Single cell RNA cluster analysis

Jinge Yu

1 Introductioin

Recently single-cell sequencing technologies is playing an important role in cellular molecular, enables people to explore heterogeneity at celluar level. There are many methods such as single-cell measurements of genome sequence, chromatin accessibility, DNA methylation, cell surface proteins, histone modifications and chromosomal conformation. Among these technologies Single-Cell RNA sequencing is one of the most widely used. scRNA technologies has emerged as a powerful tool for the unbiased and systematic characterization of the cells present in a given tissue, which helps us better understand cell heterogeneity.

Alough scRNA is good at identifing cell clusters in a tissues, their spatial organization is still missing. As to explore their spatial structure, potts model together with DMH(Double Metroplis Hasting algorithm) were applied to lung cancer pathological image analysis in Qianyun Li(2017). Besides, Reuben Moncada(2020) combines a microarray-based spatial transcriptomics method to capture the spatial patterns of gene expression. By using an array of spots, this method is easily scalable to any architecturally complex tissue. Notice that the word **spot** can be viewed as a tiny region of several cells close to each other. The gene expression level in one spot is the sum of gene expression of all cells in the spot.

The project focuses on estimating the spatial organization in the cellular level as well as constructing a link between scRNA-seq and spots data. A tumor is first divided for single-cell RNA sequencing analysis–identify the cell types in the tissue. For the remaining tissue, ST(spots) data are processed to explore region messages across tissues. We aims at finding a cluster method detecting the heterogeneity between single cells, and cluster different cell spots at the same time.

During experinments, a tumor is first divided and a single-cell suspension is generated from one ption and processed for scRNA-seq to identify the cell populations present in the tissue. From the remaining tissue, cryosection are processed transcripts across the tissue. Our approach infers specific cell types in a given tissue region by integrating single-cell RNA-seq and Bayesian based spatial models. Notice that gene expression in one spot is the sum of expression level of those cells the spot.

In section 2, we briefly introduced some preprocessing procedures, then we developed two Bayesian clustering method for scRNA data and spot data respectively. In setion 3, we provided detials for single cell RNA model, and in section 4, potts model with Double Metropolis Hasting samping method were described. Results of our model are in section 4. In section 5, we show cluster results of our models. Besides, we listede our future plan in section 6.

2 Data Preprocessing

2.1 Data resources

Our data comes from NCBI with accession number GSE111672, and we choose PDAC-B inDrop1 as scRNA data, PDAC-B ST1 as ST data. Both data set comes from single-cell suspension of tissue of pancreatic adenocarcinoma from homo sapiens. The scRNA data consists of 19831 genes and 2000 cells, the spot data consists of 16528 genes and 996 spots.

2.2 Data visiualization

Since the number of genes are too large, we prefer choosing the 500 most variable genes first to make heatmap and operate analysis later. Here are heatmaps of both data sets:

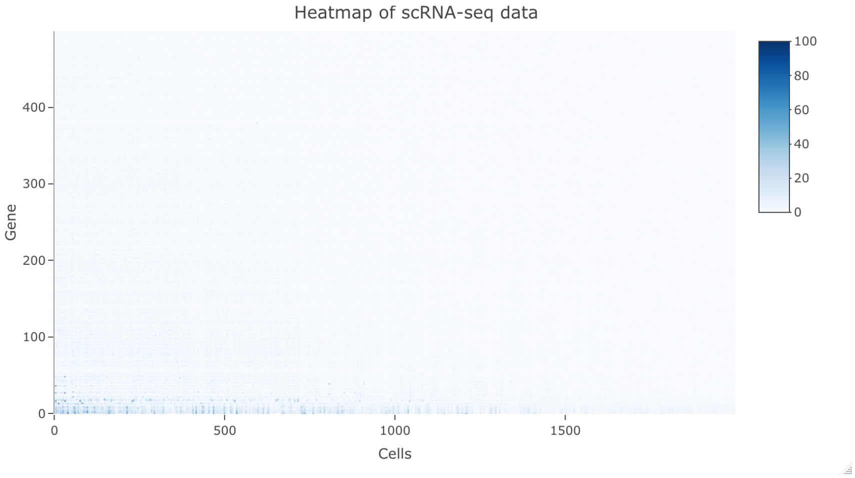

Figure 1: Heatmap of scRNA data

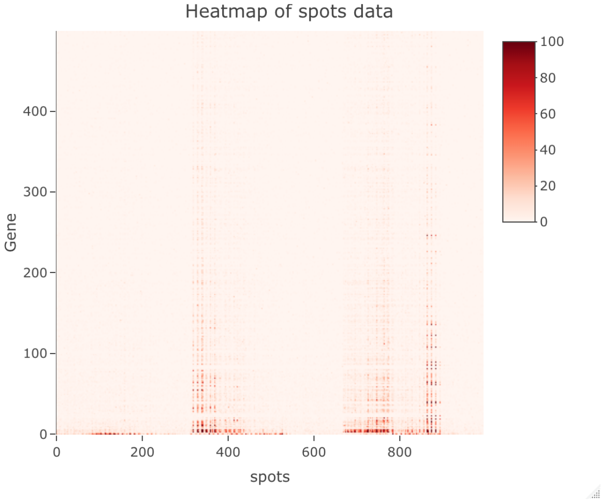

Figure 2: Heatmap of ST data

From the heatmap of both data sets we know that there are heterogeneity between different cells and spots, and the data are suitable for clustering.

2.3 Data processing

In the preprocessing procedure, the scRNA-seq data were normalized to correct for technical factors: the library size for each cell, the sum of read counts across all genes, and the median of all library sizes were calculated and the original counts were divided by its corresponding library size and multiplied the ratio by the median library size. We took the floor of normalized data. Refer to https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3405534 and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3405531 for details.

3 Mixture Gaussian model

Suppose the scRNA-seq data are collected for N cells with G genes. We denote by Z_{gi} the gene expression level for gene g with cell i. Notice that Z_{gi} is the normalized and continuously processed data. Assume there are K clusters(/cell types) of cells.

We model the data with mixture gaussian model:

$$z_{gi}|C_i = k \quad \sim \quad N(h_{gk}, \sigma_g^2), \quad (k = 1, \cdots K, i = 1, \cdots M)$$

where C_i represents cell type(cluster) of the i-th cell. And C_i, h_{gk}, σ_g^2, $\ i = 1, \cdots M$, $g = 1, \cdots, G, k = 1, \ldots, K$ are unknown parameters, which needs update. We consider Gibbs samping to update those parameters one by one. Besides, we utilize conjugate priors for all model parameters:

$$h_{gk} \sim N(\eta_h, \tau_h^2), \quad g = 1, \ldots, G, k = 1, \ldots, K$$

$$P(C_i = k) = \pi_k, \quad \pi_{1:K} \sim Dirichlet(\gamma_1, \ldots, \gamma_1)$$

$$\sigma_g^2 \sim Inv - gamm(\alpha_1, \beta_1) \quad g = 1, \cdots G$$

Given the priors, we can carry out posterior samping. In each iteratioin, the sampling scheme proceeds is as follows.("-" means given all other variables)

1. Update h_{gk} from

$$h_{gk}|- \sim N(\hat{h_{gk}}, (N_k/\sigma_g^2 + 1/\eta_h^2)^{-1}), \quad g = 1, \ldots, G, k = 1, \ldots, K.$$

   where $\hat{h_{gk}} = \frac{N_k \overline{Z}_{gk}/\sigma_g^2 + \eta_h/\tau_h^2}{N_k/\sigma_g^2 + 1/\eta_h^2}$, $N_k = \sum_{i=1}^{M} \mathbb{I}(C_i = k)$,$\overline{Z}_{gk} = \frac{1}{N_k} \sum_{i=1}^{M} Z_{gi} \mathbb{I}(C_i = k)$.

2. Update squared deviation associated with the g th gene from

$$\frac{1}{\sigma_g^2}|- \sim gamma(\frac{M}{2} + \alpha_1, \beta_1 + \frac{1}{2} \sum_{i=1}^{M} (Z_{gi} - h_{c_i g})^2).$$

3. Update the cell type parameter

$$P(C_i = k|-) = \frac{\prod_{g=1}^{G} N(Z_{gi} : h_{gk}, \sigma_g^2)}{\sum_{k=1}^{K} \prod_{g=1}^{G} N(Z_{gi} : h_{gk}, \sigma_g^2)}.$$

4 Potts Model

4.1 Spot data visualization

What is spots look like? Tere is a sketch map of spots:

Frome the figure above we can see it that each square represents a spot, and each cell only belongs to one spot. Then have a look at our spot:

The white dots are missing values in the data, maybe the corresponding spots were hard to observe or no cells was located at the spots. However, for the sake of easy calculation, we fill in all the null spots in two ways. As for the left bottom ones, we suppose the gene expression level is all 0 with all genes. The rest ones are filled with the average value of their neighbours.

Suppose the spot data are collected for N spots with G genes, and L is the row number of spots, W is the column number of spots. We denote by X_{lwg} the gene expression level for gene g with spot (l, w) where (l, w) means coordinates of a spot. Notice that X_{gi} is the normalized and continuously processed data. Assume there are S clusters(/regions) of spots.

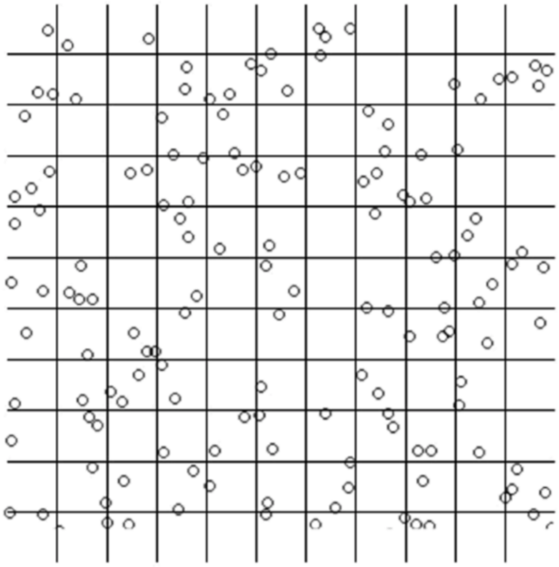

Figure 3: Spots

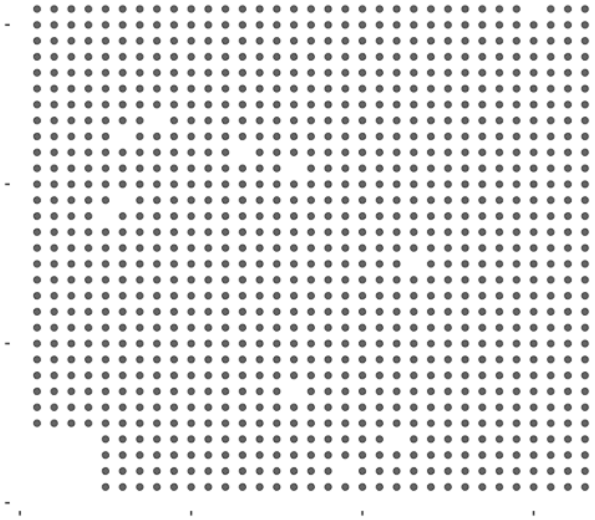

Figure 4: Spots

4.2 Model

We model the data with potts model:

$$x_{lwg}|R_{lw} = s \sim N(\mu_{sg}, \sigma_g'^2).$$

where R_{lw} is the region which spot (l, w) belongs to. $R = R_{lw}$, $1 \leq l \leq L$, $1 \leq w \leq W$. $R_{lw} \in \{1, 2, ..., S\}$.

$$Pr(R|\Theta) = \frac{1}{C(\Theta)} exp\{-H(R|\Theta)\}$$

Where $H(R|\Theta)$ is the energy function.

$$H(R|\Theta) = -\sum_{(l,w)\sim(l',w')} \theta_{R_{lw}R_{l'w'}} \mathbb{I}(R_{lw} \neq R_{l'w'}).$$

where $\Theta = (\theta ss')$ represents interaction energy matrix, whose diagnoal element are zeros.

$$H(R_{lw}|R_{-l,-w}, \Theta) = -\sum_{(l',w')\in Nei(l,w)} \theta_{R_{lw}R_{l'w'}} \mathbb{I}(R_{lw} \neq R_{l'w'})$$

Parameters need to be updated:

μ_{sg}, $\quad \sigma_g^2(1 \leq g \leq G)$, $\quad R_{lw}(1 \leq s \leq S, \quad 1 \leq w \leq W)$, $\quad \Theta = (\theta ss')$, $\quad (1 \leq s \leq S, \quad 1 \leq s' \leq S, \quad s \neq s')$.

However, there is an intractable normalizing constant $C(\Theta)$ in the posterior distributioin of $\theta_{ss'}$. So we consider double Metropolis–Hastings algorithm to sample $\theta_{ss'}$.

Besides, we utilize conjugate priors for all model parameters:

$$\theta ss' \sim N(\eta_\theta, \tau_\theta^2), \quad 1 \leq s \leq S, \quad 1 \leq s' \leq S, \quad s \neq s'.$$

$$\mu_{sg} \sim N(\eta_\mu, \tau_\mu^2), \quad 1 \leq s \leq S, 1 \leq g \leq G.$$

$$\sigma_g'^2 \sim Inv - gamma(\alpha_2, \beta_2), \quad 1 \leq g \leq G.$$

Given the priors, we can carry out posterior samping. In each iteratioin, the sampling scheme proceeds is as follows.("-" means given all other variables)

1. Update μ_{sg} from

$$\mu_{sg}|- \sim N(\hat{\mu_{gk}}, (N_s/\sigma_g'^2 + 1/\eta_\mu^2)^{-1}), \quad g = 1, \ldots, G, s = 1, \ldots, S.$$

   where $\hat{\mu_{gk}} = \frac{N_s\overline{X}_{sg}/\sigma_g'^2 + \eta_\mu/\tau_\mu^2}{N_s/\sigma_g'^2 + 1/\eta_\mu^2}$, $N_s = \sum_{l=1}^{L}\sum_{w=1}^{W} \mathbb{I}(R_{lw} = s)$,$\overline{X}_{sk} = \frac{1}{N_s}\sum_{l=1}^{L}\sum_{w=1}^{W} x_{lwg}\mathbb{I}(R_{lw} = s)$.

2. Update squared deviation associated with the g th gene from

$$\frac{1}{\sigma_g'^2}|- \sim gamma(\frac{LW}{2} + \alpha_2, \beta_2 + \frac{1}{2}\sum_{l=1}^{L}\sum_{w=1}^{W}(X_{lwg} - \mu_{R_{lw}g})^2).$$

3. Update R_{lw} from

$$\mathbb{P}(R_l w = s) = \frac{\prod_{g=1}^{G} N(x_{lwg}; \mu_{sg}, \sigma_g'^2) exp\{\sum_{l',w')\in Nei(l,w)} \theta_{sR_{l'w'}} \mathbb{I}(R_{l'w'} \neq s\}}{\sum_{s=1}^{S}\prod_{g=1}^{G} N(x_{lwg}; \mu_{sg}, \sigma_g'^2) exp\{\sum_{l',w')\in Nei(l,w)} \theta_{sR_{l'w'}} \mathbb{I}(R_{l'w'} \neq s\}}.$$

4. Update $\theta_{ss'}$, using Douoble MH algorithm.

   (a) Sample $\theta_{ss'}^\star$ from prposal distribution $N(\theta_{ss'}, \tau_0^2)$.

   (b) Simulate auxiliary variable $y \sim \mathbb{P}_{\Theta^\star}(y|R)$.

(c) Calculate the MH ratio:

$$r = \frac{N(\theta^\star_{ss'}; \eta_\theta, \tau^2_\theta) N(\theta_{ss'}; \theta^\star_{ss'}, \tau^2_0) \mathbb{P}(R|\Theta^\star) \mathbb{P}(y|\Theta)}{N(\theta_{ss'}; \eta_\theta, \tau^2_\theta) N(\theta^\star_{ss'}; \theta_{ss'}, \tau^2_0) \mathbb{P}(R|\Theta) \mathbb{P}(y|\Theta^\star)}$$

(d) Accept $\theta^\star_{ss'}$ with probability $\min\{1, r\}$.

5 Results

The cluster number parameter K from 4 to 8, S from 3 to 7. For each K and S, we ran 5000 MCMC iterations respectively, of which the first 4000 were discarded as burn-in. And we use BIC to define the best cluster number of both data sets.

5.1 scRNA data

BIC cluster results:

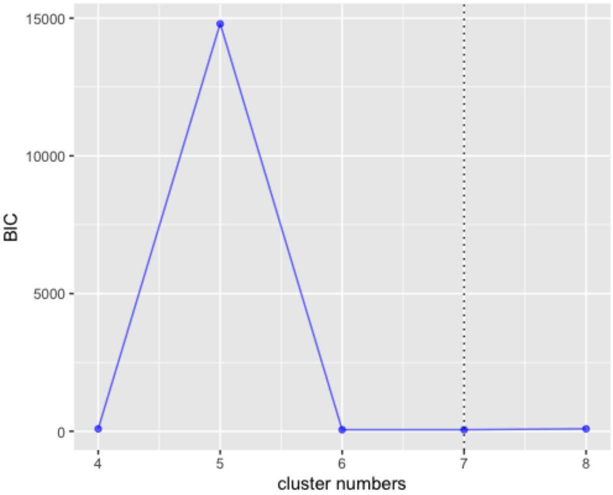

Figure 5: BIC cluster of scRNA data

Clusers	4	5	6	7	8
BIC value	87.88	14794.44	56.19	55.55	88.09

From Figure 5 and Table 1 we know that 7 is the best cluster number for scRNA data. Then we have a look at the iteration plots of two parameters $h_{20,3}$ and σ^2_{100}:

From Figure 6 and Figure 7 we can see it that parameters are stable during iteratioin, our model makes sense, and the cluster result is more convicing. I used umap method to perform dimension reduction on genes, and the result is as follows:

The picture shows one cluster is far away from the others, which may be cancer cells. And we can search for those cells in the future to explore more characteristics of cancer cells.

5.2 Spot data

BIC cluster results:

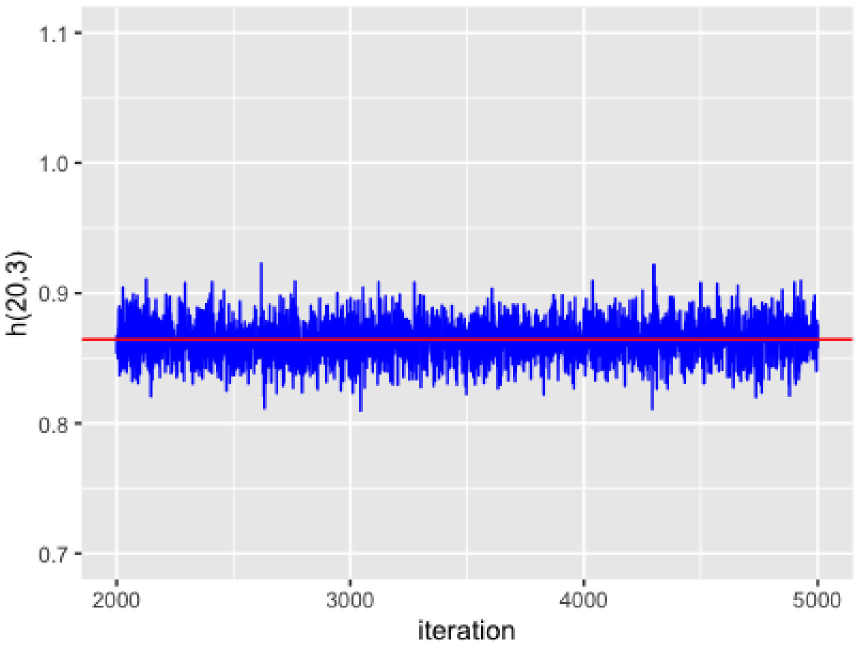

Figure 6: Iteration parameter of scRNA data

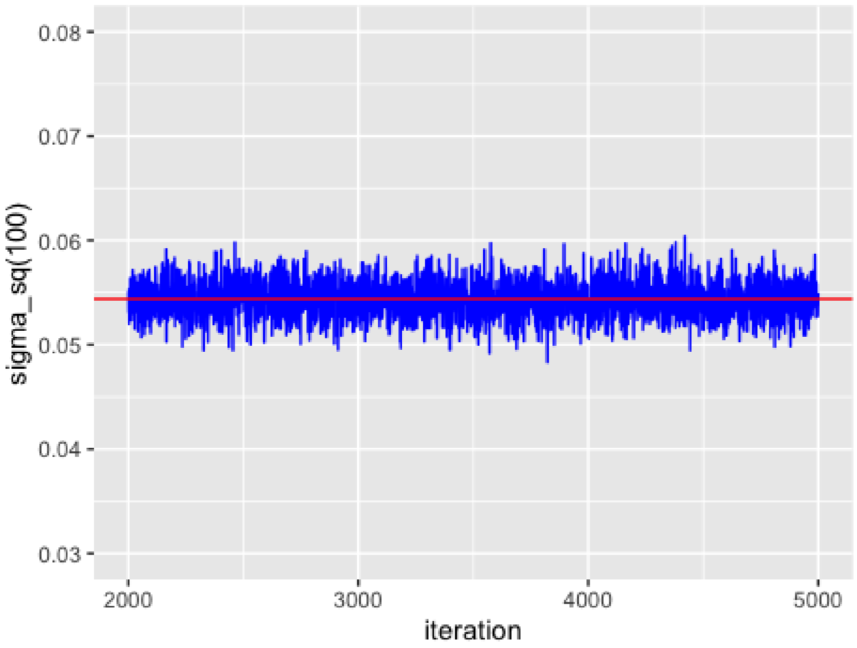

Figure 7: Iteration parameter of scRNA data

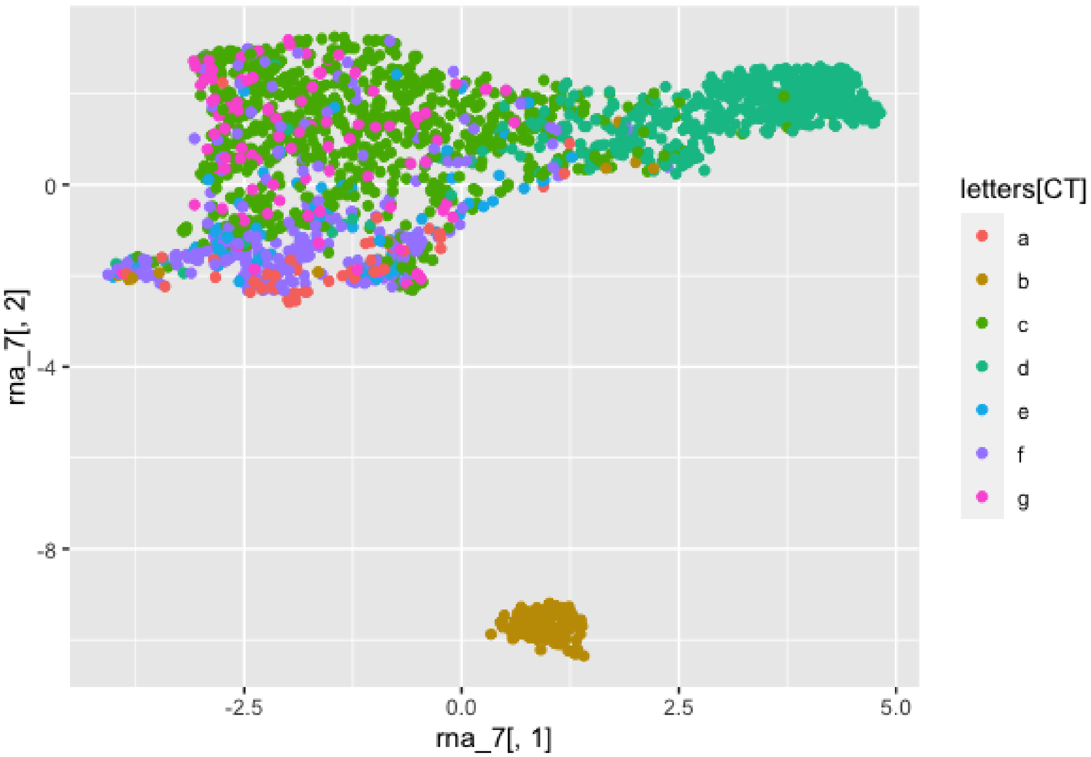

Figure 8: Iteration parameter of scRNA data

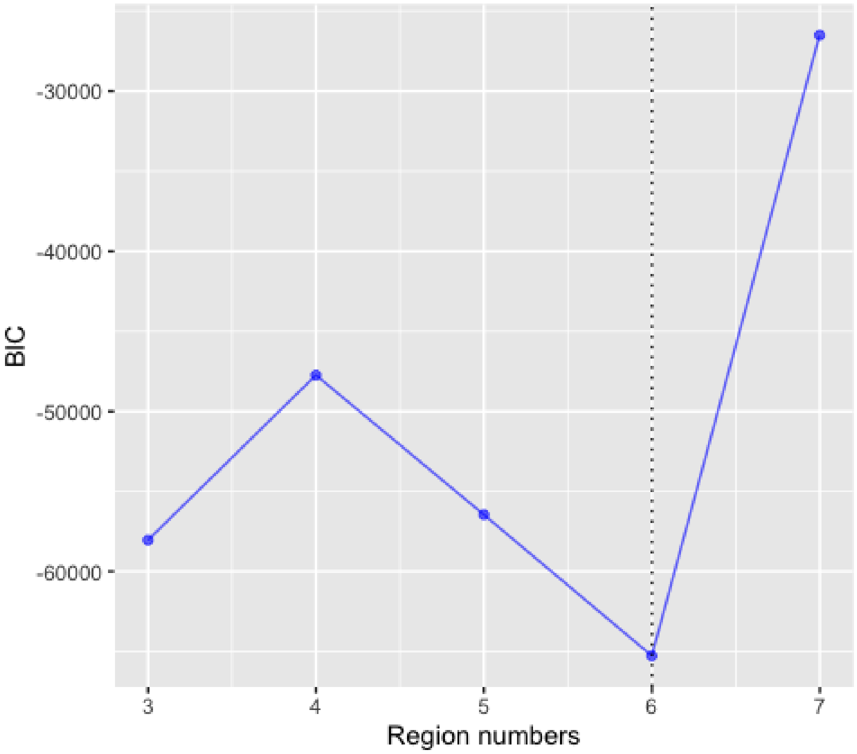

Figure 9: BIC cluster of spot data

From figure above we know that 6 is the best cluster number. Then we plot three parameters' iteration line when there are 6 regions.

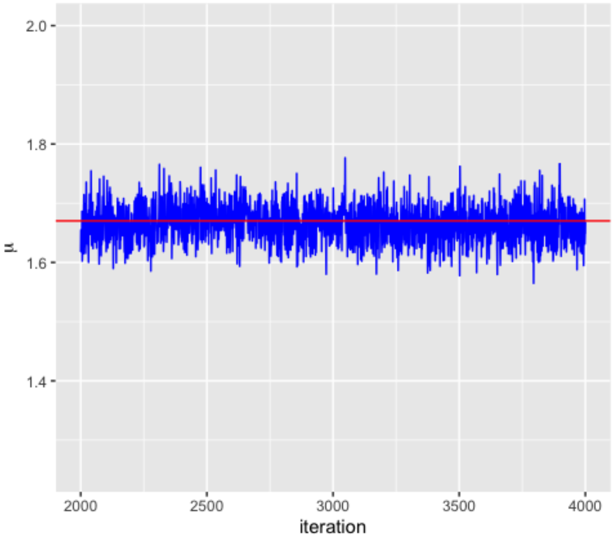

Figure 10: Iteration parameter of spot data

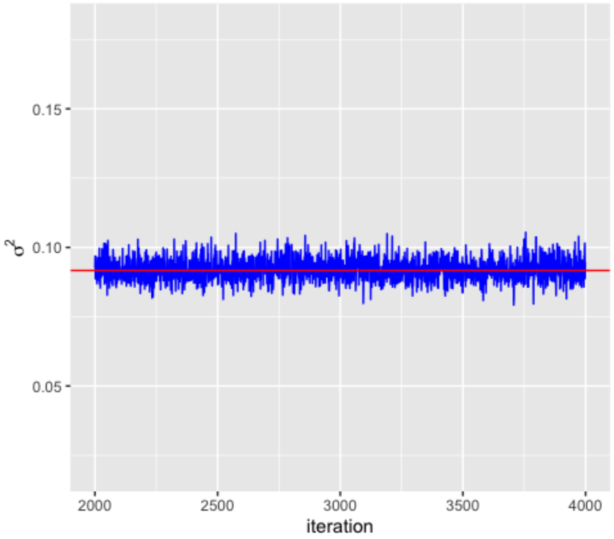

Figure 11: Iteration parameter of spot data

Figure 10 and figure 11 shows the parameter $\mu_{20,3}$ and σ'^{1}_{100} is stable during iteration, however $\theta_{3,3}$ has a large varience. But since we are interested in R rather than θ, diverge of θ is less important.

Cluster result of spot data when clustering number is 6.

Figure seems not to detect the heterogeneity between spots, however we can see a vertical blue and red regions, at which cancer cells may be located. So that the spots data provides us spatial information about different cells. For example, the molecule-targeted treatment of tumors has been widely accepted nowadays, our spatial structure will help targeted treatment.

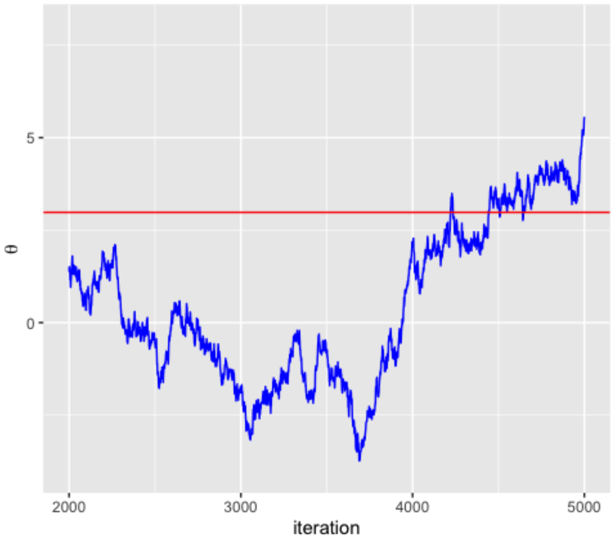

Figure 12: Iteration parameter of spot data

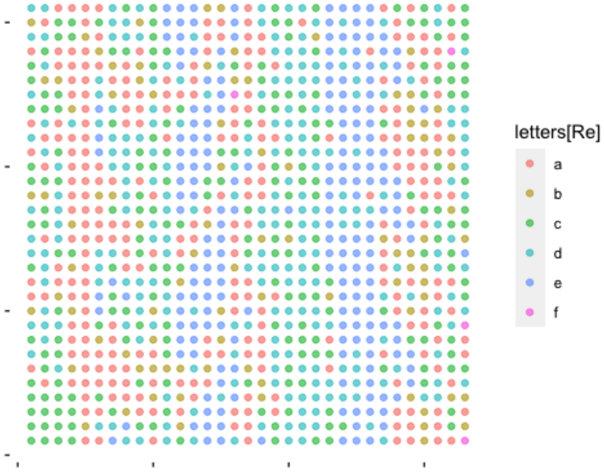

Figure 13: Cluster result of spot data

6 Future Plan

The project model single-cell RNA data and spot data separately, which lost the information of the same gene in both data sets. Therefore, we'd like to find a link between scRNA-seq data and spot data, integrate single-cell RNA-seq methods and micoarray-based spatial models.

Since the iteration takes a long time, we should consider parallel methods in the future.

We are not clear about the biological meaning of spatial clustering results, so we need to figure out the deep meaning of them.

Reference

Liang F. A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants[J]. Journal of Statistical Computation and Simulation, 2010, 80(9): 1007-1022.

Li Q, Yi F, Wang T, et al. Lung cancer pathological image analysis using a hidden potts model[J]. Cancer informatics, 2017, 16: 1176935117711910.

Moncada R, Barkley D, Wagner F, et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas[J]. Nature Biotechnology, 2020: 1-10.