section 4.2. To generate the expression counts for gene j, the latent gene expression level indicators $\boldsymbol{\xi}_{\cdot j}$'s were first generated based on Eq. 7 with three different values of $\omega_{0j} \in \{1.4, 1, 0.6\}$ and the fixed value of $\omega_{1j} = 1$. These three values of $\boldsymbol{\omega}_j$ correspond to approximately 60%, 50%, and 40% lowly-expressed cells in $\boldsymbol{\xi}_{\cdot j}$. Additionally, we set four different values of $\theta_j \in \{-2.5, -1.2, 1.9, 3.2\}$ to generate SVGs with various patterns of attraction or repulsion. These values correspond to strong attraction, weak attraction, weak repulsion, and strong repulsion patterns, respectively. For the non-SVG, $\theta_j = 0$ which indicates complete randomness and no spatial correlation. The distance threshold c_d and decay parameter λ_0 were set as 0.15 and 20, respectively. We then simulated gene expression data from a ZINB model with three different group-mean ratios $R \in \{2, 5, 10\}$ between high and low expression:

$$y_{ij}|\xi_{ij}, \eta_{ij}, \mu_{0j}, \mu_{1j}, \phi_{0j}, \phi_{1j} \sim \begin{cases} \mathrm{NB}\left(y_{ij}; s_i\beta_0 r, \phi_{1j}\right)^{\xi_{ij}} \mathrm{NB}\left(y_{ij}; s_i\beta_0, \phi_{0j}\right)^{1-\xi_{ij}}, & \eta_{ij} = 0 \\ I\left(y_{ij} = 0\right) & \eta_{ij} = 1 \end{cases},$$

where the underlying baseline expression levels $\beta_0 = 10$. In the simulation study, size factors $\boldsymbol{s} = (s_1, \ldots, s_n)^{\top}$ were generated from log-N $(0, 0.2^2)$, and dispersion parameters ϕ_{0j}, ϕ_{1j} in the NB model are generated from an exponential distribution Exp (1/10). Further, to imitate high sparsity and account for medium sparsity in real SRT data, we created three sets of sparsity levels, 0%–10%, 10%–20%, and 30%–40%, and generated extra zero parameters $\boldsymbol{\eta}_{\cdot j}$'s correspondingly. Extra zeros were randomly selected and imputed into the generated gene expression count data. Thus, we considered three group-mean ratios and three sparsity levels, which is $3 \times 3 = 9$ scenarios in total. For each scenario, we simulated 30 replicates with $p = 100$ genes in each replicate, 10 out of which were SVGs.

Before estimating the parameters using BOOST-HMI, we specified the prior distributions. Non-informative gamma priors were specified for μ_{0j}, μ_{1j}, ϕ_{0j} and ϕ_{1j}, i.e., $\mu_{0j} \sim \mathrm{Ga}(a_\mu, b_\mu)$ and $\phi_{0j} \sim \mathrm{Ga}(a_\phi, b_\phi)$. We set a_μ, b_μ, a_ϕ, and b_ϕ to 0.01, which produced a gamma distribution with mean one and variance 100. Priors for ω_{0j} and θ_j were set to control the gene expression abundance and gene expression pattern, $\omega_{0j} \sim \mathrm{N}(1, \tau_\omega^2)$ and $\theta_j \sim \mathrm{N}(0, \tau_\theta^2)$. In the simulation study and real data analysis, we set $\tau_\omega = 0.5$, where the prior distribution of ω_{0j} indicates that the latent proportion of low expression cells ranges from 1% to 100% with a probability of 95%. τ_θ was set to 3.5 such that the prior for θ_j guarantees that θ_j falls within −6 to 6 with a probability of 92%. For hyperparameters in the energy function, we set the distance threshold $c_d = 0.15$ and expected the relationship of decay parameter λ_0 and c_d to be $\exp(-\lambda_0 c_d) = 0.05$, which specifies the range of exponential decay function $\exp(-\lambda_0 d_{ii'})$ to be [0.05, 1] as $c_d \ge d_{ii'} \ge 0$; therefore, we set λ_0 as 20 correspondingly. As for the setting of the MCMC algorithm, we implemented BOOST-HMI in a gene-wise fashion. For each gene, we initialized model parameters by randomly drawing from their prior distributions. The MCMC algorithm is iterated $U = 10,000$ times after 10,000-iteration burn-in. The algorithm was implemented in `R` and `Rcpp`. As mentioned in Section 3.2, BOOST-HMI identifies SVGs based on Bayes Factors (BFs). In our study, we select a BF threshold of 10, which indicates strong evidence in favor of the $\mathcal{M}_1$ (Kass and Raftery, 1995).

We implemented the other five competing methods with their default settings. BOOST-GP selects SVGs using Bayes factors. SpatialDE, BinSpect, SPARK, and SPARK-X, use p-values to select SVGs. To control type-I error rate, the Benjamini–Hochberg (BH) (Benjamini and Hochberg, 1995) procedure was used to adjust p-values from SpatialDE and BinSpect. We specifically avoided adjusting p-values from SPARK and SPARK-X since its raw p-values are calibrated by the Cauchy combination rule (Liu et al., 2019; Sun et al., 2020). For p-values, the threshold was set to 0.05.

Our task is to evaluate the ability of each method to correctly identify underlying SVGs from the simulated dataset, which can be defined as a binary classification problem; therefore, to evaluate the performance of the five methods, we employed two performance

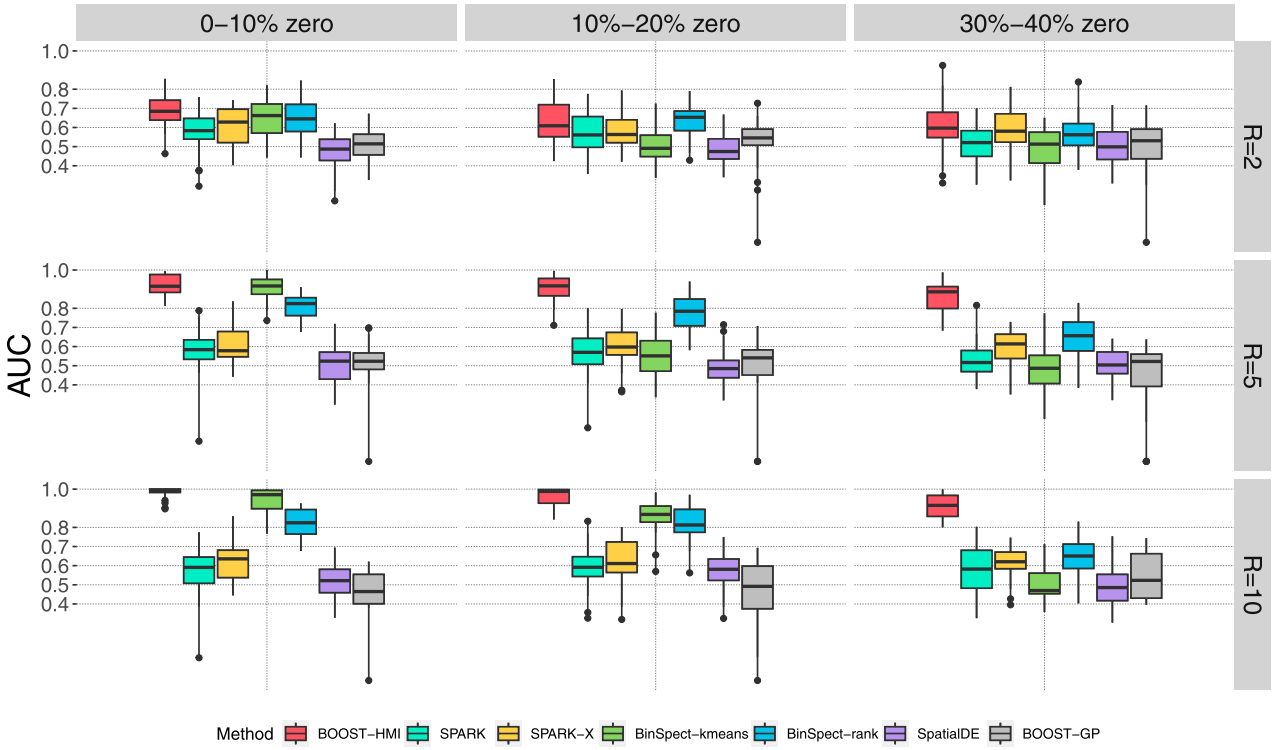

FIGURE 2
Simulation study: The boxplots of AUCs achieved by BOOST-HMI, SPARK, SPARK-X, SpatialDE, BinSpect-kmeans, BinSpect-rank, and BOOST-GP across nine scenarios.

metrics for binary classification problems: First, we used the area under the curve (AUC) (Bradley, 1997) of the receiver operating characteristic (ROC) (Fukunaga, 2013). The ROC is a plot of the true positive rate against the false positive rate for different classification thresholds. The AUC is a single value ranging from 0 to 1, with a higher value indicating better classification performance.

Figure 2 displays a boxplot of AUCs calculated by the aforementioned seven methods over 30 replicates across nine scenarios. It clearly suggests that BOOST-HMI achieves superior performance compared to the other methods, especially when there was high sparsity. BinSpect-kmeans and BinSpect-rank showed competitive performance when there was no zero-inflation, i.e., when the sparsity level was between 0% and 10%, regardless of the different group ratios; however, these methods showed decreasing AUCs as sparsity level increased. SPARK and SpatialDE suffered from a limited ability to detect SVGs from low expression variability or high zero-inflation scenarios. Between SPARK and SpatialDE, the simulation study showed that SPARK has better SVG detection power over SpatialDE, which is consistent with the conclusion drawn by Sun et al. (2020) and Jiang et al. (2022). In summary, BOOST-HMI achieved satisfactory performance and is robust against different group expression level ratios and sparsity levels.

The second metric we used is the Matthews correlation coefficient (MCC). MCC is a summary value that examines the binary classification performance under a specific cutoff, i.e., BF or p-value thresholds for our study. It has values ranging from −1 to 1, incorporating true positives, true negatives, false positives and false negatives. A larger MCC value, such as 1, corresponds to an excellent classifier, while a negative MCC indicates a strong disagreement between prediction and observation. Table 1 summarizes the average MCCs obtained in the simulation study across the five methods. The result is consistent with our conclusion from our analysis of the AUC: BOOST-HMI achieved the highest power under high zero-inflation, while all other six methods suffered from the high number of false zeros. In the scenario without zero inflation, BinSpect-kmeans stood out, and BinSpect-rank, SPARK, and BOOST-HMI showed competitive performance in identifying SVGs.

We further expanded our evaluation of BOOST-HMI's effectiveness through comprehensive analyses using simulated data. These investigations, detailed in the Supplementary Material, encompass a scalability test (Section S2), a sensitivity analysis (Section S3), an examination of performance under model mis-specification (Section S4), and an assessment of statistical power and false discovery control (Section S5).

4.2 Application to the mouse hippocampus seqFISH data

The mouse hippocampus dataset is a public seqFISH dataset with 21 field replicates collected on a third coronal section (Shah et al., 2016). Following SpatialDE and SPARK protocols, we analyzed the field 43 dataset, which contains p = 249 genes measured on 257 cells with spatial location preserved. Out of

TABLE 1 Simulation study: The averaged MCCs, with standard deviations in parentheses, achieved by BOOST-HMI, BinSpect-kmeans, BinSpect-rank, SPARK, and SpatialDE across nine scenarios.

	0%–10% zeros		
	$R = 2$	$R = 5$	$R = 10$
BinSpect-kmeans	0.182 (0.159)	0.653 (0.118)	0.714 (0.085)
BinSpect-rank	0.199 (0.193)	0.476 (0.123)	0.488 (0.139)
SPARK	0.255 (0.162)	0.502 (0.164)	0.524 (0.153)
SPARK-X	0.324 (0.153)	0.279 (0.152)	0.324 (0.153)
SpatialDE	0.009 (0.111)	−0.009 (0.063)	0.027 (0.109)
BOOST-GP	0.012 (0.060)	0.012 (0.060)	0.000 (0.000)
BOOST-HMI	0.231 (0.141)	0.449 (0.084)	0.510 (0.063)
10%–20% zeros			
	$R = 2$	$R = 5$	$R = 10$
BinSpect-kmeans	0.007 (0.092)	0.121 (0.134)	0.601 (0.142)
BinSpect-rank	0.173 (0.166)	0.482 (0.120)	0.537 (0.153)
SPARK	0.113 (0.146)	0.271 (0.153)	0.373 (0.123)
SPARK-X	0.284 (0.144)	0.224 (0.137)	0.284 (0.144)
SpatialDE	−0.039 (0.043)	0.005 (0.093)	0.031 (0.120)
BOOST-GP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
BOOST-HMI	0.160 (0.176)	0.433 (0.091)	0.473 (0.065)
30%–40% zeros			
	$R = 2$	$R = 5$	$R = 10$
BinSpect-kmeans	0.002 (0.086)	0.005 (0.083)	−0.006 (0.105)
BinSpect-rank	0.060 (0.139)	0.243 (0.157)	0.226 (0.142)
SPARK	0.014 (0.103)	0.133 (0.150)	0.144 (0.155)
SPARK-X	0.114 (0.134)	0.182 (0.136)	0.186 (0.131)
SpatialDE	−0.016 (0.071)	−0.001 (0.097)	0.016 (0.110)
BOOST-GP	−0.001 (0.007)	0.012 (0.060)	0.000 (0.000)
BOOST-HMI	0.098 (0.158)	0.406 (0.088)	0.427 (0.088)

249 genes, 214 were selected from a list of transcription factors and signaling pathway components, and the remaining 35 were selected from cell identity markers. Quality control was performed following SPARK protocol (Sun et al., 2020) and the original study. We filtered out cells with x- or y-axis values exceeding 203–822 pixels to tackle border artifacts. After filtering, $n = 131$ cells were included for the following analysis. We excluded SpatialDE due to its unsatisfactory performance in SVG detection in the simulation study. Prior settings, MCMC algorithm implementation, and significance criteria were identical to what was described in Section 4.1. Five independent MCMC chains were sampled and diagnosed for algorithm convergence. Algorithm convergence was checked based on the BF vector. BFs from five chains were highly correlated with Pearson correlation coefficients ranging from 0.90 to 0.99. We further checked algorithm convergence using the potential scale reduction factor (PSRF) (Gelman and Rubin, 1992; Brooks and Gelman, 1998) on posterior samples of θ_j's and ω_{0j}'s. If multiple chains converge to the target posterior distribution, the PSRF will be close to one. In our analysis, the PSRFs were below 1.2, suggesting convergence of the MCMC algorithms. Posterior samples obtained from the quintuplet of MCMC chains were amalgamated for subsequent analysis. Concerning efficiency, we report the execution times of our method compared to others in Supplementary Table S2 of the Supplementary Material.

As detailed in Section 2.1, the variance of a NB distribution is given by $\nu + \nu^2/\phi$, with ν and ϕ representing the mean and dispersion parameters, respectively. A low ϕ value suggests significant over-dispersion, whereas $\phi \to \infty$ implies that the mean and variance are equal. This relationship allows us to deduce over-dispersion from the posterior distribution of ϕ_j. Furthermore, we introduced a latent binary variable η_{ij} to distinguish whether a zero count y_{ij} originates from the zero or NB component. The posterior probability of η_{ij} enables us to identify zero-inflation. In our analysis, we found that the average posterior mean for ϕ_j was 14.237, with the ϕ_j's for 95% of the genes ranging from 1.577 to 89.164. This evidence strongly supports the existence of over-dispersion. Moreover, the average posterior mean of η_{ij} for zero counts was 0.9425. This indicates that approximately 94.25% of the zeros are attributable to the zero component, thereby underscoring the presence of zero-inflation. Among the $p = 249$ genes analyzed in the mouse hippocampus dataset, SPARK identified 17 SVGs, while BOOST-HMI detected 22 SVGs. Notably, BOOST-HMI successfully detected 16 cell identity markers previously presented by Shah et al. (2016), whereas SPARK identified 14 markers. In comparison, BinSpect-kmeans and BinSpect-rank were more aggressive, respectively identifying 38 and 44 SVGs. A Venn diagram in Figure 3D showcases the overlap of SVGs identified by the four methods. Among them, BOOST-HMI and SPARK shared 12 SVGs in common. Only one SVG detected by BinSpect-kmeans overlapped with that from BOOST-HMI, and none overlapped with that from SPARK. None of the SVGs detected by BinSpect-rank were detected by either SPARK or BOOST-HMI. We further visualized the spatial pattern for each SVG using the marginal PPI of the posterior samples of the hidden gene expression indicator $\boldsymbol{\xi}_{\cdot j}$. Supplementary Figure S12 visualizes the posterior distributions of θ_j of those identified SVGs by BOOST-HMI or SPARK. Supplementary Figure S13 displays the relative gene expression levels for each SVG. Figure 3A displays the spatial patterns of SVGs detected by SPARK and BOOST-HMI, while Supplementary Figures S6, S7 in the Supplementary Material depict the spatial patterns for SVGs detected by BinSpect-kmeans and BinSpect-rank, respectively. Among the 12 common SVGs identified by SPARK and BOOST-HMI, strong spatial repulsion patterns between high- and low expression genes are evident across 131 cells. Notably, a larger Bayes factor (shown in parentheses) indicates a stronger spatial pattern; genes *Foxo1*, *sst*, *mog*, *myl14*, and *ndnf* exhibited clear spatial patterns between polarized estimated hidden indicators. SPARK detected five unique SVGs, as Figure 3B shows, for which the PPIs of the estimated hidden expression indicators are close to 0.5. BOOST-HMI identified ten unique SVGs, as displayed in Figure 3C. Among these, seven genes, such as gene *Zfp423*, *slc5a7* and *palvb*, demonstrated either high or low

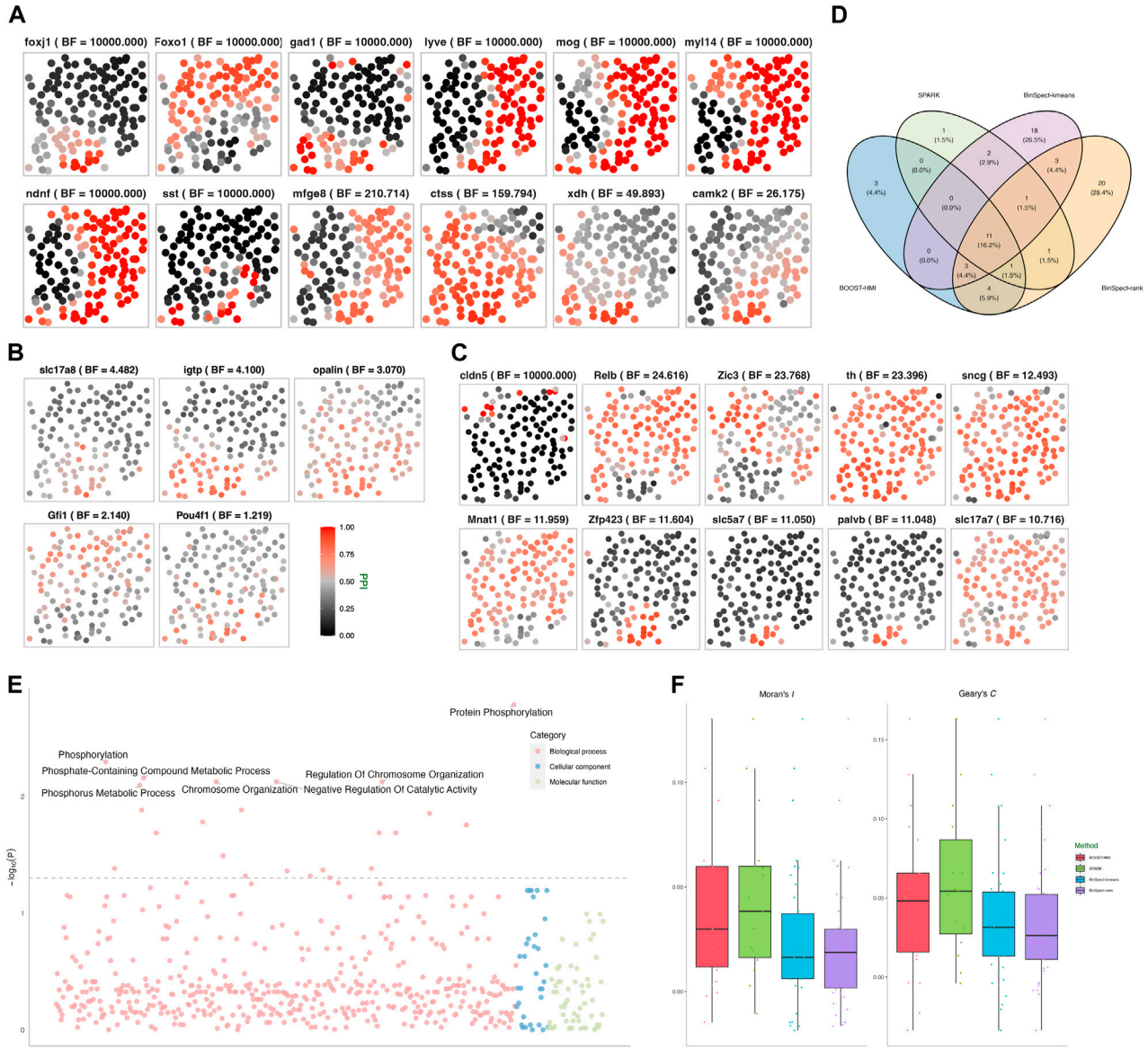

FIGURE 3
Mouse hippocampus seqFISH data: **(A)** Spatial distribution of hidden gene expressions ξ^{MAP} of the 12 SVGs detected by both SPARK and BOOST-HMI. **(B)** Spatial distribution of hidden gene expressions ξ^{MAP} of the five SVGs detected by SPARK only. **(C)** Spatial distribution of hidden gene expressions ξ^{MAP} of the ten SVGs detected by BOOST-HMI only. **(D)** Venn diagram of the overlap across SVGs identified by all four methods. **(E)** Enriched GO terms associated with SVGs detected by BOOST-HMI. **(F)** Boxplot of Moran's *I* and Geary's *C* values for SVGs across the four methods.

expression in the majority of cells. The remaining three unique SVGs, *Zic3*, *Mnat1*, and *slc17a7* delineated three distinct patterns, which may be related to novel biological mechanisms. Lim et al. (2007) previously highlighted the crucial role of *Zic3* in preserving pluripotency in embryonic stem cells, while Herman and El-Hodiri (2002) demonstrated that mutations in *Zic3* are linked to developmental abnormalities such as laterality defects, congenital heart disease, and neural tube defects.

In addition to visualizing spatial patterns, we quantified the degree of spatial attraction or repulsion pattern in gene expression across different cells using the spatial autocorrelation tests Moran's *I* and Geary's *C*. Moran's *I* quantifies the spatial clustering or dispersion by standardizing the spatial autocovariance, yielding a correlation coefficient ranging from −1 to 1. A positive Moran's *I* value corresponds to a spatial clustering pattern where the variable tends to have similar values to its neighboring cells. A Moran's *I* value close to 0 suggests a random spatial distribution of the data, while a negative value corresponds to a dispersion pattern, where the variable value tends to be dissimilar from its neighbors. To assess the spatial patterns exhibited by the SVGs, Moran's *I* and Geary's *C* values were calculated for each SVG identified by at least one of the four methods. Moran's *I* for each gene *j* was calculated by the following formula:

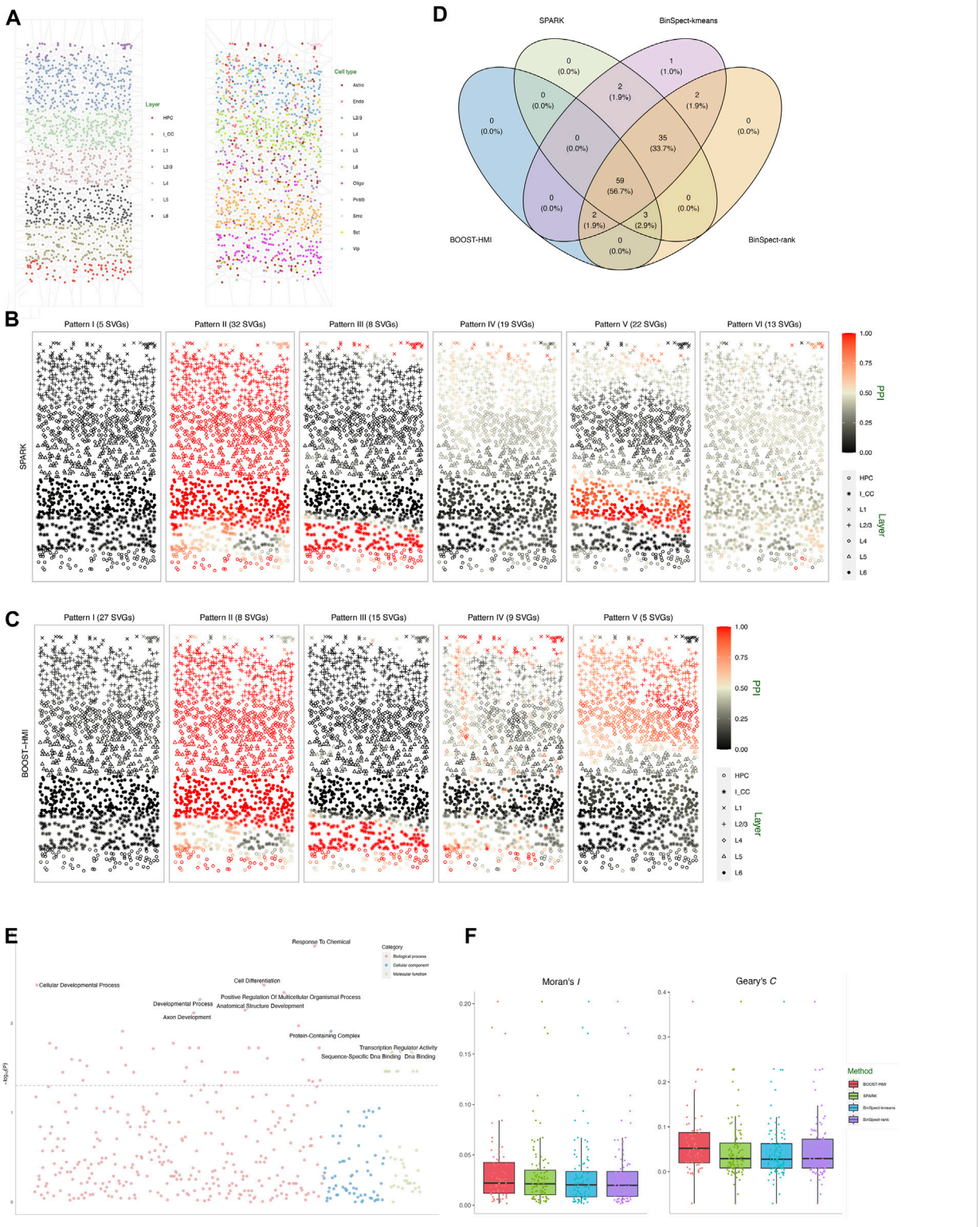

FIGURE 4
Mouse visual cortex STARmap data: **(A)** Voronoi diagrams of layer structures and cell type distribution. **(B)** Spatial distribution of the average hidden gene expressions ξ^{MAP} of the six SVG patterns detected by SPARK. **(C)** Spatial distribution of the average hidden gene expressions ξ^{MAP} of the five SVG patterns detected by BOOST-HMI. **(D)** Venn diagram of the overlap across SVGs identified by all four methods. **(E)** Enriched GO terms associated with SVGs detected by BOOST-HMI. **(F)** Boxplot of Moran's *I* and Geary's *C* values for SVGs across the four methods.

$$\text{Moran's } I = \frac{n}{\sum_i \sum_h w_{ih}} \frac{\sum_i \sum_h w_{ih}\left(y_{ij} - \bar{y}_j\right)\left(y_{hj} - \bar{y}_j\right)}{\sum_i \left(y_{ij} - \bar{y}_j\right)^2},$$

where $w_{ij} = A/(d_{ih})^m$ is the connectivity spatial weight between cell i and h. Spatial weight is a decay factor of the distance between 2 cells; in our study, we set A = 1, m =

1. y_{ij} and y_{hj} are the gene expression count of cell i and cell h, and $\bar{y}_j$ is the mean expression of gene j. Similar to Moran's I, Geary's C measures the spatial similarity or dissimilarity between neighboring cells, and is calculated with the following formula:

$$\text{Geary}'\text{s } C = \frac{n-1}{2\sum_i\sum_h w_{ih}} \frac{\sum_i\sum_h w_{ih}\left(y_{ij}-y_{hj}\right)^2}{\sum_i\left(y_{ij}-\bar{y}_j\right)^2}.$$

Geary's C ranges from 0 to 2, where a value close to 0 indicates a spatial attraction pattern, one corresponds to complete randomness, and two implies a spatial repulsion pattern. To ensure uniform interpretation of Moran's I and Geary's C, following Hu et al. (2021), we scaled Geary's C to the range [−1, 1]. The distributions of these values are depicted in Figure 3F. Remarkably, over 75% of Moran's I and Geary's C values were positive, compellingly indicating the presence of spatial patterns associated with the SVGs across the four methods implemented. Moreover, SVGs from BOOST-HMI and SPARK exhibited the highest Moran's I values, while SVGs from SPARK demonstrated the highest Geary's C values.

To explore the relevant biological functions of identified SVGs, we conducted gene ontology (GO) enrichment analysis using the `R` package `clusterProfiler` (Yu et al., 2012). As mentioned, 214 genes were selected from a list of transcription factors and signaling pathway components. As a result, genes in the background set enriched 4,622 GO terms and 10,285 relations. Figure 3E depicts the biological processes enriched by SVGs that were detected by BOOST-HMI, such as a smoothened signaling pathway (GO: 0007224), regulation of neural precursor cell proliferation (GO: 20000177), and cellular response to stress (GO: 0033554). Moreover, gene *Foxo1* enriched three significant GO terms, which may inspire further research work on *Foxo1* regulation in the mouse hippocampus. *Mnat1*, one of the SVGs, enriched cellular response to stress. Several studies have found that *Mnat1* is associated with various disease progression and regulation. Qiu et al. (2020) found that *Mnat1*, which was detected only by BOOST-HMI, contributes to the progression of osteosarcoma, and Zou et al. (2020) reported that decreased *Mnat1* expression induces degradation of an important regulator of necroptosis in endothelial cells from samples with Alzheimer's disease.

4.3 Application to the mouse visual cortex STARmap data

The second real dataset we analyzed is a STARmap dataset, which profiles the mouse visual cortex from the hippocampus to the corpus callosum, spanning six neocortical layers at single-cell resolution (Wang et al., 2018). The STARmap dataset measures the expression of 1,020 genes in 1,549 cells, including non-neuron cells such as endothelial, oligodendrocytes, astrocytes, and neuron cells, i.e., parvalbumin-expressing, vasoactive intestinal peptide-expressing, and somatostatin-expressing interneurons. Figure 4A depicts the layer structure and distribution of cell types within the tissue section as presented in the original study (Wang et al., 2018). Moreover, the STARmap dataset is highly sparse with nearly 79% zero counts. To address potential sources of variability, we performed three quality control steps: 1) cells with fewer than 100 read counts detected were filtered out; 2) genes with more than 90% zero counts were filtered out; 3) genes whose maximum count is smaller than ten were removed. After quality control, the gene expression profile measured the expression of $p = 107$ genes in $n = 1,523$ cells.

SPARK, BinSpect-kmeans, and BinSpect-rank were implemented with the same parameter settings as the simulation study. As for BOOST-HMI, we set the distance threshold $c_d = 0.05$ and the decay parameter $\lambda_0 = 60$ to satisfy the dependency $\exp(-\lambda_0 c_d) = 0.05$. We ran four independent MCMC chains with the same prior specifications and parameter settings as the simulation study, and made posterior inferences after integrating the posterior samples across the four chains. Concerning efficiency, we report the execution times of our method compared to others in Supplementary Table S2 of the Supplementary Material.

As Figure 4D shows, SPARK detected 99 SVGs, while BOOST-HMI detected 64 SVGs. Both BinSpect-kmeans and BinSpect-rank detected 101 SVGs. SPARK, BinSpect-kmeans, and BinSpect-rank detected 94 SVGs in common. Compared to other methods, BOOST-HMI was conservative, detecting 59 common SVGs with the other three methods. To further investigate the spatial patterns of detected SVGs, we visualized the estimated hidden gene expression indicator for each SVG, annotated with the corresponding Bayes factor, in Supplementary Figures S8–S11 of the Supplementary Material. Supplementary Figure S14 displays the average relative gene expression levels across SPARK and BOOST-HMI-identified SVGs in each cluster. Analysis of the detected genes reveals a noteworthy observation: all four methods identify SVGs associated with layer structures, such as *Apod*, *Apoe*, and *Egr1*, which exhibit high expression in layer L6 and HPC. In contrast, SVGs exclusively detected by the other three methods either lack a clear layer structure or are estimated to be lowly expressed across the entire tissue section, which suggests that BOOST-HMI can detect SVGs with clear spatial patterns and address potential falsification. This conclusion is strongly supported by corroborating evidence from both calculations of Moran's I and Geary's C. Figure 4F demonstrates that SVGs detected by BOOST-HMI show stronger spatial autocorrelation than those identified by the other three methods. To delve deeper into the identified spatial patterns, we performed agglomerative hierarchical clustering on the SVGs detected by SPARK and BOOST-HMI, as shown in Figure 4C. SVGs detected by SPARK were grouped into six clustered, while those detected by BOOST-HMI formed five clusters. Pattern I, II, III and IV from SPARK and BOOST-HMI demonstrate a similar spatial pattern. Spatial pattern V from BOOST-HMI delineates layers L1, L2/3, and L4, while pattern V from SPARK is associated with layer L6. Pattern VI from SPARK highlights a fraction of cells in layer L1.

To gain insights into biological processes, molecular function, and cellular components, GO enrichment analysis was performed on SVGs identified by BOOST-HMI. Figure 4E shows that SVGs detected by BOOST-HMI are implicated in biological processes such as the cellular developmental process and anatomical structure development. Additionally, these SVGs are associated with cellular components such as protein-containing complexes as well as molecular functions such as DNA binding. Notably, the gene *Bcl6* significantly enriched the cellular developmental process. Nurieva et al. (2009) found that *Bcl6* functions as a regulator of T follicular helper cell differentiation and B cell-mediated immunity. Our findings have the potential to inspire further novel biological insights.

5 Conclusion

This paper introduces BOOST-HMI, a novel method for identifying SVGs in imaging-based SRT datasets. By integrating gene expression data with spatial location, BOOST-HMI employs a ZINB mixture model to effectively handle the excessive zeros typical in SRT data. Additionally, it uses a hidden Bayesian mark interaction model to accurately quantify spatial dependencies in gene expressions.

Our approach is adaptable for analyzing sequencing-based SRT data. We validated BOOST-HMI through a simulation study and analysis of two real datasets, demonstrating its effectiveness across various SRT technologies and tissue sections. The simulation results showed that BOOST-HMI is particularly adept at identifying SVGs in data with high sparsity levels, between 30% and 40%. When analyzing the mouse hippocampus seqFISH data, BOOST-HMI achieved comparable results to SPARK, with the identified SVGs exhibiting stronger spatial patterns as quantified by Moran′s*I* and Geary′s*C*. Moreover, the SVGs identified were enriched in biologically relevant GO terms, such as smoothened signaling pathways and regulation of neural precursor cell proliferation, offering avenues for further biological investigation.

Further analysis of the mouse visual cortex STARmap dataset revealed that BOOST-HMI can identify SVGs with spatial patterns aligning with the underlying cell structure of the tissue. Additionally, GO enrichment analysis indicated that these SVGs are linked to cellular developmental processes, underscoring the potential for novel biological insights.

While our method assumes homogeneity in spatial patterns across tissue sections, this may not hold true for all cases. Future work will aim to generalize BOOST-HMI to accommodate heterogeneous spatial patterns, enhancing its practicality. Another focus will be on scaling the model to accommodate the growing size of SRT datasets, such as those generated by advanced technologies like Slide-seqV2, which can resolve over 19,600 genes from around 23,000 cells (Stickels et al., 2021). Enhanced scalability will enable BOOST-HMI to analyze datasets from various technologies like HDST, Slide-seqV2, and others, potentially leading to more groundbreaking biological discoveries.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

JY: Data curation, Investigation, Methodology, Software, Visualization, Writing–original draft, Writing–review and editing. XJ: Conceptualization, Data curation, Investigation, Methodology, Writing–original draft, Writing–review and editing. KJ: Writing–review and editing. SS: Supervision, Writing–review and editing. QL: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Science Foundation (2113674, 2210912) and the National Institutes of Health (1R01GM141519). SS was supported in part by Korean NRF grant funded by the Korea government (MSIT) (RS-2023-00243012, RS-2023-00219980), POSTECH Basic Science Research Institute Fund (Korean NRF grant 2021R1A6A1A10042944), and POSCO HOLDINGS grant 2023Q033.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1356709/full#supplementary-material

References

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* 57, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* 30, 1145–1159. doi:10.1016/s0031-3203(96)00142-2

Brooks, S. P., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statistics* 7, 434–455. doi:10.2307/1390675

Clifford, P. (1990). Markov random fields in statistics. *Disord. Phys. Syst. A volume honour John M. Hammersley*, 19–32.

Dries, R., Zhu, Q., Dong, R., Eng, C.-H. L., Li, H., Liu, K., et al. (2021). Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol.* 22, 78–31. doi:10.1186/s13059-021-02286-2

Edsgärd, D., Johnsson, P., and Sandberg, R. (2018). Identification of spatial expression trends in single-cell gene expression data. *Nat. Methods* 15, 339–342. doi:10.1038/nmeth.4634

Efremova, M., Vento-Tormo, M., Teichmann, S. A., and Vento-Tormo, R. (2020). CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc.* 15, 1484–1506. doi:10.1038/s41596-020-0292-x

Eng, C.-H. L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., et al. (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* 568, 235–239. doi:10.1038/s41586-019-1049-y

Fukunaga, K. (2013). *Introduction to statistical pattern recognition*. Elsevier.

Gelman, A. (2006). *Prior distributions for variance parameters in hierarchical models.* (comment on article by browne and draper).

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–472. doi:10.1214/ss/1177011136

Herman, G., and El-Hodiri, H. (2002). The role of ZIC3 in vertebrate development. *Cytogenet. genome Res.* 99, 229–235. doi:10.1159/000071598

Hu, J., Li, X., Coleman, K., Schroeder, A., Ma, N., Irwin, D. J., et al. (2021). SpaGCN: integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods* 18, 1342–1351. doi:10.1038/s41592-021-01255-8

Jiang, X., Xiao, G., and Li, Q. (2022). A Bayesian modified ising model for identifying spatially variable genes from spatial transcriptomics data. *Statistics Med.* 41, 4647–4665. doi:10.1002/sim.9530

Jin, S., Guerrero-Juarez, C. F., Zhang, L., Chang, I., Ramos, R., Kuan, C.-H., et al. (2021). Inference and analysis of cell-cell communication using CellChat. *Nat. Commun.* 12, 1088–1120. doi:10.1038/s41467-021-21246-9

Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *J. Am. Stat. Assoc.* 90, 773–795. doi:10.2307/2291091

Li, Q., Wang, X., Liang, F., and Xiao, G. (2019). A Bayesian mark interaction model for analysis of tumor pathology images. *Ann. Appl. Statistics* 13, 1708–1732. doi:10.1214/19-AOAS1254

Li, Q., Zhang, M., Xie, Y., and Xiao, G. (2021). Bayesian modeling of spatial molecular profiling data via Gaussian process. *Bioinformatics* 37, 4129–4136. doi:10.1093/bioinformatics/btab455

Liang, F. (2010). A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. *J. Stat. Comput. Simul.* 80, 1007–1022. doi:10.1080/00949650902882162

Liao, J., Lu, X., Shao, X., Zhu, L., and Fan, X. (2021). Uncovering an organ's molecular architecture at single-cell resolution by spatially resolved transcriptomics. *Trends Biotechnol.* 39, 43–58. doi:10.1016/j.tibtech.2020.05.006

Lim, L. S., Loh, Y.-H., Zhang, W., Li, Y., Chen, X., Wang, Y., et al. (2007). Zic3 is required for maintenance of pluripotency in embryonic stem cells. *Mol. Biol. Cell* 18, 1348–1358. doi:10.1091/mbc.e06-07-0624

Liu, Y., Chen, S., Li, Z., Morrison, A. C., Boerwinkle, E., and Lin, X. (2019). ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. *Am. J. Hum. Genet.* 104, 410–421. doi:10.1016/j.ajhg.2019.01.002

Moffitt, J. R., Bambah-Mukku, D., Eichhorn, S. W., Vaughn, E., Shekhar, K., Perez, J. D., et al. (2018). Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* 362, eaau5324. doi:10.1126/science.aau5324

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika* 93, 451–458. doi:10.1093/biomet/93.2.451

Nurieva, R. I., Chung, Y., Martinez, G. J., Yang, X. O., Tanaka, S., Matskevitch, T. D., et al. (2009). Bcl6 mediates the development of T follicular helper cells. *Science* 325, 1001–1005. doi:10.1126/science.1176676

Qiu, C., Su, W., Shen, N., Qi, X., Wu, X., Wang, K., et al. (2020). MNAT1 promotes proliferation and the chemo-resistance of osteosarcoma cell to cisplatin through regulating PI3K/Akt/mTOR pathway. *BMC Cancer* 20, 1187–1212. doi:10.1186/s12885-020-07687-3

Shah, S., Lubeck, E., Zhou, W., and Cai, L. (2016). *In situ* transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* 92, 342–357. doi:10.1016/j.neuron.2016.10.001

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 353, 78–82. doi:10.1126/science.aaf2403

Stickels, R. R., Murray, E., Kumar, P., Li, J., Marshall, J. L., Di Bella, D. J., et al. (2021). Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* 39, 313–319. doi:10.1038/s41587-020-0739-1

Sun, S., Zhu, J., and Zhou, X. (2020). Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat. Methods* 17, 193–200. doi:10.1038/s41592-019-0701-7

Svensson, V., Teichmann, S. A., and Stegle, O. (2018). SpatialDE: identification of spatially variable genes. *Nat. Methods* 15, 343–346. doi:10.1038/nmeth.4636

Wang, X., Allen, W. E., Wright, M. A., Sylwestrak, E. L., Samusik, N., Vesuna, S., et al. (2018). Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* 361, eaat5691. doi:10.1126/science.aat5691

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics A J. Integr. Biol.* 16, 284–287. doi:10.1089/omi.2011.0118

Zhao, P., Zhu, J., Ma, Y., and Zhou, X. (2022). Modeling zero inflation is not necessary for spatial transcriptomics. *Genome Biol.* 23, 118. doi:10.1186/s13059-022-02684-0

Zhu, J., Sun, S., and Zhou, X. (2021). SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biol.* 22, 184. doi:10.1186/s13059-021-02404-0

Zou, C., Mifflin, L., Hu, Z., Zhang, T., Shan, B., Wang, H., et al. (2020). Reduction of mNAT1/hNAT2 contributes to cerebral endothelial necroptosis and aβ accumulation in Alzheimer's disease. *Cell Rep.* 33, 108447. doi:10.1016/j.celrep.2020.108447