



OPEN A Bayesian nonparametric method for jointly clustering multiple spatial transcriptomic datasets and simultaneous gene selection

Donald Turner^{1✉} & Yang Ni^{1,2}

In spatial transcriptomics, many algorithms are available for clustering cells into groups based on gene expression and location, although not without limitations. Such limitations include having to know the number of clusters, limiting inference to only one donor, and being unable to identify information common to multiple donors. To address these limitations, we propose a Bayesian nonparametric clustering algorithm capable of incorporating spatial transcriptomic data from multiple donors, which can identify clusters both common to all donors and idiosyncratic for each donor, features a variable selection of informative genes, and is able to determine the number of clusters automatically. Our method makes use of a Bayesian nonparametric method for combining inference across donors and a partition distribution indexed by pairwise distance information to cluster both within and across multiple spatial transcriptomics datasets. In our simulations and a real-data application, we show that our method can outperform other commonly used clustering algorithms.

It is known that the function of many biological systems (e.g., embryos, tumors) depends on the spatial organization of the cells¹. This dependence motivates the field of spatial transcriptomics, in which gene expression and spatial information at the cellular or spot level are used to model biological systems. Advancements in sequencing technologies, such as STARmap, have enabled robust intact-tissue RNA sequencing capable of simultaneous measurement of over 1,000 genes while preserving spatial information².

Interest in identifying sub-populations of cells based on gene expression and spatial location has motivated the development of statistical clustering methods for spatial transcriptomics data. Methods such as Giotto³ and Seurat⁴ are based on dimension reduction and nearest-neighbor clustering. The Python library stLearn makes use of graph-based community detection methods for clustering on normalized cellular spatial and gene expression data⁵. Although based on robust and well-understood methodology, these methods require specification of K , the total number of clusters, which may not be known *a priori*.

There are methods that can automatically select K , which tend to be Bayesian approaches. Such methods include BayesSpace⁶, which implements a Markov random field model for clustering. Using this method, K can either be specified beforehand or obtained from the elbow of the pseudo-log-likelihood. SPRUCE⁷ employs a Bayesian mixture model for clustering, either prespecifying K or using the widely applicable information criterion to determine K . DR-SC⁸ performs dimension reduction and spatial clustering jointly using a hidden Markov random field model. Similarly to SPRUCE, K is either fixed or selected using the modified Bayesian information criterion. These methods, although capable of identifying sub-populations without specification of K , do not incorporate information from multiple related spatial transcriptomics samples, which are becoming more and more available.

Allen et al.⁹ introduced MAPLE, a method which utilizes a spatial autoencoder, neighbor network, and finite Bayesian mixture model to jointly cluster observations using multiple spatial transcriptomics samples. However, it is often of interest to compare multiple groups of spatial transcriptomics data (e.g., healthy versus diseased tissue), and determine information common to each group and idiosyncratic within each group, which MAPLE is not capable of.

To address the aforementioned limitations of the existing methods, we propose a Bayesian nonparametric mixture model capable of performing spatial clustering on multiple spatial transcriptomics datasets with a data-driven determination of K , and can be used to identify information common to all experimental groups and idiosyncratic to each group. We also include a variable selection component, which can identify the genes

¹Texas A&M University, Department of Statistics, College Station 77843, USA. ²Texas A&M University, CPRIT Single Cell Data Science Core, College Station 77843, USA. ✉email: dturner@stat.tamu.edu

most significant in the clustering process via a spike-and-slab prior. This joint clustering and variable selection approach bypasses the post-selection inference problem, which arises when one first identifies clusters and then, given the clusters, identifies significant differentially expressed genes across clusters. Simulation results show that our method outperforms other established methods such as DR-SC in terms of clustering accuracy. The simulations also show that our variable selection is accurate. We further demonstrate the proposed method with the motivating STARmap dataset² and discuss the significance of our findings.

Methods

Let $Y_j \in \mathbb{R}^{n_j \times p}$ denote gene expression for n_j cells and p genes in group/donor $j = 1, \dots, J$ and let $X_j \in \mathbb{R}^{n_j \times 2}$ denote the matching spatial coordinates of the cells. Our goal is to cluster cells into unknown cell types from multiple spatial transcriptomic datasets $(Y_j, X_j)_{j=1}^J$ and, simultaneously, select genes that are differentially expressed across cell types. We call our method, **Multiple Spatial Sparse Clustering (MSC)**. A schematic illustration of the proposed MSC is provided in Fig. 1, which depicts three key inferential goals: (i) cell types common to all groups, (ii) cell types idiosyncratic to each group, and (iii) differentially expressed genes.

To achieve those goals, we introduce three sets of latent variables. First, let $r_{ji} \in \{0, 1\}$ be a binary variable such that $r_{ji} = 0$ if cell i in group j belongs to a cell type that is common to all groups (i.e., every donor has cells of this type) and $r_{ji} = 1$ if the cell type is idiosyncratic/unique to group j (i.e., only donor j has such cell type). Let $n_0 = \sum_{j=1}^J \sum_{i=1}^{n_j} I(r_{ji} = 0)$ and $n_j^* = \sum_{j=1}^J \sum_{i=1}^{n_j} I(r_{ji} = 1)$ then be the number of cells from the common cell types and from the cell types idiosyncratic to group j , respectively. Second, let $s_{ji}^0 \in \{1, \dots, q_{n_0}\}$ be a categorical variable such that $s_{ji}^0 = \ell$ if cell i in group j belongs to the common cell type ℓ . Similarly, let $s_{ji} \in \{1, \dots, q_{n_j^*}\}$ be a categorical variable such that $s_{ji} = \ell$ if cell i belongs to cell type ℓ idiosyncratic to group j . Note that both the number of common cell types q_{n_0} and the number of idiosyncratic cell types $q_{n_j^*}$ are unknown and allowed to grow with n_0 and n_j^* , respectively. Lastly, let $z_h \in \{0, 1\}$ be a binary variable such that $z_h = 1$ if gene h is differentially expressed across cell types and $z_h = 0$ otherwise. We propose a new Bayesian nonparametric hierarchical model that allows us to make inferences on these latent variables.

Specifically, given z_h , we assume the i th row $Y_{ji} = (y_{ji1}, \dots, y_{jip})$ of Y_j follows,

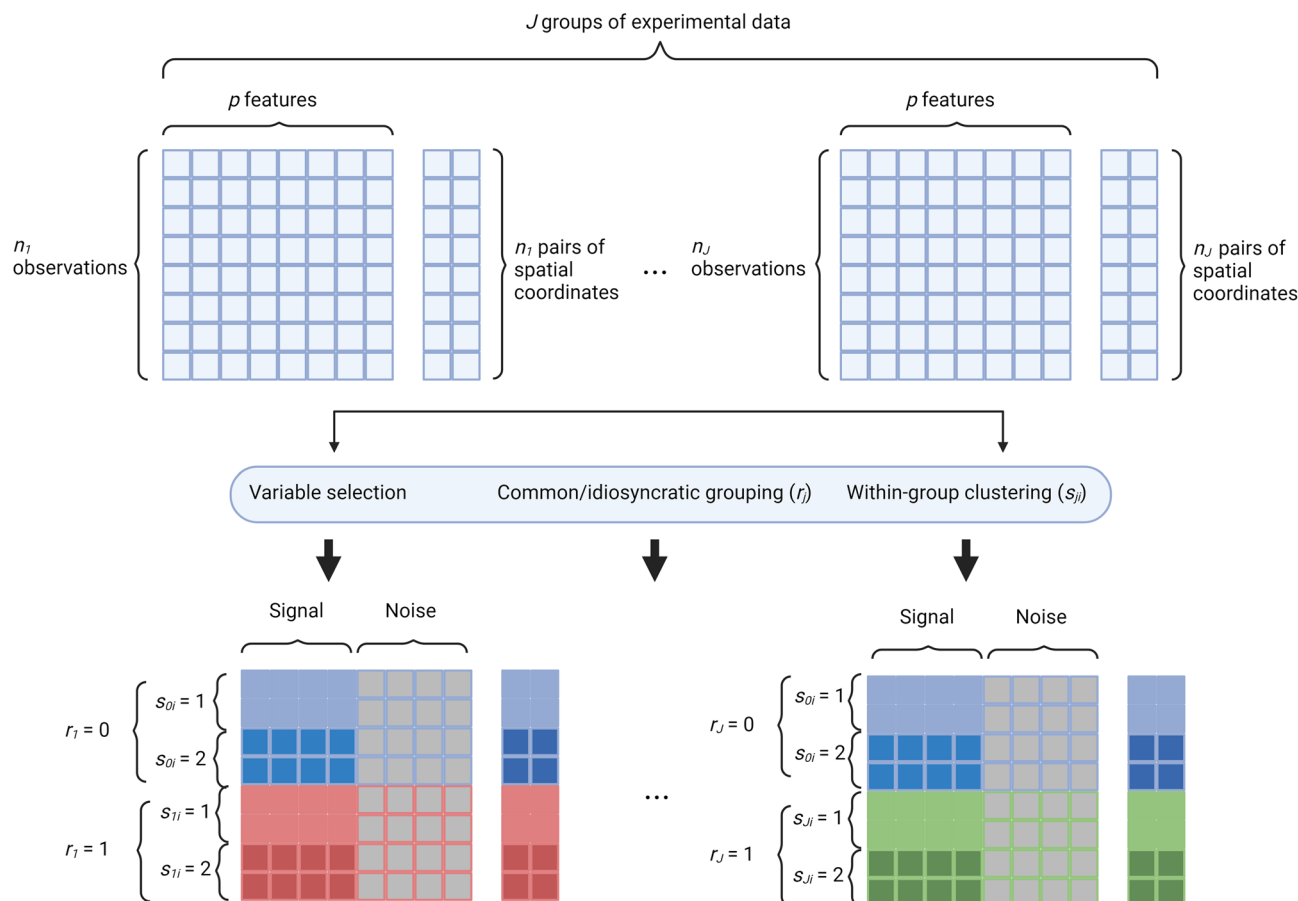


Fig. 1. Schematic illustration of the proposed MSC (created with BioRender).

$$\mathbf{Y}_{ji} \sim \prod_{h=1}^p N(y_{jih} | \theta_{jih}, \tau_h)^{z_h} N(y_{jih} | \eta_{jih}, \tau_h)^{1-z_h}.$$

Since $z_h = 1$ means that gene h is differentially expressed across cell types, a sophisticated prior model will be used for θ_{jih} to accommodate the heterogeneity whereas a simple prior model would suffice for η_{jih} .

Let $\theta_{ji} = \{\theta_{jih} | z_h = 1\}$ and let $\pi_{n_0} = \{S_1^0, \dots, S_{q_{n_0}}^0\}$ be the partition of n_0 cells from the common cell types where $S_\ell^0 = \{(j, i) | s_{ji}^0 = \ell\}$. Let $\delta_a(\cdot)$ denote a point mass at a . We assume that conditional on $r_{ji} = 0$ (i.e., common cell types),

$$\theta_{ji} | r_{ji} = 0 \sim \sum_{\ell=1}^{q_{n_0}} I(s_{ji}^0 = \ell) \delta_{\phi_\ell}(\cdot), \quad \phi_\ell \sim G_0, \quad \pi_{n_0} \sim p(\pi_{n_0}), \quad (1)$$

where G_0 is a base distribution and $p(\pi_{n_0})$ is a random partition distribution, both to be specified later. In words, (1) implies that cells of the same type (say, ℓ) share the same mean gene expression, i.e., $\theta_{ji} = \phi_\ell$ for all (j, i) such that $s_{ji}^0 = \ell$.

Similarly, let $\pi_{n_j^*} = \{S_1^j, \dots, S_{q_{n_j^*}}^j\}$ be the partition of n_j^* cells from the cell types idiosyncratic to group j where $S_\ell^j = \{(j, i) | s_{ji} = \ell\}$. We assume that conditional on $r_{ji} = 1$ (i.e., idiosyncratic cell types),

$$\theta_{ji} | r_{ji} = 1 \sim \sum_{\ell=1}^{q_{n_j^*}} I(s_{ji} = \ell) \delta_{\psi_{j\ell}}(\cdot), \quad \psi_{j\ell} \sim G_0, \quad \pi_{n_j^*} \sim p(\pi_{n_j^*}).$$

To incorporate the spatial information in partitioning the cells, we leverage the Ewens-Pitman attraction (EPA) distribution¹⁰. Specifically, let $\lambda(\cdot, \cdot)$ denote a similarity function such that $\lambda(i, i') = 1/d_{ii'}$ where $d_{ii'}$ is the Euclidean distance between cells i and i' . The EPA distribution sequentially allocates cells to clusters. The order in which cells are allocated is random, determined by a random permutation $\sigma = (\sigma_1, \dots, \sigma_n)$ of $\{1, \dots, n\}$, where the i th cell allocated is σ_i . The EPA distribution for the common partition is then given by,

$$p(\pi_{n_0}) = \prod_{j=1}^J \prod_{i:r_{ji}=0} p_{ji}(\alpha, \delta, \lambda, \pi(\sigma_1, \dots, \sigma_{n_{ji}})),$$

where δ is a discount parameter $\delta \in [0, 1)$, α is a mass parameter $\alpha > -\delta$, $n_{ji} = \sum_{j'=1}^j \sum_{i'=1}^{i-1} I(r_{j'i'} = 0)$, $p_{ji}(\alpha, \delta, \lambda, \pi(\sigma_1, \dots, \sigma_{n_{ji}})) = 1$ for (j, i) such that $r_{ji} = n_{ji} = 0$, and, for other (j, i) such that $r_{ji} = 0$ and $n_{ji} > 0$,

$$p_{ji}(\alpha, \delta, \lambda, \pi(\sigma_1, \dots, \sigma_{n_{ji}})) = P(\sigma_{n_{ji}+1} \in S | \alpha, \delta, \lambda, \pi(\sigma_1, \dots, \sigma_{n_{ji}})) \\ = \begin{cases} \frac{n_{ji} - \delta q_{n_{ji}}}{\alpha + n_{ji}} \times \frac{\sum_{s \in S} \lambda(\sigma_{n_{ji}+1}, \sigma_s)}{\sum_{s=1}^{n_{ji}} \lambda(\sigma_{n_{ji}+1}, \sigma_s)} & \text{for } S \in \pi(\sigma_1, \dots, \sigma_{n_{ji}}) \\ \frac{\alpha + \delta q_{n_{ji}}}{\alpha + n_{ji}} & \text{otherwise} \end{cases}$$

Note that $\frac{\sum_{s \in S} \lambda(\sigma_{n_{ji}+1}, \sigma_s)}{\sum_{s=1}^{n_{ji}} \lambda(\sigma_{n_{ji}+1}, \sigma_s)}$ encourages cell i in group j to be allocated to cluster S if it is spatially close to the cells that are already in cluster S . To make the inference of partition invariant to the order of allocations, Dahl et al. (2017)¹⁰ assumes a uniform prior on the permutation, $\sigma_0 = (\sigma_1, \dots, \sigma_{n_0}) \sim p(\sigma_0) \propto 1$, which we follow.

Similarly, for $j = 1, \dots, J$, the EPA distribution for the partition idiosyncratic to group j is given by,

$$p(\pi_{n_j^*}) = \prod_{i:r_{ji}=1} p_{ji}^*(\alpha, \delta, \lambda, \pi(\sigma_{j1}^*, \dots, \sigma_{jn_j^*}^*)),$$

where $n_{ji}^* = \sum_{i'=1}^{i-1} I(r_{ji} = 1)$, $p_{ji}^*(\alpha, \delta, \lambda, \pi(\sigma_{j1}^*, \dots, \sigma_{jn_j^*}^*)) = 1$ for i such that $r_{ji} = 1$ and $n_{ji}^* = 0$, and, for other i ,

$$p_{ji}^*(\alpha, \delta, \lambda, \pi(\sigma_{j1}^*, \dots, \sigma_{jn_j^*}^*)) = P(\sigma_{n_{ji}^*+1}^* \in S | \alpha, \delta, \lambda, \pi(\sigma_{j1}^*, \dots, \sigma_{jn_j^*}^*)) \\ = \begin{cases} \frac{n_{ji}^* - \delta q_{n_{ji}^*}}{\alpha + n_{ji}^*} \times \frac{\sum_{s \in S} \lambda(\sigma_{n_{ji}^*+1}^*, \sigma_s^*)}{\sum_{s=1}^{n_{ji}^*} \lambda(\sigma_{n_{ji}^*+1}^*, \sigma_s^*)} & \text{for } S \in \pi(\sigma_{j1}^*, \dots, \sigma_{jn_j^*}^*) \\ \frac{\alpha + \delta q_{n_{ji}^*}}{\alpha + n_{ji}^*} & \text{otherwise,} \end{cases}$$

and $\sigma_j^* = (\sigma_{j1}^*, \dots, \sigma_{jn_j}^*) \sim p(\sigma_j^*) \propto 1$.

We complete the specification of MSC with conjugate priors for the remaining hyperparameters,

$$\begin{aligned}\tau_h &\sim IG(a_\tau, b_\tau) \\ G_0 &= N(0, \tau_h/\kappa) \\ r_{ji} &\sim \text{Bernoulli}(\varepsilon), \quad \varepsilon \sim \text{beta}(a_\varepsilon, b_\varepsilon) \\ z_h &\sim \text{Bernoulli}(\rho_z), \quad \rho_z \sim \text{beta}(a_z, b_z)\end{aligned}$$

In all our implementation, we set $\alpha, a_\tau, b_\tau, a_\varepsilon, b_\varepsilon, a_z$ and b_z equal to 1, δ to 0, and κ to 0.008. We choose to non-informative hyperparameters rather than set hyperpriors on them to retain the conjugacy of our model, otherwise we introduce more computational complexity into our inference. $\alpha = 1$ is a typical choice for Dirichlet process mixture models¹¹². The choice of $\delta = 0$ in the EPA distribution implies that the distribution of subsets in our data partition is equivalent to a Dirichlet process mixture¹⁰. We choose a small value for κ to ensure a high variance in our likelihood, which provides a larger scope when searching for clusters across the sample space.

Posterior inference

We use the blocked Gibbs sampler to draw posterior inference. We choose to 'block' (i.e., sample simultaneously) cluster and grouping labels for computational benefits and potential help with mixing and convergence¹³. For simplicity, we assume $J = 2$ as in our real data. The sampler iteratively samples parameters $(r_{ji}, s_{ji}), \varepsilon, z_h$, and ρ_z from their respective full conditional distributions, which is outlined below.

1. *Sampling (r_{ji}, s_{ji}) .* For each observation Y_{ji} , we consider four cluster assignment options: 1) assignment to an existing idiosyncratic cluster, 2) generation of a new idiosyncratic cluster, 3) assignment to an existing common cluster, and 4) generation of a new common cluster. Let ϕ and φ denote the normal-inverse-gamma posterior predictive distribution and the marginal likelihood, respectively. Suppose we have K_j clusters in each idiosyncratic group and K_0 clusters in the common group. Then the probability for each option is given by,

$$\begin{aligned}u_{jk} &\propto (1 - \varepsilon)p(\pi_{n_j} | s_{ji} = k) \prod_{h=1}^p \phi(y_{jih} | \mathbf{Y}_{-jih}), \\ u_j^* &\propto (1 - \varepsilon)p(\pi_{n_j} | s_{ji} = K_j + 1) \prod_{h=1}^p \varphi(y_{jih}), \\ u_{0k} &\propto \varepsilon p(\pi_{n_0} | s_{0i} = k) \prod_{h=1}^p \phi(y_{jih} | \mathbf{Y}_{-jih}), \\ u_0^* &\propto \varepsilon p(\pi_{n_0} | s_{0i} = K_0 + 1) \prod_{h=1}^p \varphi(y_{jih}),\end{aligned}$$

where \mathbf{Y}_{-jih} is the set of all the other observations of gene h in the same cluster to which i belongs. Using the above probabilities, we sample:

$$(s_{ji}, r_{ji}) = \begin{cases} (k, 1) & \text{with probability } u_{jk}, \text{ for } k = 1, \dots, K_j, \\ (K_j + 1, 1) & \text{with probability } u_j^*, \\ (k, 0) & \text{with probability } u_{0k}, \text{ for } k = 1, \dots, K_0, \\ (K_0 + 1, 0) & \text{with probability } u_0^*. \end{cases}$$

2. *Sampling ε .* We sample ε directly from $\text{Beta}(a_\varepsilon + n_0, b_\varepsilon + \sum_j n_j^*)$.
3. *Sampling z_h .* Let $K = K_0 + K_1 + K_2$ be the total number of clusters across both experimental groups. For gene $h = 1, \dots, p$, let Y_{kh} be the cells in cluster $k = 1, \dots, K$ and Y_h be all the cells. We sample z_h from the following probabilities,

$$\begin{aligned}P(z_h = 1 | \rho_z, Y_h) &\propto \rho_z \prod_{k=1}^K \varphi(Y_{kh}), \\ P(z_h = 0 | \rho_z, Y_h) &\propto (1 - \rho_z) \varphi(Y_h).\end{aligned}$$

4. *Sampling ρ_z .* We sample ρ_z directly from $\text{Beta}(a_z + \sum_{h=1}^p z_h, b_z + p - \sum_{h=1}^p z_h)$.

To improve mixing, we implement an additional split-merge update for cluster (s_{ji}) and group (r_{ji}) assignments. We propose to split a common cluster into multiple idiosyncratic clusters, or merge multiple idiosyncratic clusters into one common cluster.

- *Split move.* Sample $k \in \{1, \dots, K_0\}$ uniformly. Let $\mathbf{r} = [r_1, r_2]$, and define \mathbf{Y} similarly. Then the Metropolis-Hastings ratio of accepting the splitting of the common cluster k into two new idiosyncratic clusters is given by,

$$\frac{(1/K_0) \prod_{l=1}^2 (\varphi(\mathbf{Y}_l | s_l = k, r_l = 1)(1 - \varepsilon)^{n_l} p(\pi_{n_l} | s_l = k))}{\prod_{l=1}^2 (1/K_l) (\varphi(\mathbf{Y} | \mathbf{r} = 0) \varepsilon^{n_0} p(\pi_{n_0}))},$$

If the splitting is accepted, we increase K_1 and K_2 by 1 and decrease K_0 by 1.

- *Merge move.* Sample $k_1 \in \{1, \dots, K_1\}$ and $k_2 \in \{1, \dots, K_2\}$ uniformly. Then the Metropolis-Hastings ratio of accepting the merge of idiosyncratic clusters k_1 and k_2 into one common cluster is given by

$$\frac{(1/K_1)(1/K_2)\varphi(\mathbf{Y} | s_0 \in \{k_1, k_2\}, \mathbf{r} = 0)\varepsilon^{n_0} p(\pi_{n_0} | s_0 \in \{k_1, k_2\})}{(1/K_0)\varphi(\mathbf{Y}_1 | r_1 = 0)\varphi(\mathbf{Y}_2 | r_2 = 0)(1 - \varepsilon)^{n_1 n_2} p(\pi_{n_1}) p(\pi_{n_2})}.$$

If the merge is accepted, we decrease K_1 and K_2 by 1 and increase K_0 by 1.

We give a pseudocode summary of the sampler below. Functions mentioned in the pseudocode are explained in more detail in the [supplementary materials](#).

Sampler pseudocode

```

Define  $n = \sum_{j=1}^J n_j$ 
Permute rows( $\mathbf{Y}$ )
for  $i \in 1, \dots, n$  do
  for  $j \in 1, \dots, J$  do
    for  $k \in 0, \dots, K_j$  do ▷  $k = 0$  for new cluster
      if  $k = 0$  then
         $L = \text{ML}(\mathbf{Y}_{ji} | \dots)$  ▷ ML is the Normal-Inverse-Wishart marginal likelihood
      else
         $L = \text{PP}(\mathbf{Y}_{ji} | \mathbf{Y}_{j,-i}, s_{ji} = k, r_{ji} = 1, \dots)$  ▷ PP is the Normal-Inverse-Wishart posterior predictive distribution
         $\pi_j = \text{EP}(D | s_{ji} = k, r_{ji} = 1, \dots)$  ▷ EP gives partition probabilities via the EPA distribution
         $p_{jk} = L \times \pi_j \times (1 - \varepsilon)$ 
      end if
    end for
  end for
  for  $k \in 0, \dots, K_0$  do
    if  $k = 0$  then
       $L = \text{ML}(\mathbf{Y}_{ji} | \dots)$ 
    else
       $L = \text{PP}(\mathbf{Y}_{ji} | \mathbf{Y}_{-ji}, s_i = k, r_i = 0, \dots)$ 
       $\pi_0 = \text{EP}(D | s_i = k, r_i = 0, \dots)$ 
       $p_{0k} = L \times \pi_0 \times \varepsilon$ 
    end if
  end for
end for
 $\mathbf{p}_j = \{p_{jk}\}_{k=0}^{K_j}$ 
 $\mathbf{p}_0 = \{p_{0k}\}_{k=0}^{K_0}$ 
Sample  $k^*$  from  $[K_j + 1, 1, \dots, K_j, K_0 + 1, 1, \dots, K_0]$  with probabilities  $[\mathbf{p}_j, \mathbf{p}_0]$ 
Set  $s_{ji} = k^*$  and  $\mu_{ji} = \mu_{jk}^*$ 
if  $k^* \in [K_j + 1, \dots, K_j]$  then
   $r_{ji} = 1$ 
else
   $r_{ji} = 0$ 
end if

```

```

 $N_1 = \sum r$ 
 $N_0 = n - N_1$ 
 $\varepsilon = \text{rbeta}(a_\varepsilon + N_0, b_\varepsilon + N_1)$ 
 $p_{z1} = p_{z0} = 1$ 
for  $h \in 1, \dots, p$  do
  for  $l \in \text{unique}(s)$  do
     $p_{z1} = p_{z1} \times \text{zL}(Y_h | s_h = p, \dots)$   $\triangleright$  zL is the normal marginal likelihood for selection indicators  $z_h$ 
  end for
   $p_{z1} = p_{z1} \times \pi_z$ 
   $p_{z0} = \text{zL}(Y_h | \dots) \times \pi_z$ 
  Sample  $z^*$  from  $\{0, 1\}$  with probabilities  $\{p_{z0}, p_{z1}\}$ 
   $z_h = z^*$ 
end for
 $\pi_z = \text{rbeta}(a_\pi + \sum z, b_\pi + p - \sum z)$ 
if  $K_0 > 1$  then
  Perform split update. The ratio stated in section 2.1 of the supplementary materials is given below
  
$$\frac{(1/K_0)(\prod_{l \in L} \text{ML}(Y_l | s_l = k, r_l = 1))(1 - \varepsilon)^{n_l} \text{EP}(\pi_{n_l} | s_l = k)}{(\prod_{l \in L} 1/K_l) \text{ML}(Y | r = 0) \varepsilon^{n_0} \text{EP}(\pi_{n_0})}$$

end if
if  $K_{j_1} > 0$  and  $K_{j_2} > 0$  for  $j_1, j_2 \in 1, \dots, J$  then
  Perform merge update with the given merging probability:
  
$$\frac{(1/K_{j_1})(1/K_{j_2}) \text{ML}(Y | s_0 \in \{k_1, k_2\}, r = 0) \varepsilon^{n_0} \text{EP}(\pi_{n_0} | s_0 \in \{k_1, k_2\})}{(1/K_0) \text{ML}(Y_{j_1} | r_{j_1} = 0) \text{ML}(Y_{j_2} | r_{j_2} = 0) (1 - \varepsilon)^{n_{j_1} + n_{j_2}} \text{EP}(\pi_{n_{j_1}}) \text{EP}(\pi_{n_{j_2}})}$$

end if
end for
end for

```

Real data

We demonstrate the proposed model on a publicly available real spatial transcriptomic dataset obtained from the work of Wang *et al.*, taken from the visual cortices of a number of mice using the STARmap single-cell RNA sequencing technique². The mice were placed under two experimental conditions: one given an hour of light exposure after 4 days kept in a dark environment, and the other kept in continual darkness. We randomly picked one mouse from the light exposure group ($n_1 = 837$) and another from the continual darkness group ($n_2 = 927$). We will refer to them as Mouse 1 and Mouse 2, respectively. We considered the same 17 spatially varying genes as in Chakrabarti *et al.* (2023)¹⁴. We also followed their preprocessing steps: we removed cells showing extreme expression of genes and log-normalized the data with a scaling factor equal to the median expression of total reads per cell. Genes were standardized and spatial distances were normalized to (0, 1). Our goal is to cluster cells based on gene expression and spatial location both within each experimental group and across the groups and to select differentially expressed genes across clusters. We ran a Markov chain for 5,000 iterations with a thinning factor of 1/25. No burn-in period was implemented.

Results

Of the 17 genes, we identified 12 significant differentially expressed genes: PLCXD2, RORB, Cux2, Pcp4, NRN1, Nectin3, ARX, OTOF, PROK2, Homer1, eRNA3, and SLC17A7.

We investigate the significance of the selected genes. PLCXD2 by itself has no known function in the brain, but together with another gene, GPR158, it is part of a signaling complex responsible for the development of the dendritic spine¹⁵. RORB, according to GeneCards¹⁶, is linked to circadian rhythms and plays a role in the development of epilepsy. Cux2, like PLCXD2, is involved with the development of the dendritic spine¹⁷. Pcp4 is expressed in bone marrow stem cells, in which it is associated with the deposition of calcium¹⁸. NRN1 is associated with neurodevelopment and synaptic plasticity and serves as a biomarker for schizophrenia¹⁹. Nectin3 regulates the creation of cellular adhesive molecules called nectins²⁰. ARX plays a role in the development of the forebrain, pancreas, and testes²¹. Mutations in the OTOF gene have been shown to result in auditory neuropathy²². PROK2 is a biomarker for Kallmann syndrome, which is characterized by impaired sense of smell and delayed puberty²³ due to the underdevelopment of neurons in the brain that signal the hypothalamus. Homer1, similarly to PLCXD2 and Cux2, is involved with the development of the dendritic spine²⁴. eRNA3 is known to regulate gene expression, but its function remains debated²⁵. SLC17A7 is a member of the SLC17 family of genes found in neuron-rich areas of the brain, responsible for glutamate transport²⁶.

A gene set enrichment analysis was performed on the 12 selected genes using Enrichr^{27–29}. Analysis was performed on the MGI Mammalian Phenotype Level 4 2024 ontology³⁰.

As shown in Fig. 2, the phenotype term most strongly associated with our gene set is abnormal miniature excitatory postsynaptic currents, which is a “defect in the size or duration of spontaneous currents detected in postsynaptic cells that occur in the absence of an excitatory impulse”³⁰. The associated genes were NRN1, Homer1, OTOF, and SLC17A7. This may suggest that the difference in light exposure between the mice could

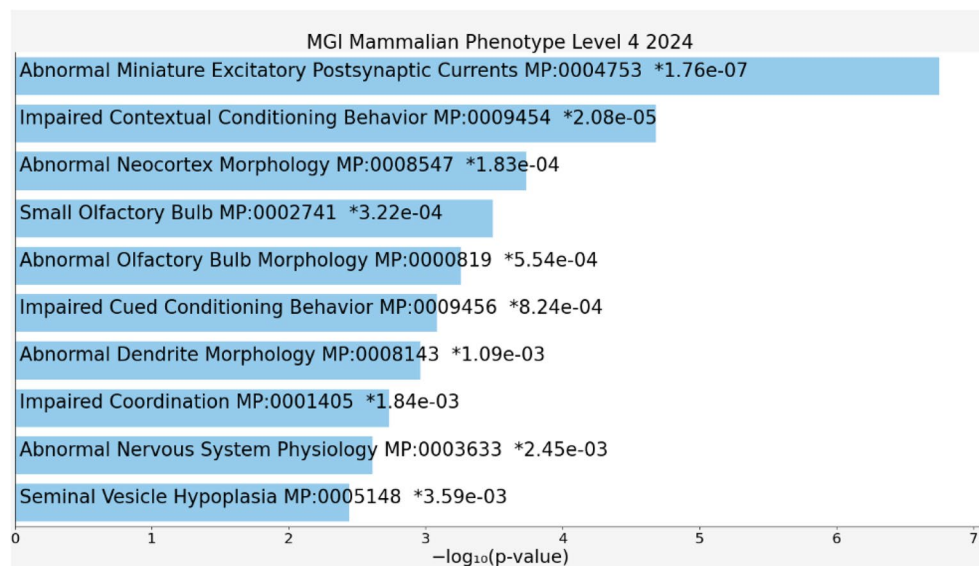


Fig. 2. Gene set enrichment results for the MGI Mammalian Phenotype Level 4 2024 ontology.

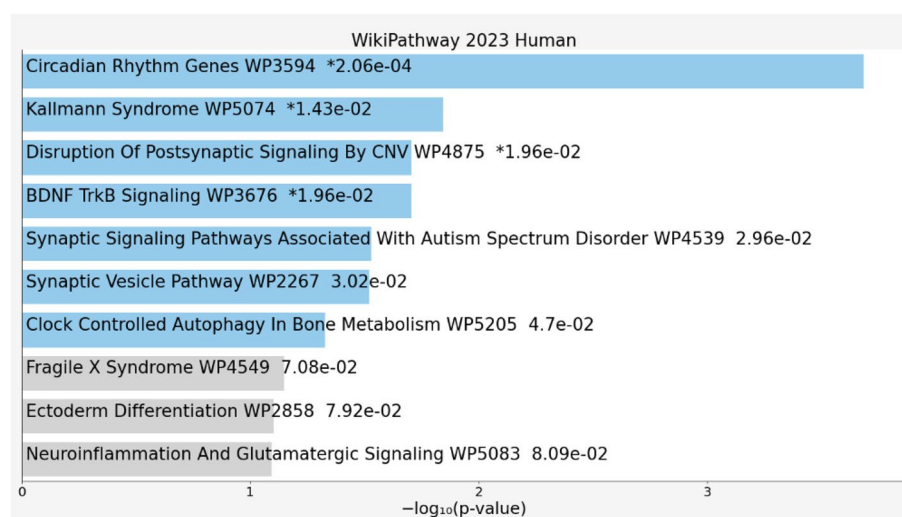


Fig. 3. Gene set enrichment results for the WikiPathway 2023 Human database. A star indicates that the p -value adjusted for multiple testing is also significant.

be related to the aforementioned defect. The second term, impaired contextual conditioning behavior, denotes impaired ability to associate an aversive experience with the neutral environment³⁰. The associated genes are NRN1, Homer1, and ARX. As with the first term, our results may indicate that light exposure or lack thereof inhibits this conditioning.

To see how such genes are functionally relevant for humans, we also performed an enrichment analysis using the WikiPathway 2023 Human database³¹. As shown in Fig. 3, our selected gene set is most closely related to the circadian rhythm genes. The overlapping genes were Homer1, PROK2, and RORB. The significance of these enriched genes in Mouse 1 and Mouse 2 indicates that the relationship between the difference in light exposure and the expression of genes associated with circadian rhythm may be worth further investigation. Also, interestingly, a significant pathway in the WikiPathway enrichment results is related to the disruption of postsynaptic signaling by copy number variations, reinforcing the connection between the 12 selected genes and postsynaptic signaling as suggested in MGI enrichment results.

A graph of protein-protein interactions was obtained from the STRING database³² and is shown in Fig. 4. The graph shows coexpression for Slc17a7 between both Homer1 and RORB, reinforcing their connection to synaptic/circadian functions in our enrichment analysis. Interestingly, no connections among NRN1, Homer1, ARX, or OTOF are shown in the graph although they are associated in the mammalian phenotype enrichment analysis.

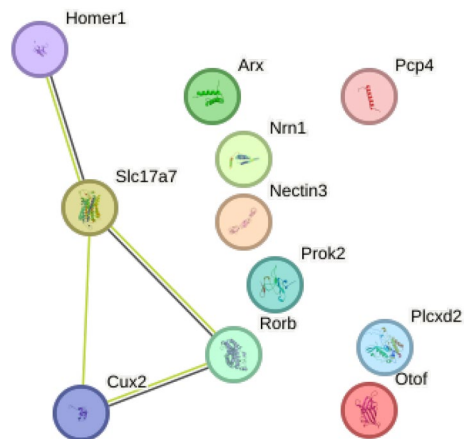


Fig. 4. Protein-protein interaction graph obtained from the STRING database. 11 of the 12 selected genes are shown (eRNA3 was not available in the database).

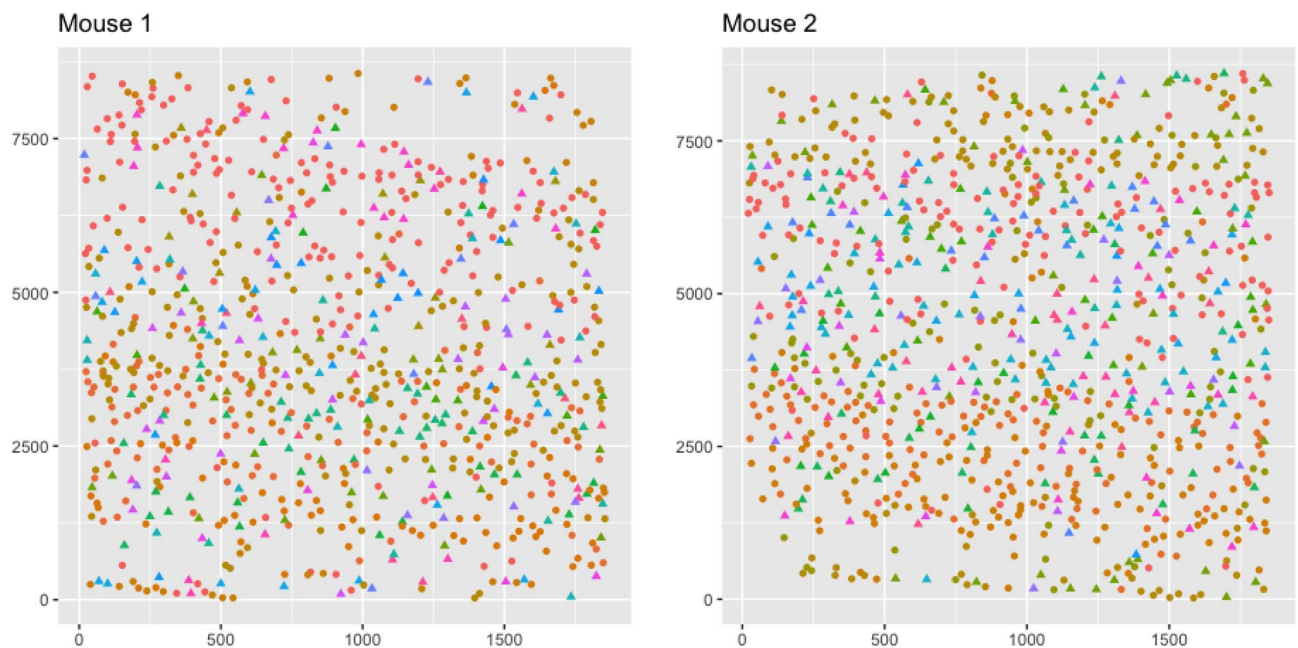


Fig. 5. Clustering results on the cellular measurement locations for each mouse. Colors represent cluster assignment, and the point type corresponds to group assignment (circles for the common model, triangles for idiosyncratic).

In terms of the clustering results, our algorithm assigned most cells to the 5 common cell types shared between the two mice (589 out of 837 cells for Mouse 1, and 610 out of 927 cells for Mouse 2). We found 26 idiosyncratic clusters for Mouse 1, and 16 idiosyncratic clusters for Mouse 2. The clustering results are visualized spatially in Fig. 5. To visualize these clusters, we applied UMAP³³ to the gene expression to reduce the data to 2 dimensions, which shows the similarity of cells in gene expression in a 2-dimensional space.

Fig. 6 displays the UMAP embeddings of cells assigned to the common clusters and idiosyncratic clusters in Groups 1 and 2. It reveals distinct clusters in the embedded space, especially for the common clusters. We also see that each idiosyncratic group covers a different region in the 2-dimensional space, which would be expected for information unique to each group. Compared to the common clusters, there are much more cell clusters in the idiosyncratic groups. There are a few potential explanations for this somewhat expected result. For example, mitochondrial content, which is known to affect gene expression³⁴, was not measured in our data. The cell cycle stages of individual cells can also affect gene expression³⁵ and hence the clustering. These would lead to higher cellular heterogeneity in terms of gene expression than the heterogeneity arising from cell types alone, which could partially explain why we found many more idiosyncratic clusters than the expected number of cell types in mouse brains.

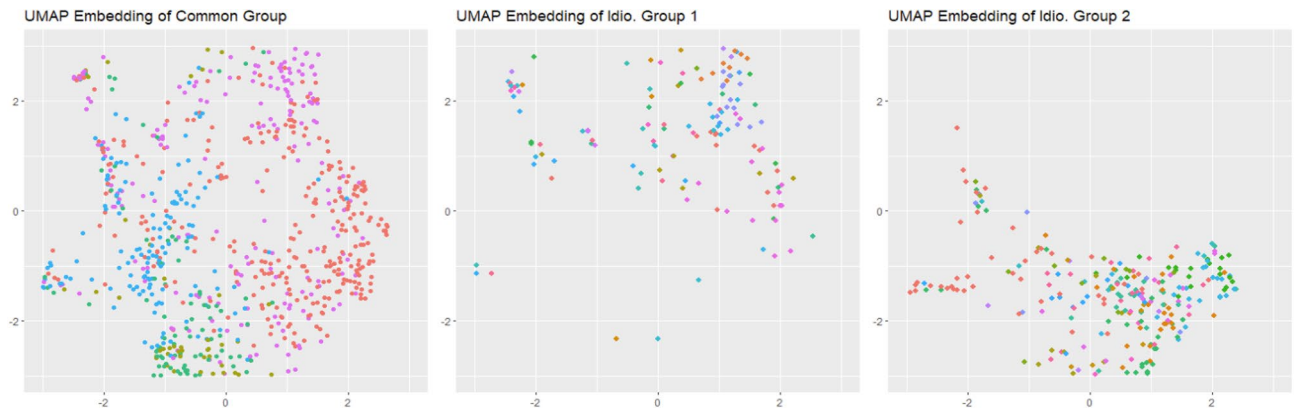


Fig. 6. UMAP embedding of cells assigned to the common clusters and idiosyncratic clusters in Groups 1 and 2 from left to right. Cluster assignments are indicated by colors.

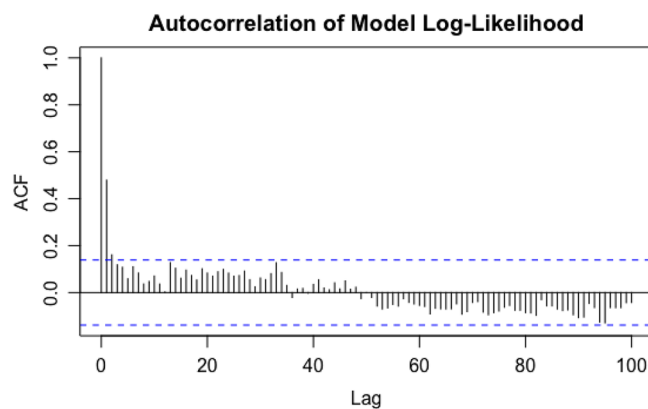


Fig. 7. Autocorrelation function for the model log-likelihood. Blue dashed lines are the boundaries of a 95% confidence interval.

To assess convergence, the autocorrelation function was calculated for the model log-likelihood, and is plotted in Fig. 7. Rapid decay toward 0 indicates that the sampler has converged and produces independent draws from the posterior distribution. The Geweke diagnostic was calculated on our log-likelihood chain, resulting in a Z-score of -0.8591 and a corresponding p -value of 0.3903 , providing further evidence of convergence.

Inference was run on a personal computer using an AMD Ryzen 5 3600 processor and 16 GB of memory. Using this machine yielded a computing time of 13 days. Although our computing time is heavy compared to existing methods, our method allows for much richer inference including spatial clustering, variable selection, comparison between experimental groups, and uncertainty quantification. Most of our code was implemented in R, aside from the function for obtaining partition probabilities from the EPA distribution (section 1.2.4 in the [supplementary materials](#)), which was coded in C++. A full C++ implementation may improve computing time, and will be a subject of our future investigations.

Simulations

In addition to real data, we also conduct simulations to evaluate the proposed method. We consider 6 scenarios with varying dimensions $p = 15, 20, 25, 30, 40, 50$. Under each scenario, we fix the number of signal variables to 5 (i.e., $p - 5$ variables are noises) and the number of samples per group to $n_j = 100$ for group $j = 1, 2$. The five signal variables are generated together with the two-dimensional spatial coordinates from a mixture of seven-dimensional multivariate Gaussian distributions, with a proportion $\varepsilon = 0.6$ of data in each group generated from a common model with three clusters, and the remaining generated from two idiosyncratic models with two clusters for each group. The mixture of Gaussian shares the same banded covariance matrix where the diagonal elements are equal to 0.15, the v th diagonals are $0.03/v$ for $v = 1, 2, 3, 4$, and the rest are 0. The mean for each cluster is sampled from a grid of seven equally spaced numbers ranging from 1 to 10. The $p - 5$ noise variables are generated from independent standard normal distributions.

A Gibbs sampler was used to obtain cluster labels s_{ji} , group assignment r_{ji} , and variable selection z_h over 3000 iterations, discarding a burn-in period of 2000 iterations. Cluster labels s_{ji} were determined via the least-square criterion³⁶, and variable selection z_h and the selection r_{ji} of common vs idiosyncratic model were determined by the mean probability model (i.e., using 0.5 as a cutoff to threshold their probabilities). For

	MSC	K-means	DR-SC	BayesSpace
$n_j = 100$, $p = 15$	0.993 (0.002)	0.877 (0.003)	0.955 (0.007)	0.895 (0.022)
$n_j = 100$, $p = 20$	0.993 (0.003)	0.867 (0.003)	0.905 (0.007)	0.932 (0.018)
$n_j = 100$, $p = 25$	0.988 (0.003)	0.854 (0.004)	0.901 (0.007)	0.945 (0.012)
$n_j = 100$, $p = 30$	0.994 (0.002)	0.843 (0.005)	0.874 (0.007)	0.933 (0.014)
$n_j = 100$, $p = 40$	0.992 (0.003)	0.819 (0.004)	0.873 (0.006)	0.949 (0.008)
$n_j = 100$, $p = 50$	0.992 (0.003)	0.800 (0.005)	0.852 (0.009)	0.937 (0.014)

Table 1. Combined group NMI from our simulations. Standard errors are given in parentheses. There are 5 signal variables in each scenario, the others being noise.

evaluation, we calculated the normalized mutual information (NMI) for cluster labels s_{ji} , misclassification rate for r_{ji} , and true positive & false discovery rates for variable selection z_h . NMI was calculated on the two groups of labels concatenated to one vector, compared to the true labels (also concatenated to one vector). We compare the clustering performance of our method to the popular K-means model, the joint dimension reduction and spatial clustering (DR-SC⁸) method, and BayesSpace⁶. Both K-means and DR-SC were applied separately to each experimental group with the true number of clusters $K = 5$ (3 common and 2 idiosyncratic clusters per experimental group), and NMI was calculated on the resulting group labels concatenated to a single vector, also compared to the true labels. BayesSpace was implemented for 5000 iterations, discarding a burn-in period of 2000 iterations.

For $p = 15$ across 50 replicates, our method achieved a mean NMI 0.993, a mean misclassification rate for r 0.06, and a true positive rate of 1 and a false negative rate of 0 for z . In comparison, K-means obtained mean Group 1 and Group 2 NMI of 0.877, DR-SC returned NMI of 0.955, and BayesSpace returned an NMI of 0.895. Our method outperforms the others similarly in terms of clustering across the other scenarios of varying data dimension, as shown by Table 1. The true positive and false negative rate of z is 1 and 0 respectively across all scenarios. For $p = \{20, 25, 30, 40, 50\}$, we obtain respective r misclassification rates of $\{0.05, 0.06, 0.03, 0.06, 0.05\}$. We emphasize that the competing method cannot make inferences about variable selection or group assignment.

Conclusion

We have introduced a Bayesian nonparametric model for clustering observations from multiple datasets and grouping observations across datasets using information common across all datasets and idiosyncratic to each dataset in this paper. Clustering, grouping, variable selection, and K are all determined automatically in the proposed Bayesian model. Our simulations and real data application demonstrate the proposed method, with simulations showing the advantage of the method over existing clustering algorithms.

We designed our model as a synthesis of a nonparametric Bayesian method for combining inference across experiments, and a partition distribution prior indexed by pairwise information. We have shown how the model is implemented in a Gibbs sampler. Our model was applied to spatial transcriptomic data for two mice, showing its ability to cluster and group observations using gene expression and pairwise distances, and its capability for selecting significant genes. We have discussed the selected genes and their biological significance.

There remain many extensions and possible future work for this model. Here, we have applied our model to two groups of data, but theoretically, this model may be used for an arbitrary number of datasets. The limiting factor in scaling this method to include more datasets is runtime. A component for data preprocessing could also be added to the model, allowing all-in-one analysis of raw transcriptomic data. In future work, we will investigate ways to improve computing efficiency and including additional components to bolster the versatility of our model.

Data availability

All data generated or analysed during this study are included in this published article [and its [supplementary information files](#)].

Received: 21 February 2025; Accepted: 11 July 2025

Published online: 24 July 2025

References

1. Moses, L. & Pachter, L. Museum of spatial transcriptomics. *Nat. Methods* **19**, 534–546. <https://doi.org/10.1038/s41592-022-0140-9-2> (2022) (Number: 5 Publisher: Nature Publishing Group).
2. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
3. Dries, R. et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol.* **22**, 78. <https://doi.org/10.1186/s13059-021-02286-2> (2021).

4. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29. <https://doi.org/10.1016/j.cell.2021.04.048> (2021).
5. Pham, D. et al. stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *BioRxiv* 2020–05 (2020). Publisher: Cold Spring Harbor Laboratory.
6. Zhao, E. et al. Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology* **39**, 1375–1384. <https://doi.org/10.1038/s41587-021-00935-2> (2021) (Number: 11 Publisher: Nature Publishing Group.).
7. Allen, C. et al. A Bayesian multivariate mixture model for high throughput spatial transcriptomics. *Biometrics* **79**, 1775–1787. <https://doi.org/10.1111/biom.13727> (2023). eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.13727>.
8. Liu, W. et al. Joint dimension reduction and clustering analysis of single-cell RNA-seq and spatial transcriptomics data. *Nucleic Acids Res.* **50**, e72. <https://doi.org/10.1093/nar/gkac219> (2022).
9. Allen, C., Chang, Y., Ma, Q. & Chung, D. MAPLE: A Hybrid Framework for Multi-Sample Spatial Transcriptomics Data, <https://doi.org/10.1101/2022.02.28.482296> (2022). Pages: 2022.02.28.482296 Section: New Results.
10. Dahl, D. B., Day, R. & Tsai, J. W. Random partition distribution indexed by pairwise information. *J. Am. Stat. Assoc.* **112**, 721–732 (2017).
11. McAuliffe, J. D., Blei, D. M. & Jordan, M. I. Nonparametric empirical bayes for the dirichlet process mixture model. *Stat. Comput.* **16**, 5–14 (2006).
12. Petrone, S., Guindani, M. & Gelfand, A. E. Hybrid dirichlet mixture models for functional data. *J. R. Stat. Soc. Series B Stat. Methodol.* **71**, 755–782 (2009).
13. Jensen, C. S., Kjærulff, U. & Kong, A. Blocking gibbs sampling in very large probabilistic expert systems. *Int. J. Hum. Comput. Stud.* **42**, 647–666 (1995).
14. Chakrabarti, A., Ni, Y. & Mallick, B. K. Bayesian flexible modelling of spatially resolved transcriptomic data (2023). 2305.08239.
15. Amado, M. L. D. P. F. *Deciphering the molecular mechanisms that mediate postsynaptic maturation*. Master's thesis (2022).
16. Stelzer, G. et al. The GeneCards suite: From gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinformatics* **54**, 1.30.1–1.30.33. <https://doi.org/10.1002/cpbi.5> (2016). <https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/cpbi.5>.
17. Cubelos, B. et al. Cux1 and cux2 regulate dendritic branching, spine morphology, and synapses of the upper layer neurons of the cortex. *Neuron* **66**, 523–535 (2010).
18. Xiao, J. et al. Expression of Pcp4 gene during osteogenic differentiation of bone marrow mesenchymal stem cells in vitro. *Mol. Cell. Biochem.* **309**, 143–150. <https://doi.org/10.1007/s11010-007-9652-x> (2008).
19. Chandler, D. et al. Impact of neuritin 1 (nrn1) polymorphisms on fluid intelligence in schizophrenia. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **153B**, 428–437. <https://doi.org/10.1002/ajmg.b.30996> (2010). <https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajmg.b.30996>.
20. Xu, F. et al. Nectin-3 is a new biomarker that mediates the upregulation of MMP2 and MMP9 in ovarian cancer cells. *Biomed. Pharmacother.* **110**, 139–144. <https://doi.org/10.1016/j.biopha.2018.11.020> (2019). Place: France.
21. Géczy, J., Cloosterman, D. & Partington, M. ARX: a gene for all seasons. *Curr. Opin. Genet. Dev.* **16**, 308–316. <https://doi.org/10.1016/j.gde.2006.04.003> (2006).
22. Varga, R. et al. Non-syndromic recessive auditory neuropathy is the result of mutations in the otoferlin (OTOF) gene. *J. Med. Genet.* **40**, 45–50. <https://doi.org/10.1136/jmg.40.1.45> (2003). Publisher: BMJ Publishing Group Ltd eprint: <https://jmg.bmj.com/content/40/1/45.full.pdf>.
23. Dodé, C. & Rondard, P. PROK2/PROKR2 Signaling and Kallmann Syndrome. *Front. Endocrinol.* **4**, <https://doi.org/10.3389/fendo.2013.00019> (2013).
24. Yamazaki, H. & Shirao, T. Homer, Spikar, and Other Drebrin-Binding Proteins in the Brain. *Adv. Exp. Med. Biol.* **1006**, 249–268. https://doi.org/10.1007/978-4-431-56550-5_14 (2017) (Place: United States.).
25. Carullo, N. V. N. et al. Enhancer RNAs predict enhancer-gene regulatory links and are critical for enhancer function in neuronal systems. *Nucleic Acids. Res.* **48**, 9550–9570. <https://doi.org/10.1093/nar/gkaa671> (2020) (Place: England.).
26. Reimer, R. J. SLC17: A functionally diverse family of organic anion transporters. *Mol. Aspects Med.* **34**, 350–359. <https://doi.org/10.1016/j.mam.2012.05.004> (2013).
27. Chen, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
28. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
29. Xie, Z. et al. Gene set knowledge discovery with enrichr. *Curr. Protoc.* **1**, e90 (2021).
30. Smith, C. L. & Eppig, J. T. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **1**, 390–399 (2009).
31. Agrawal, A. et al. WikiPathways 2024: next generation pathway database. *Nucleic Acids Res.* **52**, D679–D689. <https://doi.org/10.1093/nar/gkad960> (2023). <https://academic.oup.com/nar/article-pdf/52/D1/D679/55040703/gkad960.pdf>
32. Szklarczyk, D. et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).
33. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction (2020). [arXiv:1802.03426](https://arxiv.org/abs/1802.03426).
34. Osorio, D. & Cai, J. J. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. *Bioinformatics* **37**, 963–967 (2021).
35. Riba, A. et al. Cell cycle gene regulation dynamics revealed by RNA velocity and deep-learning. *Nat. Commun.* **13**, 2865 (2022).
36. Dahl, D. B. Model-based clustering for expression data via a Dirichlet process mixture model. *Bayesian inference for gene expression and proteomics* **4**, 201–218 (2006).

Author contributions

D.T. wrote the main manuscript text, including tables and figures. Y.N. served as advisor to D.T. in the writing process, provided rewrites within the Methods section and edited the Simulations section. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-11693-5>.

Correspondence and requests for materials should be addressed to D.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025