Proposal

Jinge Yu

4/7/2020

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. There is a picture as explanation:

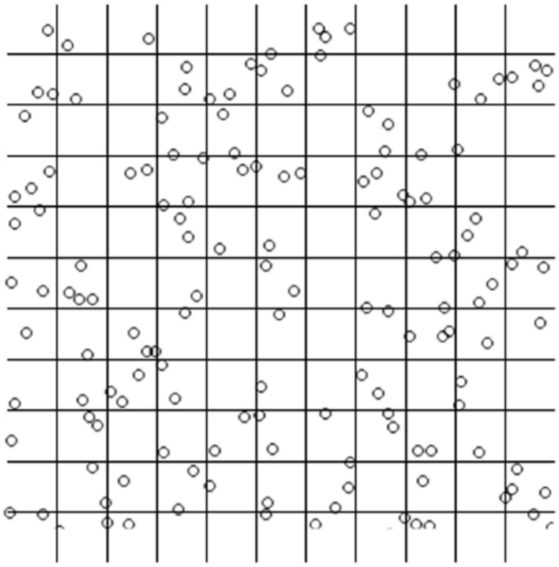


Figure 1: Spots

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.

EDA

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. 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.

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.

Data information

```r
rm(list = ls(all = TRUE))
library(ggplot2)
library(dplyr)
library(factoextra)
library(plotly)
library(cluster)

scRNA <- read.table('GSM3405531_PDAC-B-indrop1.tsv', header=TRUE, sep="\t")
ST <- read.table('GSM3405534_PDAC-B-ST1.tsv', header = TRUE, sep = '\t', quote = "")
rownames(scRNA) <- scRNA[,1]
scRNA <- scRNA[,-1]
rownames(ST) <- ST[,1]
ST <- ST[,-1]
dim(ST)
dim(scRNA)
```

```
# 996 16528
# 19831  2000
```

Qucik view of the two datasets:

```r
head(ST[,1:5])
#      A1BG A1CF A2M A2ML1 A4GALT
# 7x2     0    0   0     0      0
# 8x2     0    0   0     0      0
# 9x2     0    0   0     0      0
# 10x2    0    0   0     0      0
# 11x2    0    0   0     0      0
# 6x3     0    0   1     0      0
```

The rows of ST data represents different spots, columns stands for different genes, values are sum of gene expression in the corresponding spot. Column names means the coordinates of spots. Take **0** in the first row and the first column for example, we know that gene A1BG did not expressed in spot $(7, 2)$.

```r
head(scRNA[,1:4])
#  cell_209.199 cell_009.261 cell_277.373 cell_350.199 cell_049.210
# A1BG               0            0            0            0            0
# A1CF               0            0            0            0            0
# A2M                0            0            0            0            0
# A2ML1              0            0            0            0            0
# A3GALT2            0            0            0            0            0
# A4GALT             0            0            0            0            0
```

The row labels of scRNA data are genes, and the columns are different cells.

Since there are amount of zeros in our data, we what to see if there are unexpressed gene in all cells :

```
t_f <- apply(scRNA != 0, 2, sum)
index <- which(t_f == 0)
length(index)
# 0
```

So that all genes are expressed in our data. Next we explore zero-inflated effects:

```
effect = colSums(scRNA == 0)/dim(scRNA)[1]
head(effect)
# ell_209.199  cell_009.261  cell_277.373 cell_350.199 cell_049.210  cell_376.199
#  0.8789774    0.9392870      0.9068630    0.9566840    0.9540618    0.9214866
```

```
effect = rowSums(ST == 0)/dim(ST)[2]
head(effect)
#       7x2       8x2       9x2        10x2      11x2       6x3
# 0.9894119    0.9713819 0.9835431  0.9896539 0.9889279   0.9810019
```

The results above show that only a few genes are expressed in many cells and spots. Are there all 0 or 1?

```
sum(ST>10)     #9579
sum(ST > 100)  #232
sum(ST > 200)  #60
sum(ST > 500)  #7
sum(ST > 1000) #0
```

It seems that some some spots enjoy a high gene expression level, and this maybe a feature to distinguish theses spots from other ones, cancer cells for example.

Data Processing

Data normalization

Both datasets need to be normalized to fit in our model to remove scale effects.

```
tmp = median(colSums(scRNA))/colSums(scRNA)
RNA = floor(sweep(scRNA,2,tmp,'*'))
tmp = median(rowSums(ST))/rowSums(ST)
ST_n = floor(sweep(ST,1,tmp,'*'))
```

In some context, we'd like to transform count data to continuous ones:

```
RNAC <- log(2*(RNA +1))
STC <- log(2*(ST_n +1))
```

Data visiualization

Heatmap

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

```
sd_rna = apply(RNA, 1, sd)
sd_st = apply(ST_n, 2, sd)
```

```r
rna_varible = scRNA[order(sd_rna, decreasing = T),] %>% head(500)
st_varible = t(ST[,order(sd_st, decreasing = T)]) %>% head(500)
max(rna_varible) #723
max(st_varible) #775

vals <- unique(scales::rescale(c(volcano)))
o <- order(vals, decreasing = FALSE)
cols <- scales::col_numeric("Blues", domain = NULL)(vals)
colz <- setNames(data.frame(vals[o], cols[o]), NULL)
plot_ly(z = as.matrix(rna_varible)[,1:100], zmax= 500 ,zmin=0, colorscale = colz, type = "heatmap")%>%
  layout(title = "Heatmap of scRNA-seq data",
         xaxis = list(title = "Cells"),
         yaxis = list(title = "Gene"))
```

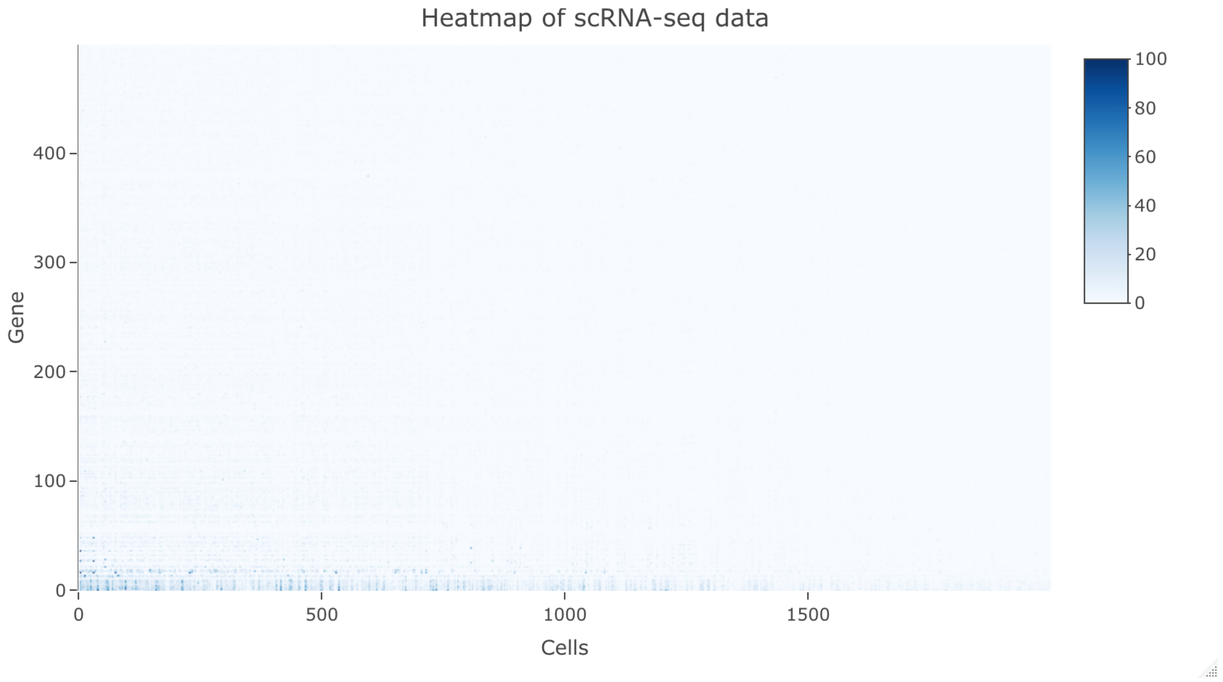


Figure 2: Heatmap of scRNA data

```r
cols2 <- scales::col_numeric("Reds", domain = NULL)(vals)
colz2 <- setNames(data.frame(vals[o], cols2[o]), NULL)
plot_ly(z = as.matrix(st_varible), zmax= 100 ,zmin=0, colorscale = colz2, type = "heatmap")%>%
  layout(title = "Heatmap of spots data",
         xaxis = list(title = "spots"),
         yaxis = list(title = "Gene"))
```

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

Kmeans Cluster

I use the most variable 500 genes as features to perform Kmeans cluster with cells and spots.

```r
set.seed(1996)
df_1 <- t(RNA[order(sd_rna, decreasing = T),] %>% head(500))
```

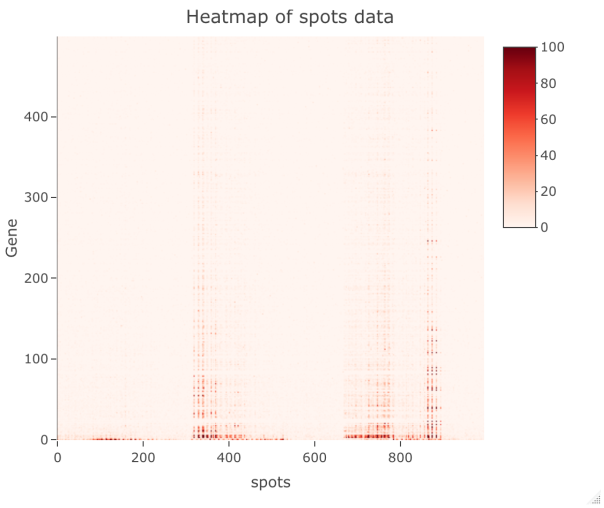
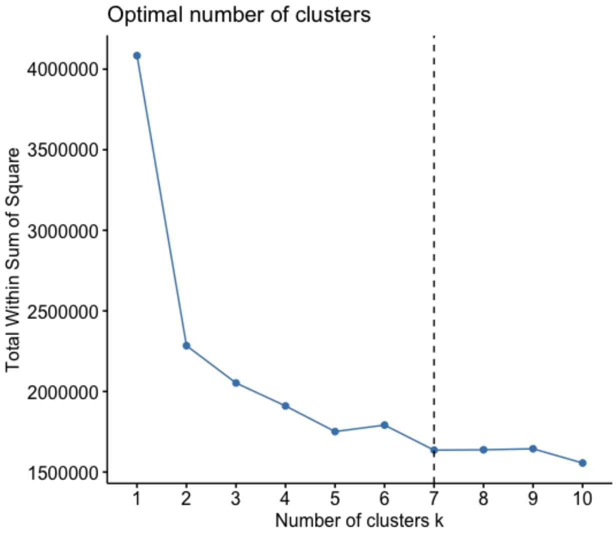
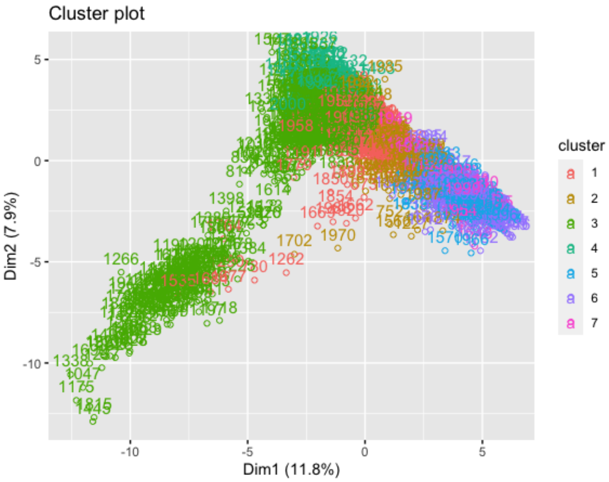


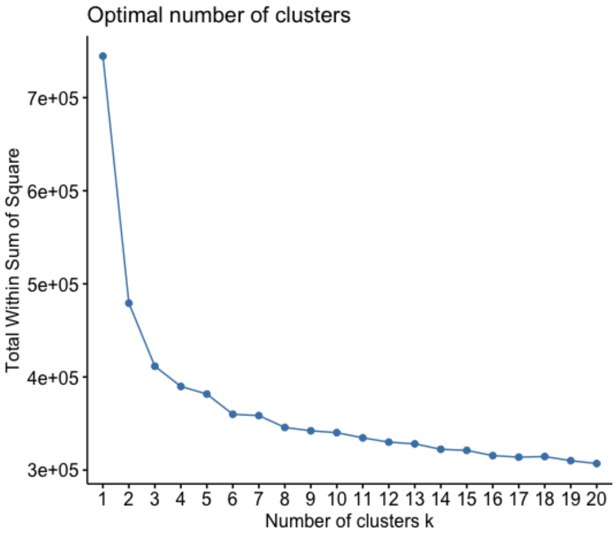
Figure 3: Heatmap of ST data

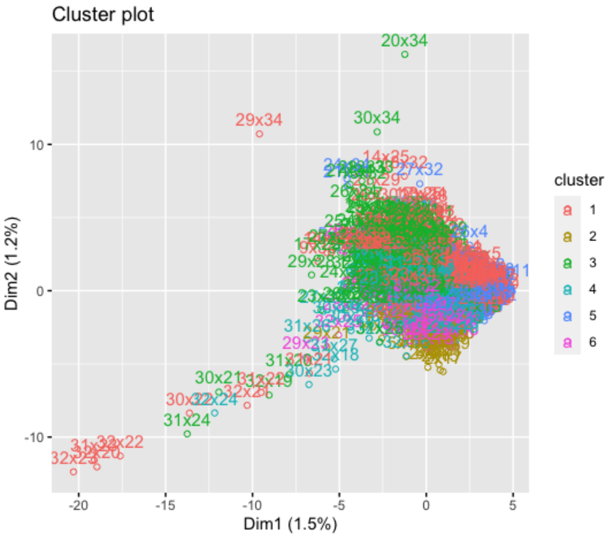
```
fviz_nbclust(df_1, kmeans, method = "wss",k.max=10) + geom_vline(xintercept = 7, linetype = 2)
km_result <- kmeans(df_1, 7)
fviz_cluster(km_result, data = df_1,
             ellipse = F,
             show.clust.cent = F,
             shape = 1,
             repel = F,
             ggtheme = theme_grey()
)
```





```r
df_2 <- ST_n[,order(sd_st, decreasing = T)][,1:500]
fviz_nbclust(df_2, kmeans, method = "wss",k.max=20)
km_result <- kmeans(df_2, 6)
fviz_cluster(km_result, data = df_2,
             ellipse = F,
             show.clust.cent = F,
             shape = 1,
             repel = F,
             ggtheme = theme_grey()
)
```





Figures above show that both scTNA-seq and spots data fails to cluster cells properly, especially for spots data. Then I do data clustering assessment.

For scRNA-seq data:

```r
res1 = get_clust_tendency(df_1, 50, graph = TRUE, gradient = list(low = "steelblue", high = "white"))
res1$hopkins_stat
# 0.9073394
res1$plot
```

The value of Hopkins statistic is 0.9073394, which is larger than 0.5, so that scRNA data is weaky clusterable. From the plot we can turn to the same conclusion.

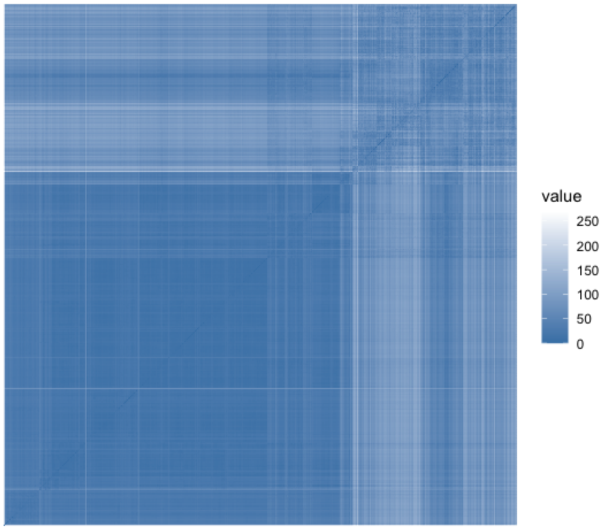


Figure 4: Dissimilarity matrix of scRNA data

For spots data:

```r
res2 = get_clust_tendency(df_2, 50, graph = TRUE, gradient = list(low = "steelblue", high = "white"))
res2$hopkins_stat # 0.9044891
res2$plot
```

Hopkins statistic of spots data is 0.9831106, which is larger than 0.5 and 0.9073394, so that spots data is weaker clusterable.

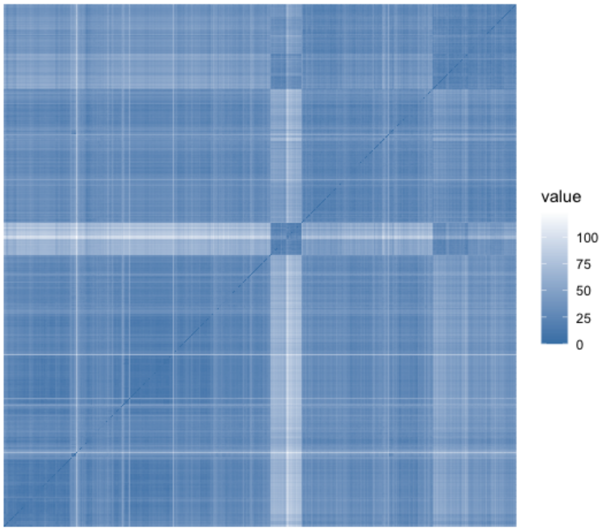


Figure 5: Dissimilarity matrix of spot data

Hopkins statistics and pictures of Kmeans cluster indicating common clustering method can not meet our requirements. So we wish to find a more accurate method for scRNA-seq and spots clustering. Besides, we also interested in the connection between scRNA-seq data and spots data.

Preliminary Model

The preliminary model focus on spots model, the main goal of it is to cluster different spots.

Notations:

(l, w), coordinate of a spot. We divide the spatial into a matrix of $L \times W$.

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

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

$X = \{x_{lwg]} : 1 \leq l \leq L, \quad 1 \leq w \leq W, \quad 1 \leq g \leq G\}$, observed data, where G stands for total number of genes.

Model assumption.

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

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

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

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

Priors: $\theta ss' \sim N(\eta_\theta, \tau_\theta^2)$,

$\mu_{sg} \sim N(\eta_\mu, \tau_\mu^2)$,

$\sigma_g^2 \sim Inv - gamma(\alpha, \beta)$.