ISSN 2616-5775 Vol. 7, Issue 5: 123-129, DOI: 10.25236/AJCIS.2024.070516

A Comparison of Spatial Transcriptomics Clustering Methods Based on Three Different Approaches

Jingyao Huo*

School of Statistics, University of International Business and Economics, Beijing, 100029, China
**Corresponding author: 13264169300@163.com*

Abstract: *This study employed spatial transcriptomics and the BayesSpace method to investigate melanoma, enabling accurate classification of tissue cells based on their location. The BayesSpace method, combined with clustering analysis, effectively examined spatial transcriptome data. Data preprocessing and PCA were conducted to reduce dimensionality, followed by clustering using k-means, GMM, and BayesSpace. Among these methods, BayesSpace proved to be the optimal clustering method. Marker gene staining verified the clustering results, demonstrating high accuracy and enabling precise identification of cell types. This study highlights the superiority of the BayesSpace method in spatial transcriptomic analysis and its potential for annotating cell types in biomedical research. The findings contribute to melanoma diagnosis and treatment through the identification of characteristic cells, marker genes, and therapeutic targets.*

Keywords: *Spatial Transcriptomics, BayesSpace Clustering, Melanoma*

1. Introduction

Spatial transcriptomics is a relatively new field that aims to sequence and analyze gene expression levels in relation to the spatial organization of cells within tissues. It provides valuable insights into the spatial distribution of cellular gene expression, allowing for a more comprehensive understanding of tissue biology.

There are several techniques used in spatial transcriptomics[1], including Laser capture microscopy, Fluorescence in Situ hybridization (FISH), Fluorescence in Situ Sequencing (FISSEQ), and spatial transcriptomic techniques based on spatial information capture[2]. Each technique has its advantages and limitations, and researchers choose the most appropriate one based on their specific experimental needs.

One well-known and commercialized technology in spatial transcriptomics is 10x Genomics' Visium technology[3]. In the Visium workflow, tissue slices are obtained and affixed to a glass slide. The slices are then stained to facilitate imaging of the expressed products. Permeabilization treatment is applied to allow mRNA to interact with oligonucleotide chains in the spatial capture region, resulting in cDNA synthesis. This cDNA is then used to construct a library, which is subsequently sequenced. By analyzing the sequencing data, software can determine the original position and gene expression profiles of cells within the tissue.

As the field of spatial transcriptomics has advanced, various institutions and research groups have collected and organized datasets, resulting in databases of significant scale. These datasets typically include image information, positional information (such as rows and columns of the spots), and data on the expression levels of different genes in tissue cells. In this case, relevant datasets that focus on melanoma tissue cells would provide information on the spatial organization of gene expression within melanoma tissue.

K-means and GMM are the two most classic methods and have been validated by previous scholars as having better clustering performance. However, in the clustering process of the above two methods, the location information of cells was not considered, which can easily lead to inaccurate clustering. The BayesSpace method introduces spatial coordinate information, so this article compares the three clustering methods mentioned above and examines the effectiveness of different clustering methods.

ISSN 2616-5775 Vol. 7, Issue 5: 123-129, DOI: 10.25236/AJCIS.2024.070516

2. Methods

2.1 K-means clustering

K-means is a method of clustering a set of samples based on distance measurement, which can be measured by Euclidean distance or statistical distance. It classifies samples that are closer together by calculating the distance between different samples. This method was proposed by James MacQueen in 1967 and is currently widely used in the fields of signal processing and machine learning.

Using the k-means clustering method, different data points are randomly divided into k categories in advance, and then the categories are adjusted one by one based on the similarity between the samples and different clusters; Divide each sample data into the cluster closest to it, and update the centroid mean of the newly obtained cluster; Repeat the previous step until the number of iterations is reached or until the points within each cluster are as closely connected as possible, which is the best clustering result and the final result is obtained. The closeness of samples within a cluster around the mean is defined as follows:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \| x - \mu_i \|_2^2 \tag{1}$$

Among them, k is the number of clusters, and C_i is an element in class i, μ_i is the class mean in class i.

When selecting the number of clusters, elbow method and contour coefficient method can be used.

A. Using the elbow method, it is necessary to calculate the SSE of clustering results for different numbers of clusters k. As the number of clusters increases, the classification becomes finer, and the similarity of sample data within the same cluster increases. Therefore, SSE will inevitably decrease, and its decreasing trend is to first rapidly decrease and then slowly decrease. Therefore, only the inflection point of the descent speed needs to be selected, which is the most suitable number of clusters for the data point situation.

B. To use the contour coefficient method, the Gap statistical method is required:

$$Gap(K) = E(\log D_K) - \log D_K \tag{2}$$

Among them, D_K is the loss function, $E(\log D_K)$ is the expectation of D_K logarithm. The K that maximizes $Gap(K)$ is the optimal number of clusters.

The clustering results based on the k-means method can be obtained by using the determined number of clusters. Based on the above discussion, k-means have the advantages of low computational complexity, intuitiveness, and easy implementation. At the same time, it also makes the selection of initial clustering centers more important and lacks robustness.

2.2 GMM clustering (Gaussian Mixture Model)

The Gaussian Mixture based Model (GMM) considers data as a mixed model of multiple Gaussian distributions, treating different Gaussian distributions in the mixed model as a cluster, and using the maximum likelihood method to determine sample classification based on the probability of different samples being located in each classification cluster. This method was proposed by Friedman and Russell in 1997 and improved by Zoran Zivkovic in 2004. It is currently widely used in speech recognition, image segmentation, financial data analysis, and bioinformatics.

The clustering method based on Gaussian finite mixture model can use the EM algorithm. Firstly, there is a logarithmic likelihood function:

$$L(\Theta \mid X) = \log P(X \mid \Theta) = \sum_{i=1}^{n} \left[\log \left(\sum_{l=1}^{k} \alpha_l N(x_i \mid \mu_l, \Sigma_l) \right) \right] \tag{3}$$

Where $\Theta = \{\alpha_1, \ldots, \alpha_k; \mu_1, \ldots, \mu_k; \Sigma_1, \ldots, \Sigma_k\}$, α_l represents Gaussian weight, and $\sum_{l=1}^{k} \alpha_l = 1$.

The purpose is to find:

$$\Theta_{MLE} = \underset{\Theta}{\operatorname{argmax}} \{\log P(X \mid \Theta)\} \tag{4}$$

Using the EM algorithm, there is an iterative relationship as follows:

$$\Theta^{(g+1)} = \underset{\Theta}{\operatorname{argmax}} \left(\int_z P\left(Z \mid X, \Theta^{(g)}\right) \log P(X, Z \mid \Theta) dz \right) \tag{5}$$

ISSN 2616-5775 Vol. 7, Issue 5: 123-129, DOI: 10.25236/AJCIS.2024.070516

Among them, $Z = \{z_1, \dots, z_n\}$ is a hidden variable.

In the end, we obtained Θ_{MLE} is as follows:

$$\alpha_l^{(g+1)} = \frac{\sum_{i=1}^{n} p\left(l|x_i, \Theta^{(g)}\right)}{n} \tag{6}$$

$$\mu_l^{(g+1)} = \frac{\sum_{i=1}^{n} x_i p\left(l|x_i, \Theta^{(g)}\right)}{\sum_{i=1}^{n} p\left(l|x_i, \Theta^{(g)}\right)} \tag{7}$$

$$\Sigma_l^{(g+1)} = \frac{\sum_{i=1}^{n} \left[x_i - \mu_l^{(i+1)}\right]\left[x_i - \mu_l^{(i+1)}\right]^T p\left(l|x_i, \Theta^{(g)}\right)}{\sum_{i=1}^{n} p\left(l|x_i, \Theta^{(g)}\right)} \tag{8}$$

Based on the above discussion, the GMM method considers the conditional probability of different clustering clusters for each data point, while relaxing the assumption of data distribution and being able to handle more complex structured data. The maximum likelihood method provides interpretable statistical significance. However, this method still has certain shortcomings, as its clustering effect on high-dimensional data is not good. If you want to use this method, you need to first reduce the dimensionality of high-dimensional data to obtain more accurate results. At the same time, due to its high computational complexity, the efficiency is low when dealing with large amounts of data.

2.3 BayesSpace clustering

BayesSpace is a Bayesian statistical method that can enhance spatial transcriptome data and make clustering results more accurate. Compared with traditional methods such as FISH, this method uses multivariate t-distribution to model the low dimensional gene expression matrix. Meanwhile, by applying the Potts model to merge spatial information, it ensures that adjacent points are divided into different clusters. This method draws on mature spatial statistical methods in the field of image analysis, effectively utilizing spatial information to improve resolution to the subplot level, enhance the performance of spatial data clustering, and improve the problem of low resolution in traditional methods. This method was proposed by the Raphael Gottardo team[4]. A method for spatial clustering based on Markov random field and MCMC method[5].

2.3.1 Data preprocessing

Firstly, perform logarithmic transformation on the original gene expression. The reason for performing logarithmic transformation is that there is a significant difference in the expression quantity of different genes, and there is a large amount of gene expressions are around 0. Using logarithmic transformation can reduce the gaps among different gene expressions.

The data used in the article was obtained through the 10xGenomics website (https://www.10xgenomics.com/), which is one of the most commonly used gene sequencing platforms.

2.3.2 Constructing a spatial clustering model

We use a spatial clustering method with a Markov random field, which has the following formula:

$$y_i|z_i = k, w_i \sim N(y_i; \mu_k, w_i^{-1}\Lambda^{-1}) \tag{9}$$

Among them, y_i represents the low dimensional representation of the gene expression vector, $z_i \in \{1, \dots, q\}$ represents different clustering clusters, μ_k represents the average vector of the k-th cluster, Λ is the accuracy matrix, and w_i is the unknown factor. Meanwhile, we assume that the common precision matrix is unconstrained.

For the number of clusters q, it is usually inferred based on biological knowledge, combined with clustering image features, to select the number near the elbow. For μ_k, Λ, and w_i, there is prior information as follows:

$$\mu_k \sim N(\mu_0, \Lambda_0^{-1}) \tag{10}$$

$$\Lambda \sim Wishart_d(\alpha, \text{diag}(\beta)_d^{-1}) \tag{11}$$

$$w_i \sim \Gamma\left(\frac{v}{2}, \frac{v}{2}\right) \tag{12}$$

Among them, μ_0 is the empirical mean vector, and Λ_0, α, and β are fixed parameters as weak priors.

ISSN 2616-5775 Vol. 7, Issue 5: 123-129, DOI: 10.25236/AJCIS.2024.070516

Additionally, it is assumed that y_i and w_i are independent of each other. Under this assumption, it can be observed that when w_i is marginalized, the normal likelihood follows a multidimensional t-distribution with an average value of 0 and a covariance matrix of $\frac{v}{v-2}\Lambda^{-1}$.

When estimating parameters, the MCMC method can be used. The final formula for the Markov random field is as follows:

$$\pi(z_i) = e^{\frac{\gamma}{|<ij>|}\times 2\sum_{<ij>} I(z_i=z_j)} \tag{13}$$

Among them, $|< ij >|$ represents all points j in the i neighborhood, and γ is a fixed parameter that controls the smoothness intensity.

2.3.3 Enhanced spatial clustering model

Due to the roughness of the unenhanced model, it is enhanced to improve resolution by dividing the original point into several sub points. In the iteration, the proposed values are generated as follows:

$$y_{ij}^{*\prime} = y_{ij}^{*} + \varepsilon_{ij} \tag{14}$$

$$\varepsilon_{ij} \sim N(0, \sigma^2 I_d)$$

Among them, σ^2 represents a fixed parameter and has $\sum_j \varepsilon_{ij} = 0$. Thus, the potential expression level y fluctuates within a relatively fixed range.

Finally, linear or nonlinear regression can be used to map it to the logarithmically normalized gene expression space of the original data.

Compared with the first two methods, BayesSpace incorporates spatial coordinate information into the Markov random field, making the model cover more data variable information and consider more about the spatial distribution of cells. Therefore, compared to traditional methods that only consider gene expression levels, the BayesSpace method can better utilize the spatial information, comprehensively use multiple data sources, and comprehensively grasp cell characteristics and functions, thereby improving the clustering results of tissue cells. However, there are still some possible issues with this method, such as it requires high preprocessing of transcriptome data, and this method also has high model complexity.

3. Results

The dataset used in this study is sequencing data of melanoma, which includes the horizontal and vertical coordinates of spots and gene expression levels, and is processed into a gene expression matrix. By using this data, we can explore the cell types in specific pathological tissues and identify gene types with high expression in tissue cells, which can be helpful for studying new gene therapy methods and finding therapeutic targets.

3.1 Data preprocessing

Due to significant differences in gene expression levels, using data from different dimensions for clustering and downstream analysis can easily make it difficult to capture cellular information with lower gene expression levels. Therefore, it is necessary to first perform data preprocessing and logarithmically standardize the data for ease of the following analyses.

For high-dimensional datasets, appropriate dimensionality reduction is usually required to simplify calculations. The most commonly used method is PCA. In this article, PCA is chosen to reduce the dimensionality of the data. Due to the significantly different expression levels of highly variable genes in different cells, different cell characteristics can be significantly distinguished. The first 7 principal components and the first 2000 highly variable genes are selected to significantly reduce the original data dimension, reduce computational and storage complexity, and remove redundant information to screen for features that can effectively distinguish cell differences, which is conducive to further analysis in the future.

ISSN 2616-5775 Vol. 7, Issue 5: 123-129, DOI: 10.25236/AJCIS.2024.070516

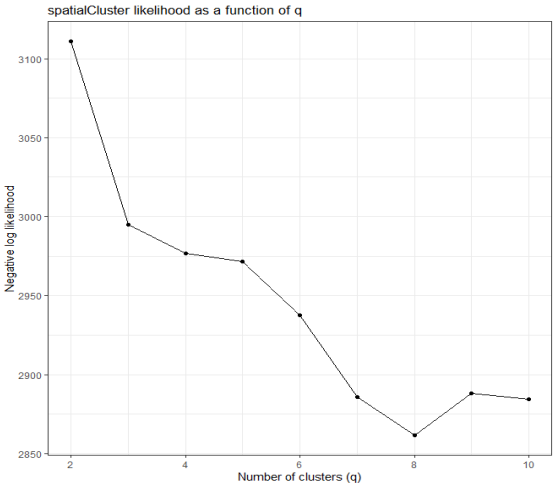

Figure 1: Negative log likelihood under different number of clusters

To determine the optimal number of clusters, the negative logarithmic likelihood function was calculated for different numbers of clusters. As shown in Figure 1, the horizontal axis represents the number of clusters, and the vertical axis represents the nearest log likelihood. Usually, based on the image, the elbow method is used to select the abscissa value of the point where the curve descent amplitude changes from steep to flat as the optimal number of clusters. In this study, the corresponding number of clusters is 3 to 6.

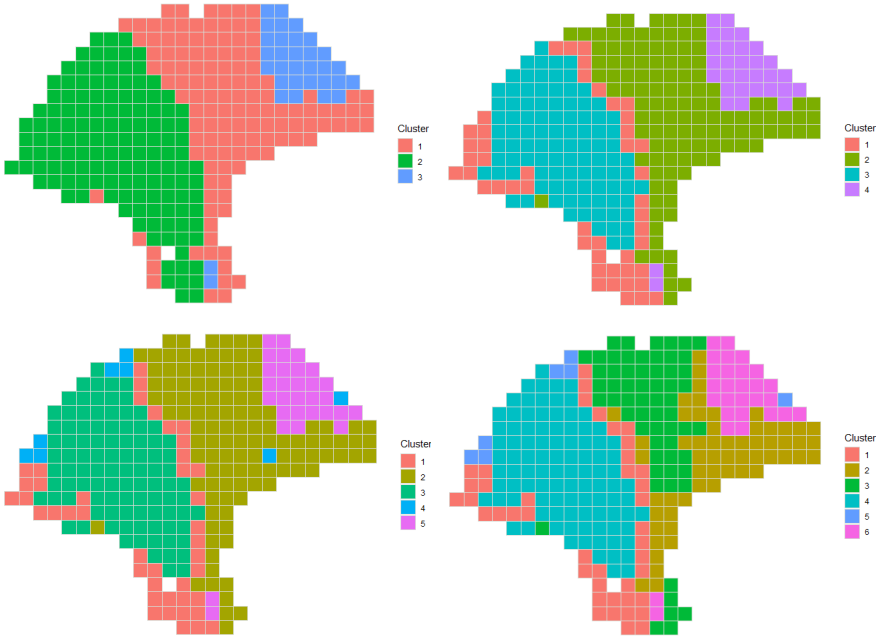

Figure 2: Cluster effect diagram under different number of clusters

To further determine the optimal number of clusters, each scenario is clustered separately, and by observing the clustering effect graph, the final optimal number of clusters is selected as 4 categories. When the number of clusters is 6, as shown in the bottom right corner of Figure 2, it can be seen that the distribution of cells in category 5 (blue) is relatively scattered and adjacent to category 6 (pink purple). Therefore, it is speculated that category 5 (blue) and category 6 (pink purple) are of the same cell type, resulting in an excessive number of clusters in category 6. When the number of clusters is 5, as shown in the bottom left of Figure 2, it is also found that the distribution of cells in category 4 (blue) is very scattered, so it is considered redundant to cluster into 5 categories. Comparing the three clusters, as shown in the upper left of Figure 2 and the four clusters, as shown in the upper right of Figure 2, it is found that the clustering effect is good. However, upon closer observation, it can be observed that Category 2 (green) in Cluster 3 is almost identical to Category 1 (pink) and Category 3 (blue-green) in Cluster 4, and Category 1 (pink) surrounds Category 3 (blue-green). Based on biological knowledge, it is speculated that there may

ISSN 2616-5775 Vol. 7, Issue 5: 123-129, DOI: 10.25236/AJCIS.2024.070516

be different cell types, Therefore, it is more reasonable to have 4 clusters.

3.2 Cluster analysis

Divide each point into multiple sub points, and use k-means method, GMM method, and BayesSpace method for clustering after enhancement. The results are shown in Figure 3, which are k-means method, GMM method, and BayesSpace method from left to right. After comparison, it was found that in the clustering results based on the k-means method, the distribution of the second category (green) is very unclear and scattered. From the analysis of location information, the possibility of belonging to the same cell type is relatively small, so using the k-means method for clustering is not accurate; In the clustering results based on GMM method and spatial clustering method, the number of isolated samples in the second category (green) of the former is more and more scattered, and there are more small areas that are not connected to the concentrated areas in the second category (green). Compared with BayesSpace method, the clustering results of GMM method are mixed with the first category (pink) and third category (blue-green) areas, resulting in poor classification performance. Therefore, using the BayesSpace method for clustering results in higher accuracy and better performance.

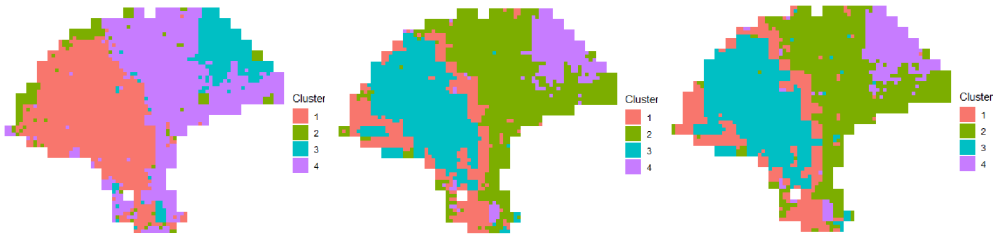

Figure 3: Cluster effect diagram with 4 clusters under different enhanced clustering methods

3.3 Marker gene staining

Based on biological knowledge, identify marker genes in the cluster 1-4 and label them[6]. Among them, PMEL is highly expressed in melanoma, CD2 is highly expressed in T cells, CD19 is highly expressed in B cells, and COL1A1 is highly expressed in fibroblasts. Visualize the four marker genes as shown in figure 4, where darker colors indicate higher expression levels of these genes, while lighter colors indicate lower expression levels of these genes.

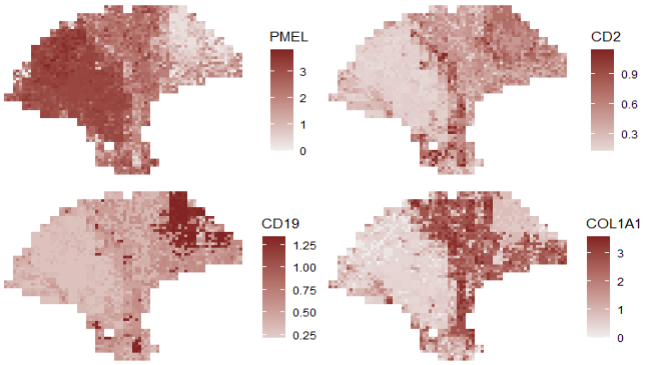

Figure 4: Expression levels of different marker genes under the BayesSpace method

From Figure 4, it can be seen that through gene expression related data, the expression information of different genes in different cells can be clearly visualized. Based on this feature, different cells can be spatially clustered to identify highly expressed genes in different categories of cells, thus accurately distinguishing cell types at different positions.

Compare the gene expression staining map enhanced by sub points (the first row in the figure) with the gene expression staining map without enhancement (the second row in the figure), as shown in Figure 5.

ISSN 2616-5775 Vol. 7, Issue 5: 123-129, DOI: 10.25236/AJCIS.2024.070516

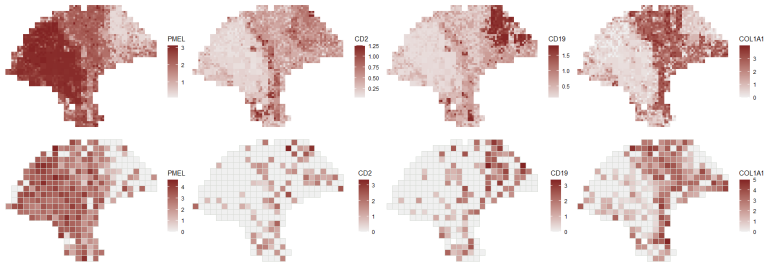

Figure 5: The expression levels of marker genes before and after enhancement

From Figure 5, it can be seen that the color areas of the unenhanced stained image are relatively chaotic, with no obvious dark areas. About two-thirds of the regions in the PMEL gene staining image are dyed dark; In the CD2 and CD19 gene staining images, the distribution of darker color blocks is disorderly and scattered, without significant features; More than half of the regions in the COL1A1 gene staining image are dyed dark. Therefore, the data augmentation method of dividing one site into multiple sub points can accurately identify highly expressed genes, thereby effectively improving clustering performance.

By comparing k-means, GMM, and BayesSpace methods, it can be found that the BayesSpace method performs better when considering cellular spatial coordinate information. In contrast, traditional methods such as k-means and GMM, although performing well in some cases, cannot fully consider the spatial distribution characteristics of cells, resulting in a lack of practicality and accuracy in clustering results. The BayesSpace method, combined with spatial location information and gene expression data, can better determine cell types, especially in identifying highly variable genes. This accurate classification result provides convenience for further research, helps to explore the characteristics of cell types more deeply, and promotes the development of related fields.

4. Conclusions

This article is based on the spatial transcriptome data of melanoma using BayesSpace method and retrograde clustering analysis. Firstly, perform data preprocessing to make the data a modelable gene expression matrix. Then, the PCA method is used to reduce the dimensionality of the data, and the optimal number of clusters is selected through the logarithmic likelihood function. Next, three clustering methods are used: k-means method, GMM method, and BayesSpace method, respectively, to cluster cells. Based on the enhanced classification images, the best clustering method is BayesSpace, and the marker genes will be stained to obtain the clustering results. This method has higher clustering accuracy, thereby more accurately identifying cell types.

This study achieved high clustering accuracy by using multiple clustering methods to partition spatial transcriptome cell clusters. By comparing with other methods, this article proves the superiority of the BayesSpace method at the method level. In the field of biomedicine, the analysis results of this article annotate the cell types of different classification clusters in downstream analysis. Further research can be conducted on this basis to explore the characteristic cells and marker genes that have diagnostic significance for melanoma, and to promote the search for therapeutic targets for melanoma.

References

[1] Williams, C.G., Lee, H.J., Asatsuma, T.et al. An introduction to spatial transcriptomics for biomedical research [J]. Genome Medicine, 2022, 14(1): 68.
[2] Asp, M., Bergenstråhle, J., Lundeberg. Spatially resolved transcriptomes—next generation tools for tissue exploration [J]. BioEssays, 2020, 42(10), 1900221.
[3] Nikhil Rao, Sheila Clark, and Olivia Habern. Bridging Genomics and Tissue Pathology[J]. Genetic Engineering & Biotechnology News, 2020, 40(2): 50-51.
[4] Zhao, E., Stone, M.R., Ren, X. et al. Spatial transcriptomics at subspot resolution with BayesSpace[J]. Nature Biotechnology, 2021, 39(11): 1375-1384.
[5] Yi Yang, Xingjie Shi, Wei Liu et al. SC-MEB: spatial clustering with hidden Markov random field using empirical Bayes [J]. Briefings in Bioinformatics, 2022, 23(1).
[6] Pérez-Cobas A E, Gomez-Valero L, Buchrieser C. Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses[J]. Microbial genomics, 2020, 6(8).