frontiers in Genetics

METHODS
published: 26 April 2021
doi: 10.3389/fgene.2021.656637

Recovering Spatially-Varying Cell-Specific Gene Co-expression Networks for Single-Cell Spatial Expression Data

*Jinge Yu and Xiangyu Luo**

Institute of Statistics and Big Data, Renmin University of China, Beijing, China

Recent advances in single-cell technologies enable spatial expression profiling at the cell level, making it possible to elucidate spatial changes of cell-specific genomic features. The gene co-expression network is an important feature that encodes the gene-gene marginal dependence structure and allows for the functional annotation of highly connected genes. In this paper, we design a simple and computationally efficient two-step algorithm to recover spatially-varying cell-specific gene co-expression networks for single-cell spatial expression data. The algorithm first estimates the gene expression covariance matrix for each cell type and then leverages the spatial locations of cells to construct cell-specific networks. The second step uses expression covariance matrices estimated in step one and label information from neighboring cells as an empirical prior to obtain thresholded Bayesian posterior estimates. After completing estimates for each cell, this algorithm can further predict or interpolate gene co-expression networks on tissue positions where cells are not captured. In the simulation study, the comparison against the traditional cell-type-specific network algorithms and the cell-specific network method but without incorporating spatial information highlights the advantages of the proposed algorithm in estimation accuracy. We also applied our algorithm to real-world datasets and found some meaningful biological results. The accompanied software is available on https://github.com/jingeyu/CSSN.

Keywords: Bayesian posterior estimates, cell-specific, gene co-expression network, prediction, single-cell spatial expression, neighborhood

OPEN ACCESS

Edited by:
Jiebiao Wang,
University of Pittsburgh, United States

Reviewed by:
Jinjin Tian,
Carnegie Mellon University, United States
Hao Dai,
Chinese Academy of Sciences (CAS), China

****Correspondence:***
Xiangyu Luo
xiangyuluo@ruc.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *21 January 2021*
Accepted: *18 March 2021*
Published: *26 April 2021*

Citation:
Yu J and Luo X (2021) Recovering Spatially-Varying Cell-Specific Gene Co-expression Networks for Single-Cell Spatial Expression Data. Front. Genet. 12:656637. doi: 10.3389/fgene.2021.656637

1. INTRODUCTION

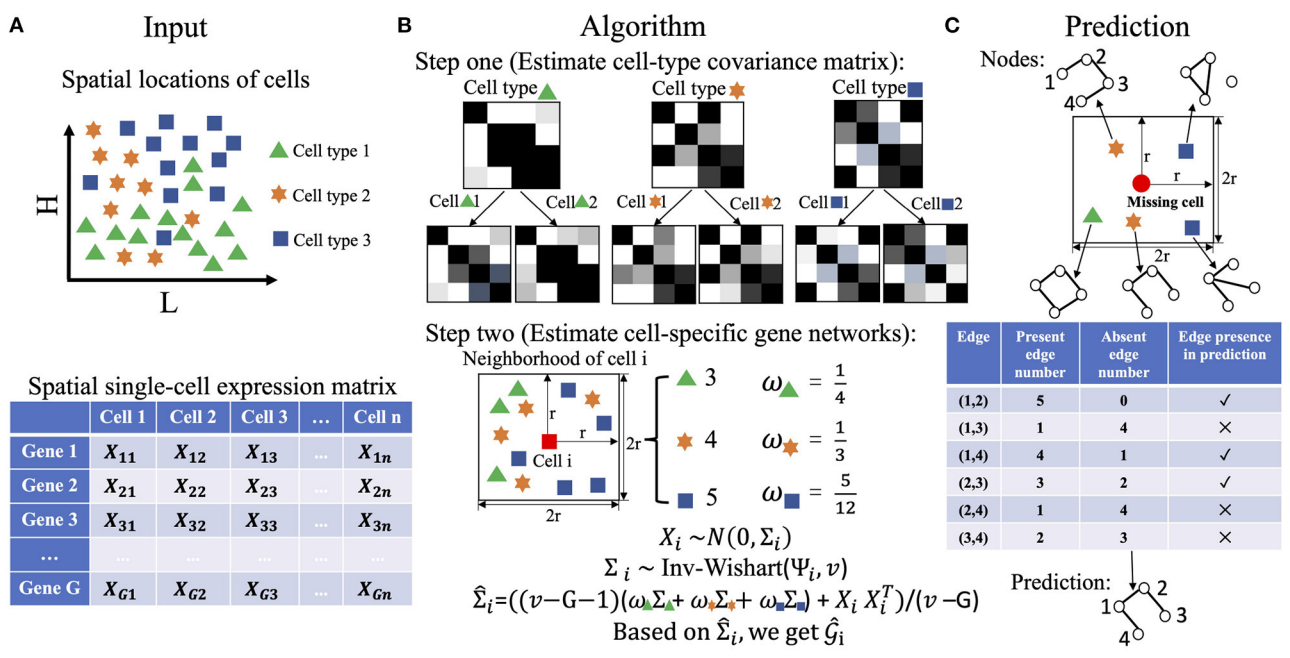
The last decade witnesses that the single-cell RNA-sequencing has revolutionized the focus of genomic analyses from bulk samples to single cells, but the technology loses important cell spatial information during tissue dissociation. Fortunately, recent technological advances have allowed for measurements of the gene expression levels at single-cell resolution while retaining the coordinates of cells in the tissue section (Chen et al., 2015; Moffitt et al., 2018; Wang et al., 2018). Specifically, various spatially resolved transcriptomic techniques have been developed to profile single-cell expression with cells' spatial information, including MERFISH (Chen et al., 2015), seqFISH (Lubeck et al., 2014), and FISSEQ (Lee et al., 2014), just to name a few. They are mainly based on either *in situ* hybridization or *in situ* sequencing. Fluorescence *in situ* hybridization (FISH) based approaches can measure hundreds of preselected marker genes, while *in situ* sequencing based approaches

can measure thousands of transcripts. Moreover, different techniques may have different strategies to capture transcriptomic spatial information. For example, MERFISH adopts an imaging-based way to map transcriptomic spatial organization for a three-dimensional tissue region. Usually, the region needs to be first sectioned into evenly spaced slices, and MERFISH is then performed on these slices, resulting in two-dimensional localization information. The information makes it possible to investigate spatial and functional organization of cells.

The amazing biological progress also offers rich opportunities to investigate the spatial patterns of cell-specific genomic features (Zhang et al., 2020). When features are genes, Sun et al. (2020) developed a statistical method to identify genes with spatially differential expressions. Li D. et al. (2020) utilized an expert system to predict signaling gene expression using information from nearby cells. However, as observed gene expressions may suffer from systematic biases (Köster et al., 2019) and are dynamically driven by an underlying regulation system, it is of more interest to study a more stable feature—gene co-expression network—(Dai et al., 2019) and learn its spatial pattern from one cell to another.

The gene co-expression network (Butte and Kohane, 2000; Stuart et al., 2003; Carter et al., 2004) can be encoded in an undirected graph, where nodes correspond to genes and an edge between nodes A and B indicates a significant association between expressions of the genes A and B. It has important biological applications including functional annotation for a set of unknown but highly connected genes (Serin et al., 2016) and single cell expression simulation (Tian et al., 2021). The pipeline to construct gene co-expression networks usually consists of two steps (Zhang and Horvath, 2005). In step one, we adopt a similarity measure (e.g., the absolute value of Pearson correlation) and calculate the similarity for all pairs of genes. In step two, we choose a threshold and genes with similarity larger than the threshold are thought of as co-expressed. Following the pipeline, Dai et al. (2019) proposed a hypothesis testing based approach to estimate cell-specific gene co-expression network, which is a breakthrough from "cell-type-specific" to "cell-specific" since most computational network methods for single-cell expression are restricted to a group of cells and ignore cell heterogeneity. Li L. et al. (2020) extends the approach to a conditional cell-specific network situation. Unfortunately, the method (Dai et al., 2019) does not incorporate the spatial information of cells and thus may lose power in estimating cell-specific gene co-expression structures, let alone carry out network prediction given a new cell location in the tissue.

To overcome the challenges, we present an easy-to-implement and computationally efficient two-step algorithm to recover cell-specific gene co-expression networks for single-cell spatial expression data. The input of the proposed algorithm is comprised of the spatial locations of cells, cell labels, as well as the gene-cell expression matrix (**Figure 1A**). If cell label information is not available, we can first carry out clustering using single-cell expression data clustering tools (Butler et al., 2018; Stuart et al., 2019). In step one, we estimate the sample expression covariance



	Cell 1	Cell 2	Cell 3	...	Cell n
Gene 1	X_{11}	X_{12}	X_{13}	...	X_{1n}
Gene 2	X_{21}	X_{22}	X_{23}	...	X_{2n}
Gene 3	X_{31}	X_{32}	X_{33}	...	X_{3n}
...	...	...	...	...	...
Gene G	X_{G1}	X_{G2}	X_{G3}	...	X_{Gn}

Edge	Present edge number	Absent edge number	Edge presence in prediction
(1,2)	5	0	✓
(1,3)	1	4	×
(1,4)	4	1	✓
(2,3)	3	2	✓
(2,4)	1	4	×
(3,4)	2	3	×

FIGURE 1 | An illustration of the two-step algorithm. **(A)** The input of the algorithm, including spatial coordinates of cells, the gene-cell expression matrix, and cell class information. Different shapes and colors represent different types of cells. **(B)** The two steps in the proposed algorithm. We first estimate cell-type covariance matrices, and then we use those estimates to refine cell-specific gene co-expression networks. **(C)** Gene co-expression network prediction based on the estimates from **(B)**.

matrix for each cell type, which serves as the "average" of the cell-specific covariance matrices in a given cell type (**Figure 1B**). In step two, for any given cell, we find its appropriate neighborhood and combine the cell label proportions in the neighborhood and the cell-type covariance matrices estimated in step one to assign an empirical prior to the covariance matrix of that cell. Subsequently, we apply the Bayes' rule to obtain the posterior mean estimates, transform it to the correlation matrix, and select a threshold to shrink absolute values of correlations less than it to zero, resulting in the cell's gene co-expression network (**Figure 1B**). After completing the estimates for each cell, we can further predict the network structures for a position where cells are not detected. We set a neighborhood of the location like in the estimation step two, and then an edge is present if and only if this edge appears more than or equal to half times among the gene networks of its neighboring cells (**Figure 1C**).

In the following, we introduce our proposed algorithm in detail in section 2. Section 3 provides the simulation study to compare the two-step algorithm against competing methods including traditional network construction methods (Zhang and Horvath, 2005) based on a group of cells and the cell-specific network construction approach (Dai et al., 2019). We use MERFISH data to demonstrate the good utility of the algorithm in section 4 and conclude the paper with a discussion in section 5.

2. METHOD

We first give some notations to clearly express the data preprocessing and our algorithm. Suppose that expression levels of G genes in n cells are measured and the expression of gene g in cell i is denoted by X_{gi}. We let $\mathbf{X} = (X_{gi})_{G\times n}$ represent the gene-cell expression matrix and use $\mathbf{X}_i$ to denote the *ith* column vector. The coordinates of cell i in the tissue section are denoted by (ℓ_i, h_i). We further assume that cells are from K distinct cell types and C_i indicates the membership of cell i. In other words, $C_i = k$ $(k = 1, \ldots, K)$ implies that cell i belongs to cell type k. Notice that the cell labels $\mathbf{C} = (C_1, \ldots, C_n)$ are assumed to be known in advance, and in case the cell label information is not available we can cluster cells using off-the-shelf single-cell expression tools. n_k is the cell number in cell type k, and $\mathbf{S}_k$ represents the index set $\{i : C_i = k\}$.

During data preprocessing, we need to normalize raw read count data to reduce the effects of different library sizes and other systematic biases. As we are interested in the pairwise gene correlations, the normalized expression values are further centered to zero and scaled to variance one within each cell type. If we still use X_{gi} to represent the normalized expression, then the transformed value is as follows. When $C_i = k$,

$$\widetilde{X}_{gi} = \frac{X_{gi} - \frac{1}{n_k}\sum_{j\in\mathbf{S}_k} X_{gj}}{\sqrt{\frac{1}{n_k-1}\sum_{j\in\mathbf{S}_k}(X_{gj} - \frac{1}{n_k}\sum_{j\in S_k} X_{gj})^2}}.$$

Next, we utilize the scaled expression matrix $\widetilde{\mathbf{X}} = (\widetilde{X}_{gi})_{G\times n}$ and its *ith* column vector $\widetilde{\mathbf{X}}_i$ in our algorithm.

In step one, we derive the sample expression covariance matrix for each cell type, which serves as the "average" of all cell-specific expression covariance matrices in that cell type and hence can be treated as an initial and coarse-grained estimate of the expression covariance matrix for each cell. Specifically, for cell type k, its sample expression covariance matrix is estimated by $\widehat{\mathbf{\Sigma}}^{(k)} := \frac{1}{n_k-1}\sum_{i\in\mathbf{S}_k}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^T$ $(1 \le k \le K)$.

In step two, suppose the gene expression covariance matrix of cell i is denoted by $\mathbf{\Sigma}_i$. Biologically, $\mathbf{\Sigma}_i$ depends on both cell i's cell type as well as cell i's spatial circumstances. Taking this into account, we assume the following Bayesian statistical model for the observations,

$$\widetilde{\mathbf{X}}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_i) \quad (1)$$

$$\mathbf{\Sigma}_i \sim \mathcal{W}^{-1}(\mathbf{\Psi}_i, \nu), \quad (2)$$

where $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_i)$ is a multivariate normal distribution with mean vector zero and covariance matrix Σ_i, and $\mathcal{W}^{-1}(\mathbf{\Psi}_i, \nu)$ is an inverse-Wishart distribution with scale matrix $\mathbf{\Psi}_i$ and ν degrees of freedom.

Equation (1) corresponds to the data-generating mechanism in which cell i's observation is sampled from its own distribution parameterized by $\mathbf{\Sigma}_i$. In the normal distribution, a zero element in $\mathbf{\Sigma}_i$ indicates that the corresponding two genes are independent, so $\mathbf{\Sigma}_i$ fully captures the gene co-expression network structure of cell i. Equation (2) reflects that we need to provide prior information for $\mathbf{\Sigma}_i$ to stabilize the estimate of $\mathbf{\Sigma}_i$; otherwise, only one sample is available, making the common maximal likelihood estimate very sensitive. We employ the inverse-Wishart distribution here as it is conjugate to the multivariate normal distribution (Gelman et al., 2013), which can enhance fast calculation of posterior estimates. Accordingly, we aim to borrow information from cell i's neighbors to define the hyper-parameter in the prior—the scale matrix $\mathbf{\Psi}_i$.

For each cell, we define its neighborhood as a square region with side length $2r$ and center at the location of the cell (**Figure 1B**). The choice of r depends on the cell density in the tissue section and our knowledge about the number of informative neighboring cells. We define the cell density as the ratio of the cell number (n) to the area where cells locate (A). As the area shape is often like a rectangle, we estimate A by $\hat{A} := (\max_i \ell_i - \min_i \ell_i)(\max_i h_i - \min_i h_i)$. If we believe that on average each cell has m_{info} informative neighboring cells, we then have the relationship $n/\hat{A} \times 4r^2 = m_{info}$, leading to $r = 0.5\sqrt{m_{info}\hat{A}/n}$. Based on our experience, we set $m_{info} = 70$ throughout our paper. Subsequently, we count the number of cells in this square region for each cell type and calculate proportions $(\omega_{i1}, \ldots, \omega_{iK})$ with $\omega_{ik} \ge 0$ and $\sum_{k=1}^K \omega_{ik} = 1$, where ω_{ik} is the proportion of type k cells in the neighborhood of cell i.

Next, we assign the weighted value $\sum_{k=1}^K \omega_{ik}\widehat{\mathbf{\Sigma}}^{(k)}$ to the prior mean of $\mathbf{\Sigma}_i$, which is $\mathbf{\Psi}_i/(\nu - G - 1)$, resulting in the scale matrix $\mathbf{\Psi}_i = (\nu - G - 1)\sum_{k=1}^K \omega_{ik}\widehat{\mathbf{\Sigma}}^{(k)}$. This prior reflects the information of nearby cells and helps stabilize the estimate of $\mathbf{\Sigma}_i$. We remark that the choice of the hyper-parameter $\mathbf{\Psi}_i$ depends on the data we are analyzing, so strictly speaking the approach is not fully Bayesian (Gelman et al., 2013).

Given the assigned prior, we estimate $\mathbf{\Sigma}_i$ by the posterior mean,

$$\widehat{\mathbf{\Sigma}}_i := \mathbb{E}(\mathbf{\Sigma}_i|\widetilde{\mathbf{X}}_i) = \frac{1}{\nu - G}(\mathbf{\Psi}_i + \widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^T)$$
$$= \frac{1}{\nu - G}((\nu - G - 1)\sum_{k=1}^{K}\omega_{ik}\widehat{\mathbf{\Sigma}}^{(k)} + \widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^T),$$

where we set ν to $2G$ depending on the number of genes. $\widehat{\mathbf{\Sigma}}_i$ is then transformed to its corresponding correlation matrix $\widehat{\mathbf{R}}_i = \text{diag}(\widehat{\mathbf{\Sigma}}_i)^{-1/2}\widehat{\mathbf{\Sigma}}_i\text{diag}(\widehat{\mathbf{\Sigma}}_i)^{-1/2}$, where $\text{diag}(\widehat{\mathbf{\Sigma}}_i)$ is a diagonal matrix with diagonal elements the same as those of $\widehat{\mathbf{\Sigma}}_i$. Finally, we select a threshold d ($0 < d < 1$), and if the (g_1, g_2) element of the matrix $\widehat{\mathbf{R}}_i$, $\widehat{R}_{i,g_1g_2}$, has an absolute value larger than d, then we believe there is an edge between gene g_1 and g_2 in the gene co-expression network of cell i. Algorithm 1 displays the two-step estimation procedure.

Algorithm 1: Two-step gene co-expression network estimation.

Input: normalized gene expression matrix $\mathbf{X}$, cell labels $\mathbf{C}$, cell coordinates (ℓ_i, h_i), $1 \le i \le n$, and hyper-parameters (m_{info}, ν, d).
Output: cell-specific gene co-expression networks $\mathcal{G}_i$ $(1 \le i \le n)$.
Preprocessing:
for i *in* $1 : n$ **do**
  for g *in* $1 : G$ **do**
    $\widetilde{X}_{gi} = \frac{X_{gi} - \frac{1}{n_k}\sum_{j\in\mathbf{S}_k} X_{gj}}{\sqrt{\frac{1}{n_k-1}\sum_{j\in\mathbf{S}_k}(X_{gj} - \frac{1}{n_k}\sum_{j\in\mathbf{S}_k} X_{gj})^2}}$ when $C_i = k$
  end
end
Step 1: Obtain cell-type-specific covariance matrix:
for k *in* $1 : K$ **do**
  $\widehat{\mathbf{\Sigma}}^{(k)} = \frac{1}{n_k - 1}\sum_{i\in\mathbf{S}_k}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^T$
end
Step 2: Estimate cell-specific gene co-expression networks.
for i *in* $1 : n$ **do**
  $\widehat{\mathbf{\Sigma}}_i = \frac{1}{\nu - G}((\nu - G - 1)\sum_{k=1}^{K}\omega_{ik}\widehat{\mathbf{\Sigma}}^{(k)} + \widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^T)$
  $\widehat{\mathbf{R}}_i = \text{diag}(\widehat{\mathbf{\Sigma}}_i)^{-1/2}\widehat{\mathbf{\Sigma}}_i\text{diag}(\widehat{\mathbf{\Sigma}}_i)^{-1/2}$
  for g_1 *in* $1 : G$ **do**
    for g_2 *in* $(g_1 + 1) : G$ **do**
      $\mathcal{G}_{i,g_1g_2} = \begin{cases} 0 \text{ if } |\widehat{\mathbf{R}}_{i,g_1g_2}| < d \\ 1 \text{ if } |\widehat{\mathbf{R}}_{i,g_1g_2}| \ge d \end{cases}$
    end
  end
end

After completing the network structure estimates for all cells, we can take advantage of the estimates to predict the gene co-expression network for any missing cell with a position in the studied tissue section area. If we are interested in an undetected cell at a new location (ℓ^*, h^*), its gene co-expression network is constructed as follows. We first find all detected cells in the neighborhood of (ℓ^*, h^*), and then we believe an edge between genes g_1 and g_2 in the prediction if there are more connections than disconnections for this pair of genes among the gene networks of (ℓ^*, h^*)'s neighboring detected cells. Algorithm 2 shows the steps of making gene co-expression network predictions.

Algorithm 2: Gene co-expression network prediction for a new cell position.

Input: Gene network estimates from Algorithm 1, cell coordinates (ℓ_i, h_i) for $1 \le i \le n$, hyper-parameter m_{info}, and a new cell position (ℓ^*, h^*).
Output: Gene co-expression network $\mathcal{G}_{(\ell^*,h^*)}$ for cell (ℓ^*, h^*).
Step 1: Find all cells in the neighborhood of (ℓ^*, h^*), denoted by $\text{Nei}_{(\ell^*,h^*)} := \{i \in \{1, 2, \ldots, n\} : (\ell_i, h_i)$ is in the neighborhood of $(\ell^*, h^*)\}$.
Step 2: Obtain gene co-expression network for cell (ℓ^*, h^*):
for g_1 *in* $1 : G$ **do**
  for g_2 *in* $(g_1 + 1) : G$ **do**
    $\mathcal{G}_{(\ell^*,h^*),g_1g_2} =$
    $\begin{cases} 0 \text{ if } \#\{j \in \text{Nei}_{(\ell^*,h^*)} : \mathcal{G}_{j,g_1g_2} = 0\} > \#\{j \in \text{Nei}_{(\ell^*,h^*)} : \mathcal{G}_{j,g_1g_2} = 1\} \\ 1 \text{ if } \#\{j \in \text{Nei}_{(\ell^*,h^*)} : \mathcal{G}_{j,g_1g_2} = 0\} \le \#\{j \in \text{Nei}_{(\ell^*,h^*)} : \mathcal{G}_{j,g_1g_2} = 1\} \end{cases}$
  end
end

3. SIMULATION STUDY

In this section, we used simulated data to evaluate the performance of the proposed two-step algorithm. We set the gene number $G = 100$, the cell-type number $K = 5$, and the cell number for each cell type $(n_1, n_2, n_3, n_4, n_5) = (394, 373, 428, 274, 529)$. We chose a rectangle area as the tissue section with length $L = 1{,}000$ and width $H = 750$, where a total of $n = \sum_{k=1}^{K} n_k = 1998$ cells distribute on the section and display clear spatial patterns (**Figure 2A**). For example, cells from cell-type 5 concentrate on the left side, while cells from cell-type 1 enrich on the right side.

We then generated cell-type-specific covariance matrices $\Sigma^{(k)}$ for $k = 1, \ldots, K$. Genes that work together often form a gene module, which can exhibit a block structure in the covariance matrix. Hence, the covariance matrix of each cell type was set as a block diagonal matrix, where each block was a 20×20 positive definite matrix. Five different modules were used for this purpose and were as follows.

- In module 1 ($\mathcal{M}_1$), its (i, j) element $\sigma_{ij} = \rho^{|i-j|} + 0.5\mathbf{I}(i = j)$ for $1 \le i \le 20$ and $1 \le j \le 20$, where $\mathbf{I}(A)$ is an indicator function of event A. We took $\rho = 0.7$.

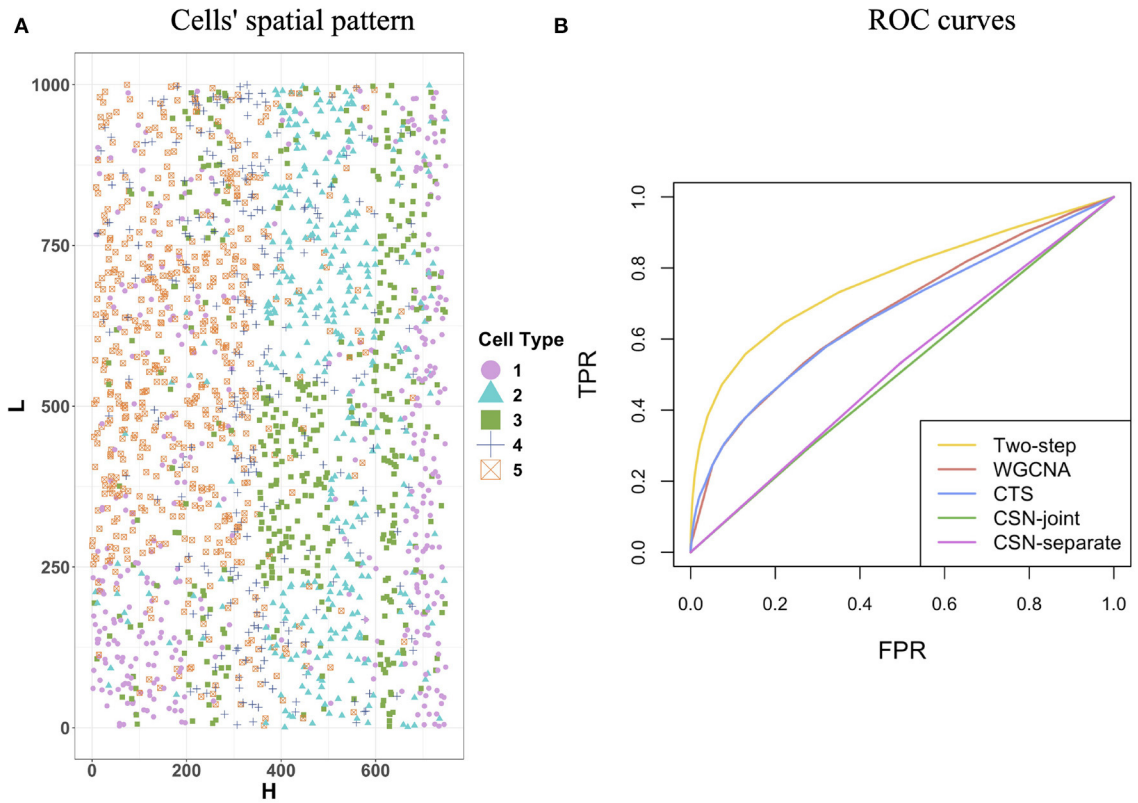


FIGURE 2 | Cells' spatial pattern and performance comparisons on the network structure recovery. **(A)** Cells' spatial pattern. Different shapes and colors correspond to different types of cells. **(B)** ROC curves of the two-step algorithm, WGCNA, CTS, CSN-joint, and CSN-separate.

- In module 2 ($\mathcal{M}_2$), $\sigma_{ij} = (1 - \frac{|i-j|}{10})_+$, which forms a banded matrix. The function $(x)_+$ equals x for $x \geq 0$ and zero for $x < 0$.
- In module 3 ($\mathcal{M}_3$), $\sigma_{ij} = \rho \mathbf{I}(|i-j| = 1) + 1.3\mathbf{I}(i = j)$ for $\rho = -0.3$.
- In module 4 ($\mathcal{M}_4$), $\sigma_j = (1 - \frac{|i-j|}{k})_+$, where $k = \lfloor G/2 \rfloor$.
- In module 5 ($\mathcal{M}_5$), the block was $F + \epsilon I_{20\times 20}$. $I_{20\times 20}$ is an identity matrix. $F = (f_{ij})_{20\times 20}$ is a symmetric matrix with independent upper triangle elements $f_{ij} = unif(-0.2, 0.8) \times Ber(1, 0.2)$, where $unif(-0.2, 0.8)$ is a random variable uniformly distributed on$(-0.2, 0.8)$, and $Ber(1, 0.2)$ is a Bernoulli random variable with the success probability 0.2. We set $\epsilon = \max\{-\lambda_{\min}(F), 0\} + 0.01$ to ensure that B is positive definite, where $\lambda_{\min}(F)$ is the smallest eigenvalue of F.

If we denote a block diagonal matrix with diagonal blocks being $\mathcal{M}_{i_1}, \mathcal{M}_{i_2}, \mathcal{M}_{i_3}, \mathcal{M}_{i_4}, \mathcal{M}_{i_5}$ in the order from the upper left to the lower right by $(\mathcal{M}_{i_1}, \mathcal{M}_{i_2}, \mathcal{M}_{i_3}, \mathcal{M}_{i_4}, \mathcal{M}_{i_5})$, then we specify $\mathbf{\Sigma}^{(1)}$=$(\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4, \mathcal{M}_5)$, $\mathbf{\Sigma}^{(2)}$=$(\mathcal{M}_1, \mathcal{M}_3, \mathcal{M}_2, \mathcal{M}_4, \mathcal{M}_5)$, $\mathbf{\Sigma}^{(3)}$=$(\mathcal{M}_1, \mathcal{M}_3, \mathcal{M}_2, \mathcal{M}_5, \mathcal{M}_4)$, $\mathbf{\Sigma}^{(4)}$=$(\mathcal{M}_3, \mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_5, \mathcal{M}_4)$, and $\mathbf{\Sigma}^{(5)}$=$(\mathcal{M}_3, \mathcal{M}_2, \mathcal{M}_5, \mathcal{M}_1, \mathcal{M}_4)$.

Next, we generated the cell-specific gene expression covariance matrix for each cell i. We first obtained the neighborhood of cell i using $r = 80$, then calculated cell-type proportions $q_{ik}, 1 \leq k \leq K$ in the neighborhood, and sampled Σ_i from the inverse-Wishart distribution $\mathcal{W}^{-1}(\sum_{k=1}^{K} q_{ik}\mathbf{49}\Sigma^{(k)}, G + 50)$. Moreover, to make the network sparse and covariance matrix positive definite, non-diagonal elements in the $\mathbf{\Sigma}_i$ with absolute values less than 0.5 were shrunk to zero, and the diagonal elements in Σ_i were added by five. Finally, we sampled the observed gene-cell expression matrix $\mathbf{X}_i = (X_{1i}, \ldots, X_{Gi})^T$ from the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \Sigma_i)$ for $1 \leq i \leq n$.

To show the advantage of our algorithm in estimating cell-specific gene expression matrix, we compared it against the weighted gene co-expression network analysis (denoted by WGCNA) (Zhang and Horvath, 2005), the traditional hard-thresholding cell-type-specific network estimation approach (denoted by CTS), and the cell-specific gene network estimation method that does not make use of cell spatial information (denoted by CSN, Dai et al., 2019). Specifically, in WGCNA, we first calculated pairwise gene expression similarity using the absolute values of Pearson correlations, then utilized the "soft" power adjacency function to convert the similarity matrix, and finally obtain the topological overlap matrix based on the adjacency matrix. Regarding CTS, we used the cell-type-level gene network as the estimate for each cell in that cell type. For CSN, we adopted two versions: in the joint version (CSN-joint), we used the gene-cell expression matrix for all cells as the input of the CSN method; and in the separate version (CSN-separate), we

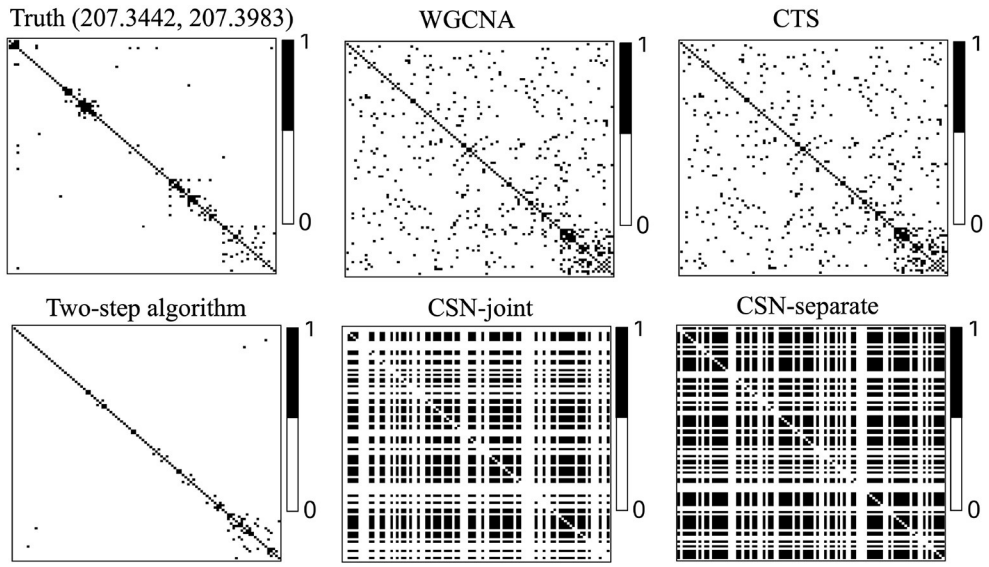


FIGURE 3 | Heatmaps of estimated gene co-expression networks for different methods and underlying truth. The network heatmaps of the cell on location (207.3442, 207.3983) was used for illustration.

only input the gene-cell expression matrix for cells coming from one cell type, repeat the procedure for each cell type, and also obtain cell-specific network estimates. In other words, for CSN-separate, the estimations for one cell only rely on the information of cells from the same cell type.

Figure 2B provides the receiver operating characteristic (ROC) curves for network structure recovery of the proposed algorithm (denoted by two-step algorithm) and other four competing approaches (WGCNA, CTS, CSN-joint, CSN-separate). The horizontal axis represents the false positive rate (FPR), which equals the ratio of the number of edges that were wrongly detected by the method for all cells to the number of absent edges in the underlying true networks for all cells, while the vertical axis corresponds to the true positive rate (TPR), describing the ratio of the number of edges that were correctly detected by the method for all cells to the number of edges in the underlying true networks for all cells. It is observed that the ROC curve of our algorithm is uniformly over the ROC curves of the other four approaches, indicating that given any FPR the TPR of the proposed algorithm is always higher than that of the other four competing methods. As WGCNA also estimates cell-type-specific networks, it does not outperform our algorithm but is slightly better than traditional CTS.

Figure 3 displays heatmaps of gene co-expression matrix of the cell with the coordinates (207.3442, 207.3983), both true and estimated gene co-expression matrix by two-step algorithm, WGCNA, CTS, CSN-joint, and CSN-separate are shown (the results of CSN-separate are similar to CSN-joint's). From **Figure 3**, we can observe that our two-step algorithm outperforms the other four methods in estimating cell-specific gene co-expression networks. To further quantify the network recovery error for these methods, we used the following error term $E := 1/n \cdot \sum_{i=1}^{n} \sum_{g_1<g_2} |\mathcal{G}_{i,g_1g_2} - \mathcal{G}^{\text{true}}_{i,g_1g_2}|$. For WGCNA, we chose the truncation value 0.0001 for the topological overlap matrix; for the proposed algorithm and CTS, the threshold d for the gene-gene correlations was chosen as 0.1; for the two CSN methods, the significance level was set at 0.01. **Table 1** shows the errors based on ten replicates and indicates that the proposed method is more accurate than the others in terms of the network structure recovery.

TABLE 1 | Mean errors and corresponding standard deviations of five methods.

Methods	Two-step algorithm	WGCNA	CTS	CSN-joint	CSN-separate
Mean error	288.62	484.05	484.44	1869.60	2462.09
(standard deviation)	(14.15)	(18.78)	(18.80)	(347.49)	(95.57)

The degree of a gene is the number of edges connected to that gene. We investigated the degree distributions of the estimated cell-specific gene co-expression network and compared it to truth and other competing approaches on one gene for each cell type. **Figures 4A–E** show the violin plots of the degrees of gene 91 for each cell type. We can see that the distribution created by our proposed algorithm is much closer to the underlying truth than CSN-separate and CSN-joint, while WGCNA and CTS's distributions are just horizontal line segments as their network estimates are identical for all cells in one cell type.

Figure 4F shows the violin plot for the computation time in second for these methods based on ten replicates. It is reasonable that WGCNA and CTS have the minimum computing time as they only estimate K cell-type-specific gene-gene network, but their performances are obviously not good. The proposed algorithm has a similar computing time to CSN-separate

and is faster than CSN-joint. Hence, our algorithm not only performs well in estimating networks but also has relatively fast computing.

Given the network estimates by our method, we can easily use algorithm 2 to predict network structures for a new location. We randomly generated 50 new coordinates as the locations of 50 missing cells, simulated the true gene network of these 50 new cells following the data-generating procedure above, and then applied the prediction algorithm. The prediction error is 347.84 (in terms of E). WGCNA, CTS, CSN-joint, and CSN-separate do not have the ability to predict gene co-expression networks of missing cells, so the proposed algorithm provides an extra important function to make network predictions.

4. REAL APPLICATION

4.1. MERFISH Mouse Hypothalamus Data

Moffitt et al. (2018) combined single-cell RNA-sequencing and a single-cell transcriptome imaging method called MERFISH to obtain expression profiles at the cellular level as well as x-y coordinates of centroid positions for cells in the mouse hypothalamic preoptic region. In the MERFISH mouse hypothalamus data, class information of cells are also available. The single-cell spatial expression data can be downloaded from https://datadryad.org/stash/dataset/doi:10.5061/dryad.8t8s248.

We chose the expression data with animal id 35 and location 0.26 of the slice in bregma coordinates and removed cells labeled

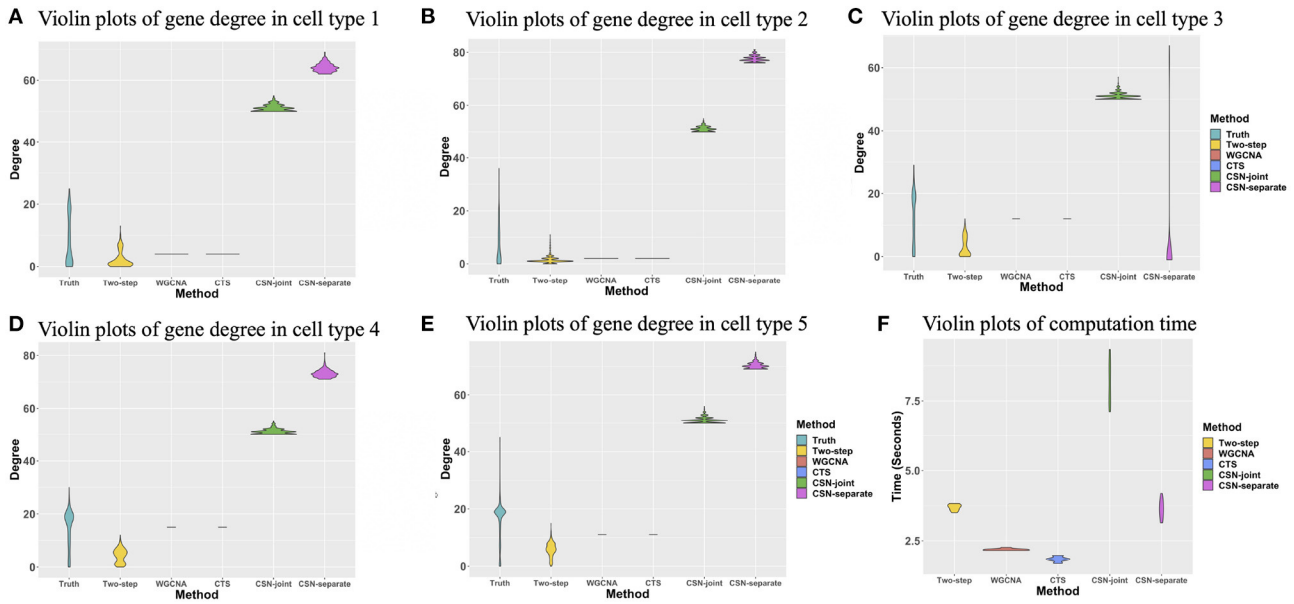


FIGURE 4 | (A–E) Violin plots of gene degrees of gene 91 in the five cell types in the simulation study. **(F)** Violin plots of the computational time (in seconds) for the five approaches based on ten replicates.

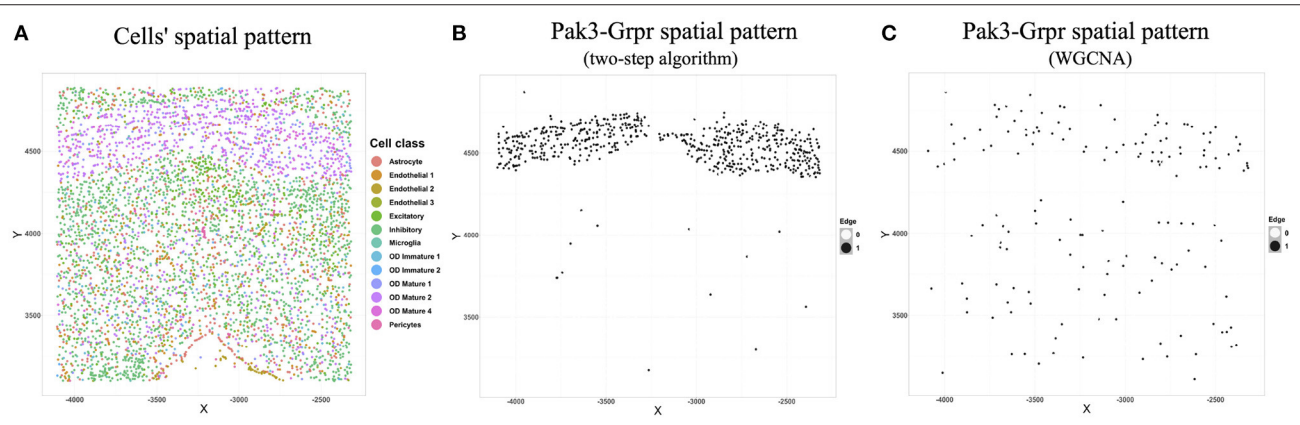


FIGURE 5 | (A) Cells' spatial distribution pattern, where different colors of points correspond to different cell classes. **(B)** The spatial pattern of the Pak3-Crpr connection obtained by two-step algorithm. The black point reflects that an edge exists between "Pak3" and "Crpr" in that cell. **(C)** The spatial pattern of the Pak3-Crpr connection obtained by WGCNA.

"Ambiguous" as well as cell types that contain less than 10 cells, resulting in 13 cell classes. The spatial pattern of the selected cells was displayed in **Figure 5A**. We further removed "blank" genes and genes whose expressions are zero across all the cells in one cell type, resulting in $G = 147$ genes and $n = 4,682$ cells. Subsequently, we applied the proposed two-step algorithm with informative neighboring cell number $m_{info} = 70$ and threshold parameter $d = 0.1$. We randomly selected two cells from cell classes "inhibitory neurons" and "excitatory neurons," respectively, and the gene co-expression networks of the two cells were shown in **Figure 6**. It is observed that the two gene co-expression networks have similar functional gene modules on the

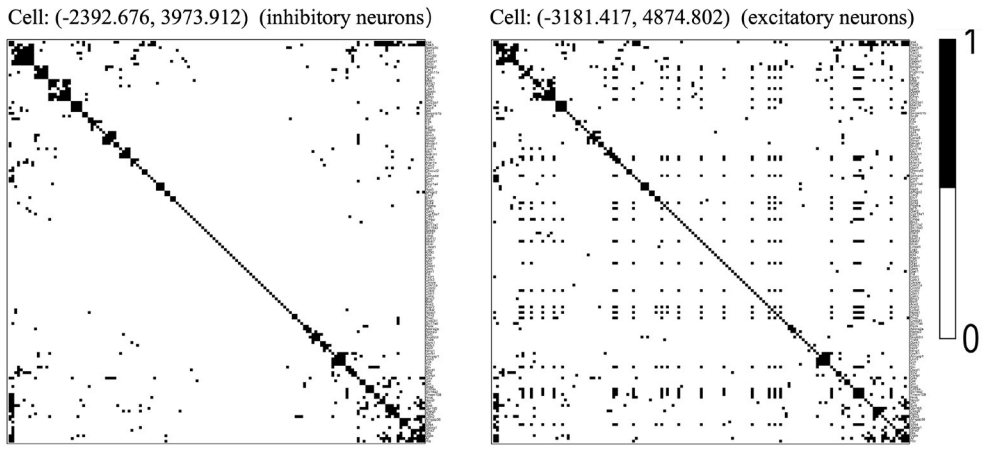


FIGURE 6 | Estimated gene co-expression networks of two selected cells in the MERFISH mouse hypothalamus data: the left panel corresponds to an inhibitory cell, while the right panel corresponds to an excitatory cell.

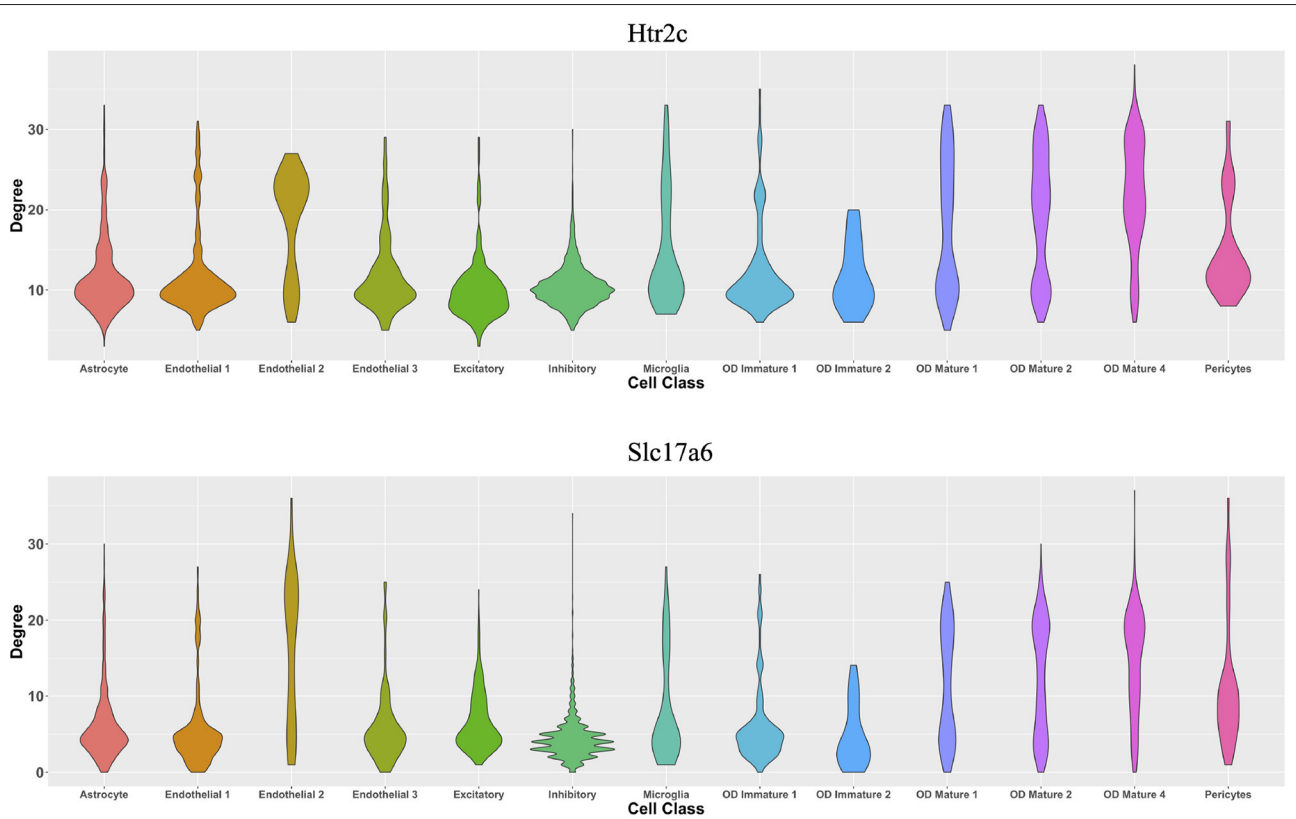


FIGURE 7 | Violin plots of two genes' degree distributions across thirteen cell classes in the MERFISH mouse hypothalamus data.

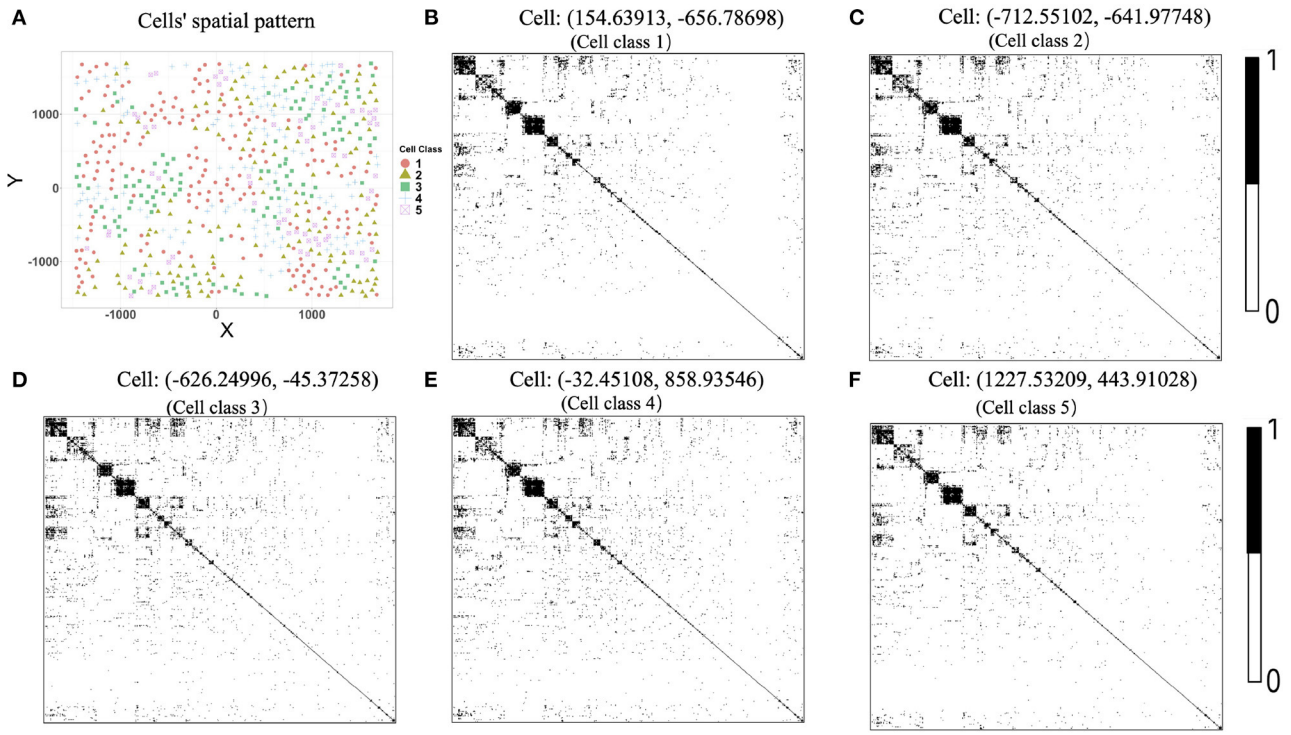


FIGURE 8 | (A) Cells' spatial distribution pattern, where different colors and shapes of points correspond to different cell classes. **(B–F)** Estimated gene co-expression networks of five randomly selected cells in the MERFISH U-2 OS cell line data.

diagonal possibly because both of them are neurons. Moreover, the network of the cell in excitatory neurons is denser than the network in inhibitory neurons, and the reason may be that the gene activity in cells controlling excitement is more active than that in cells controlling inhibition.

Cell-specific gene co-expression networks can provide insightful information about how genes' degrees vary in each cell type. To show that, in excitatory neuron cells, we selected 15 genes with the most variable degrees: Sln, Baiap2, Tmem108, Oprk1, Slc17a6, Nos1, Htr2c, Irs4, Gpr165, Slc18a2, Vgf, Pgr, Ar, Gabrg1, and Gabra1. To validate the functions of the gene set, we conducted gene set enrichment analysis (Subramanian et al., 2005) based on the gene ontology (GO) database (Gene Ontology Consortium, 2004). We found several significant annotations related to the excitatory neurons including GO_MODULATION_OF_EXCITATORY_POSTSYNAPTIC_POTENTIAL(biologicalprocess),GO_EXCITATORY_SYNAPSE (cellular component), and GO_NEURON_PROJECTION (cellular component). In terms of the inhibitory neurons, we identified 15 genes with the most variable degrees: Baiap2, Sox6, Irs4, Ar, Gda, Oprk1, Isl1, Cyr61, Prlr, Glra3, Gabra1, Dgkk, Tmem108, Sln, and Ano3. Using GO annotations, the gene set is associated with inhibitory neurons-related activities including GO_INHIBITORY_EXTRACELLULAR_LIGAND_GATED_ION_CHANNEL_ACTIVITY (molecular function) and GO_NEURON_PROJECTION (cellular component). These observations show that estimated cell-specific networks have the potential to find genes with variable degrees for each cell type, which cannot be accomplished by cell-type-specific approaches.

We next illustrated the spatial feature of estimated gene co-expression networks in terms of gene-gene connections. We calculated the median degree for each gene. Gene Pak3 with the maximum median degree (31) and gene Grpr with the minimum median degree (0) were chosen for demonstration. **Figure 5B** shows that the Pak3-Grpr connection mainly appears in the region where "mature oligodendrocytes" are enriched. The observation indicates that the two genes may tend to work together in the mature oligodendrocytes. Actually, mutations on gene Pak3 are related to intellectual disability diseases, and its expression decreases in mature oligodendrocytes and may regulate oligodendrocyte precursor cell differentiation, as reported in a previous study (Renkilaraj et al., 2017). To demonstrate the advantage of estimating cell-specific networks, we further applied WGCNA (Zhang and Horvath, 2005) with truncation level 0.1 to obtain cell-type-specific networks. However, **Figure 5C** indicates that the cell-type-specific estimations by WGCNA cannot reveal the pattern provided by cell-specific estimations.

From the perspective of cell types, **Figure 7** demonstrates the cell-type-specific degree distributions of two genes, Htr2c and Slc17a6 (Campbell et al., 2017; Chen et al., 2017), which have the most degree variances across cells. It is observed that the degree distribution of one gene varies across cell types, and

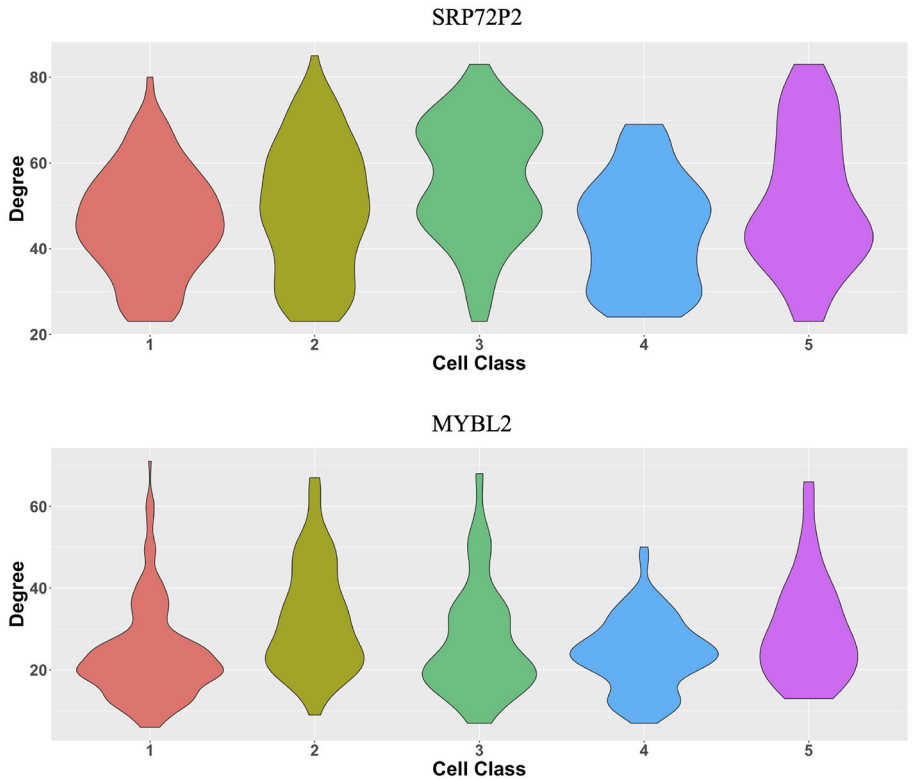


FIGURE 9 | Violin plots of two genes' degree distributions across five cell classes in the MERFISH U-2 OS cell line data.

this cannot be observed by traditional cell-type-specific gene co-expression networks.

4.2. MERFISH U-2 OS Data

We further provided some simple results of the proposed algorithms on another single-cell spatial expression dataset. Xia et al. (2019) carried out the MERFISH experiments on human osteosarcoma (U-2 OS) cells, and we downloaded the expression count data from https://www.pnas.org/content/116/39/19490/tab-figures-data. The data contain expression profiles for 10,050 genes and 1,368 cells in three batches. To avoid possible influences caused by batch effects, our analysis focuses on the batch one. We first removed "blank" genes, resulting in $n = 645$ cells and $G = 10,050$ genes. Since there is no cell-type annotation information, we first performed cell clustering procedure using Seurat (Butler et al., 2018; Stuart et al., 2019). By setting the resolution at 0.8 in Seurat clustering procedure, we obtained $K = 5$ cell classes, which is consistent with the cell type number in Xia et al. (2019). **Figure 8A** shows the cells' spatial distribution.

The original expression data were count data, so we normalized the data following the formula $x_{gi} \leftarrow \frac{10^6}{\sum_g x_{gi}} x_{gi}$, where x_{gi} is the expression level of gene g in cell i and then selected the most variable 500 genes to perform the proposed two-step algorithm. The informative neighboring cell number m_{info} was set to 70, and the threshold parameter d was set to 0.3. Accordingly, we randomly selected five cells from the five cell classes, respectively, and the gene co-expression networks of the five chosen cells were shown in **Figures 8B–F**. It is observed that the five gene networks from different cell types have similar gene modules. Moreover, we showed the degree distributions across five cell types for two genes, SRP72P2 and MYBL2, which have the most degree variances across cells. **Figure 9** tells us that the degree distributions of the two genes not only have variation within one cell type but also change from one cell type to another.

5. DISCUSSION

Recent technology advances enable us to gain deep insights into spatial cell-specific gene expressions. In this paper, we developed a simple and computationally efficient two-step algorithm to recover spatially-varying cell-specific gene co-expression networks. The simulation study shows that the proposed algorithm outperforms the traditional cell-type-specific gene network approach and cell-specific gene network estimation methods that do not employ spatial information. The application to the MERFISH data provides some interesting biological findings. In the meanwhile, there are some limitations in the proposed

algorithm we aim to improve in the future work. For example, we choose a hard threshold to identify a gene-gene connection, but an adaptive threshold selection needs to be derived.

We also acknowledge that using normal distributions to fit normalized gene expression data can lose power and be suboptimal compared to directly modeling the sequencing count data via Poisson distributions (Sun et al., 2017). Fortunately, in several previous bioinformatics works, using continuous multivariate normal distributions to model normalized single-cell sequencing data (Pierson and Yau, 2015; Chen and Zhou, 2017; Wang et al., 2020) or spatial single-cell expression data (Li D. et al., 2020) can still provide key biological findings. Moreover, in terms of computation, multivariate Poisson distributions (Karlis, 2003) largely increase the computational burden. Statistically, the covariance matrix in the multivariate Poisson distribution does not have a standard conjugate prior, thus failing to obtain an analytical form of the posterior mean. In real data, the cell number is often large (∼4,000 in our real application), which actually guarantees a satisfying normal approximation. Considering these issues, we chose the multivariate normal as the data distribution, but it is very interesting and challenging to extend the algorithm to directly model raw count data and we leave it for future work.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. MERFISH mouse hypothalamus data can be downloaded from https://datadryad.org/stash/dataset/doi:10.5061/dryad.8t8s248, and MERFISH U-2 OS cell line data is available via the link https://www.pnas.org/content/116/39/19490/tab-figures-data.

CODE AVAILABILITY STATEMENT

The codes that can reproduce results in simulation and real application are available on GitHub, https://github.com/jingeyu/CSSN_data_code. The associated CSSN package is available on GitHub, https://github.com/jingeyu/CSSN.

AUTHOR CONTRIBUTIONS

XL conceived the study. JY and XL developed the method, analyzed the real data, and wrote the paper. JY implemented the algorithm, prepared the software, and conducted simulation. Both authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by National Natural Science Foundation of China (11901572), the start-up research fund at Renmin University of China, the Fundamental Research Funds for the Central Universities, the Research Funds of Renmin University of China (19XNLG08), and the fund for building world-class universities (disciplines) of Renmin University of China.

ACKNOWLEDGMENTS

We are very grateful to the Editor, Associate Editor, and reviewers for their constructive comments which greatly improve the paper. We also thank the High-performance Computing Platform of Renmin University of China for providing computing resources.

REFERENCES

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420. doi: 10.1038/nbt.4096

Butte, A., and Kohane, I. (2000). "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (Honolulu, HI), 418–429.

Campbell, J. N., Macosko, E. Z., Fenselau, H., Pers, T. H., Lyubetskaya, A., Tenen, D., et al. (2017). A molecular census of arcuate hypothalamus and median eminence cell types. *Nat. Neurosci.* 20, 484–496. doi: 10.1038/nn.4495

Carter, S. L., Brechbühler, C. M., Griffin, M., and Bond, A. T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* 20, 2242–2250. doi: 10.1093/bioinformatics/bth234

Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., and Zhuang, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 348:aaa6090. doi: 10.1126/science.aaa6090

Chen, M., and Zhou, X. (2017). Controlling for confounding effects in single cell RNA sequencing studies using both control and target genes. *Sci. Rep.* 7, 1–14. doi: 10.1038/s41598-017-13665-w

Chen, R., Wu, X., Jiang, L., and Zhang, Y. (2017). Single-cell RNA-seq reveals hypothalamic cell diversity. *Cell Rep.* 18, 3227–3241. doi: 10.1016/j.celrep.2017.03.004

Dai, H., Li, L., Zeng, T., and Chen, L. (2019). Cell-specific network constructed by single-cell RNA sequencing data. *Nucleic Acids Res.* 47:e62. doi: 10.1093/nar/gkz172

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press.

Gene Ontology Consortium (2004). The gene ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32(Suppl_1), D258–D261. doi: 10.1093/nar/gkh036

Karlis, D. (2003). An EM algorithm for multivariate Poisson distribution and related models. *J. Appl. Stat.* 30, 63–77. doi: 10.1080/0266476022000018510

Köster, J., Brown, M., and Liu, X. S. (2019). A Bayesian model for single cell transcript expression analysis on MERFISH data. *Bioinformatics* 35, 995–1001. doi: 10.1093/bioinformatics/bty718

Lee, J. H., Daugharthy, E. R., Scheiman, J., Kalhor, R., Yang, J. L., Ferrante, T. C., et al. (2014). Highly multiplexed subcellular RNA sequencing *in situ*. *Science* 343, 1360–1363. doi: 10.1126/science.1250212

Li, D., Ding, J., and Bar-Joseph, Z. (2020). Identifying signaling genes in spatial single cell expression data. *Bioinformatics*. doi: 10.1101/2020.07.27.221465. [Epub ahead of print].

Li, L., Dai, H., Fang, Z., and Chen, L. (2020). CCSN: single cell RNA sequencing data analysis by conditional cell-specific network. *bioRxiv [Preprint]*. doi: 10.1101/2020.01.25.919829

Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M., and Cai, L. (2014). Single-cell *in situ* RNA profiling by sequential hybridization. *Nat. Methods* 11:360. doi: 10.1038/nmeth.2892

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

Pierson, E., and Yau, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* 16, 1–10. doi: 10.1186/s13059-015-0805-z

Renkilaraj, M. R. L. M., Baudouin, L., Wells, C. M., Doulazmi, M., Wehrlé, R., Cannaya, V., et al. (2017). The intellectual disability protein PAK3 regulates oligodendrocyte precursor cell differentiation. *Neurobiol. Dis.* 98, 137–148. doi: 10.1016/j.nbd.2016.12.004

Serin, E. A., Nijveen, H., Hilhorst, H. W., and Ligterink, W. (2016). Learning from co-expression networks: possibilities and challenges. *Front. Plant Sci.* 7:444. doi: 10.3389/fpls.2016.00444

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302, 249–255. doi: 10.1126/science.1087447

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M. III., et al. (2019). Comprehensive integration of single-cell data. *Cell* 177, 1888–1902. doi: 10.1016/j.cell.2019.05.031

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Sun, S., Hood, M., Scott, L., Peng, Q., Mukherjee, S., Tung, J., et al. (2017). Differential expression analysis for RNAseq using Poisson mixed models. *Nucleic Acids Res.* 45:e106. doi: 10.1093/nar/gkx204

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

Tian, J., Wang, J., and Roeder, K. (2021). ESCO: single cell expression simulation incorporating gene co-expression. *Bioinformatics*. doi: 10.1093/bioinformatics/btab116. [Epub ahead of print].

Wang, J., Devlin, B., and Roeder, K. (2020). Using multiple measurements of tissue to estimate subject-and cell-type-specific gene expression. *Bioinformatics* 36, 782–788. doi: 10.1093/bioinformatics/btz619

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

Xia, C., Fan, J., Emanuel, G., Hao, J., and Zhuang, X. (2019). Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 116, 19490–19499. doi: 10.1073/pnas.1912459116

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4:Article17. doi: 10.2202/1544-6115.1128

Zhang, M., Sheffield, T., Zhan, X., Li, Q., Yang, D. M., Wang, Y., et al. (2020). Spatial molecular profiling: platforms, applications and analysis tools. *Brief. Bioinform.* doi: 10.1093/bib/bbaa145. [Epub ahead of print].

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.