Jointly define cell types from single cell multi-omics datasets

Jinzhuang Dou

6/28/2020

This guide will demonstrate the usage of bindSC to jointly define cell types by leveraging mulitple single-cell modalities. Here we integrated scRNA-seq and scATAC-seq to define cell types that incorporate both gene expression and chromatin accessibility data.

Stage 1: Load the scRNA-seq and scATAC-seq data

For convenience, we have prepared the pre-processed data which are ready to use. User can refer to `A549_preprocess.ATAC.html` and `A549_preprocess.RNA.html` for the details of running the pre-processing workflow (It will take 10 ~ 20 mins).

```
library("bindSC")

A549_RNA <- readRDS("../../../data/A549_rna.rds")
A549_ATAC <- readRDS("../../../data/A549_atac.rds")

summary(A549_RNA)
summary(A549_ATAC)
```

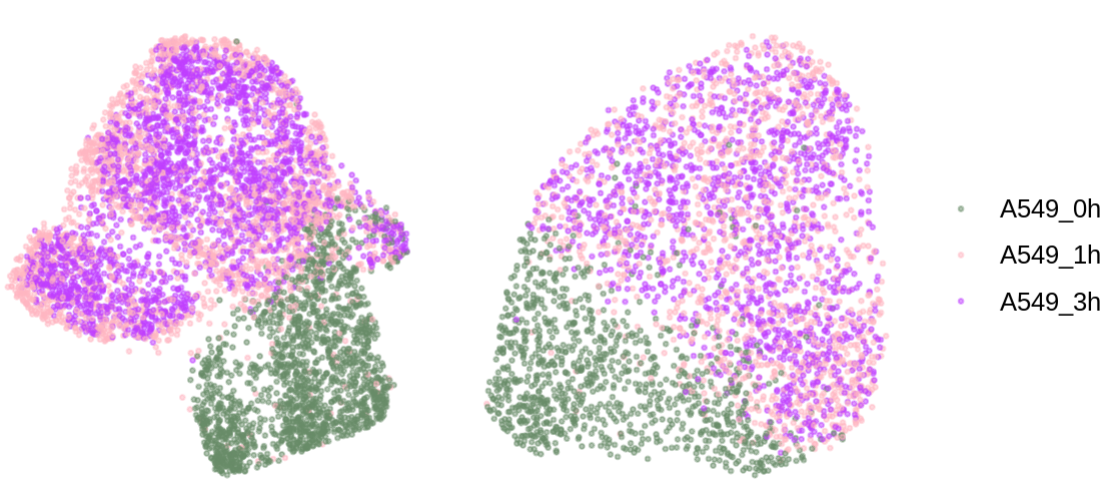
We then visualize each cell, colored by cell type, from two technologies. Left is from scRNA-seq and right is from scATAC-seq. For both technologies, cells from 0 h and 1/3 h can be well separated in 2d-UMAP.

```
library(ggpubr)
p1<-ggscatter(A549_RNA$RNA_meta, x = "UMAP_1", y = "UMAP_2",
   color = "cell_type", palette = c("darkseagreen4","lightpink","darkorchid1"),
   repel = FALSE,size=0.5, alpha=0.5,legend.title = "", title="scRNA-seq", font.title=16)

p2<-ggscatter(A549_ATAC$ATAC_meta, x = "UMAP_1", y = "UMAP_2",
   color = "group", palette = c("darkseagreen4","lightpink","darkorchid1"),
   repel = FALSE,size=0.5, alpha=0.5,legend.title = "", title="scATAC-seq", font.title=16)

p1<-p1+rremove("axis") + rremove("ticks") + rremove("xylab")+ rremove("axis.text")
p2<-p2+rremove("axis") + rremove("ticks") + rremove("xylab")+ rremove("axis.text")

p<-ggarrange(p1, p2, nrow = 1, common.legend = TRUE, legend = "right")
p
```



Stage 2: Run bindSC

We next perform bindSC on A549 data. bindSC requires three matrices as input: - `X` : gene expression matrix from scRNA-seq data - `Y` : peak matrix from scATAC-seq data - `Z0` : gene activity matrix from scATAC-seq data The gene activity matrix `Z0` can be estimated by counting the total number of ATAC-seq reads within the gene body/promoter region of each gene in each cell. More details can be seen in `A549_preprocess.ATAC.html`

```
X <- A549_RNA$X
Y <- A549_ATAC$Y
Z0 <- A549_ATAC$Z0
dim(X)
dim(Y)
dim(Z0)
```

Make sure `X` and `Z0` have matched gene features, `Y` and `Z0` have matched cell names.

```
gene.overlap <- intersect(rownames(X), rownames(Z0))
cell.overlap <- intersect(colnames(Y), colnames(Z0))

X <- as.matrix(X[gene.overlap,])
Z0 <- as.matrix(Z0[gene.overlap, cell.overlap])
Y <- as.matrix(Y[,cell.overlap])
```

Key parameters to run bindSC:

- `num.X` : Number of canonical vectors to calculate for pair matrics (X, Z0) [default `5`]
- `num.Y` : Number of canonical vectors to calculate for pair matrics (Z0, Y) [default `5`]
- `bigMemory` : Whether use the bigMemory mode. This will reduce memory usage when there are > `30K` cells/features. [Default `TRUE`]
- `block_size` : Sample size for each block. This option works only when `bigMemory` is set to `TRUE`
- `ncore` : Number of thread used [default `1`]
- `num.iteration` : Maximal number of iteration [default `100`]
- `tolerance` : Relative change ratio for `Z` during iteration [default `0.05`]
- `save` : Save the temporay files [default `FALSE`]

This process will take 30 mins with maximal memory usage being 15G.

```
out <- BiCCA(X=X, Z0=Z0, Y=Y,
       num.X = 6, num.Y = 7,
       num.iteration = 100,
       temp.path= "./tp",
       tolerance = 0.05,
       save = TRUE,
       bigMemory = TRUE, block.size = 1000)
#> 2020-07-08 09:04:59  Started!
#> 2020-07-08 09:04:59  Dimension Check: X[3774x6005] Y[24953x3628] Z0[3774x3628]
#>
#> 2020-07-08 09:48:04  Done! The decomposition is converged.
```

Stage 3: Visualization of cells from two data types in latent space

bindSC returns the list with five matrices:

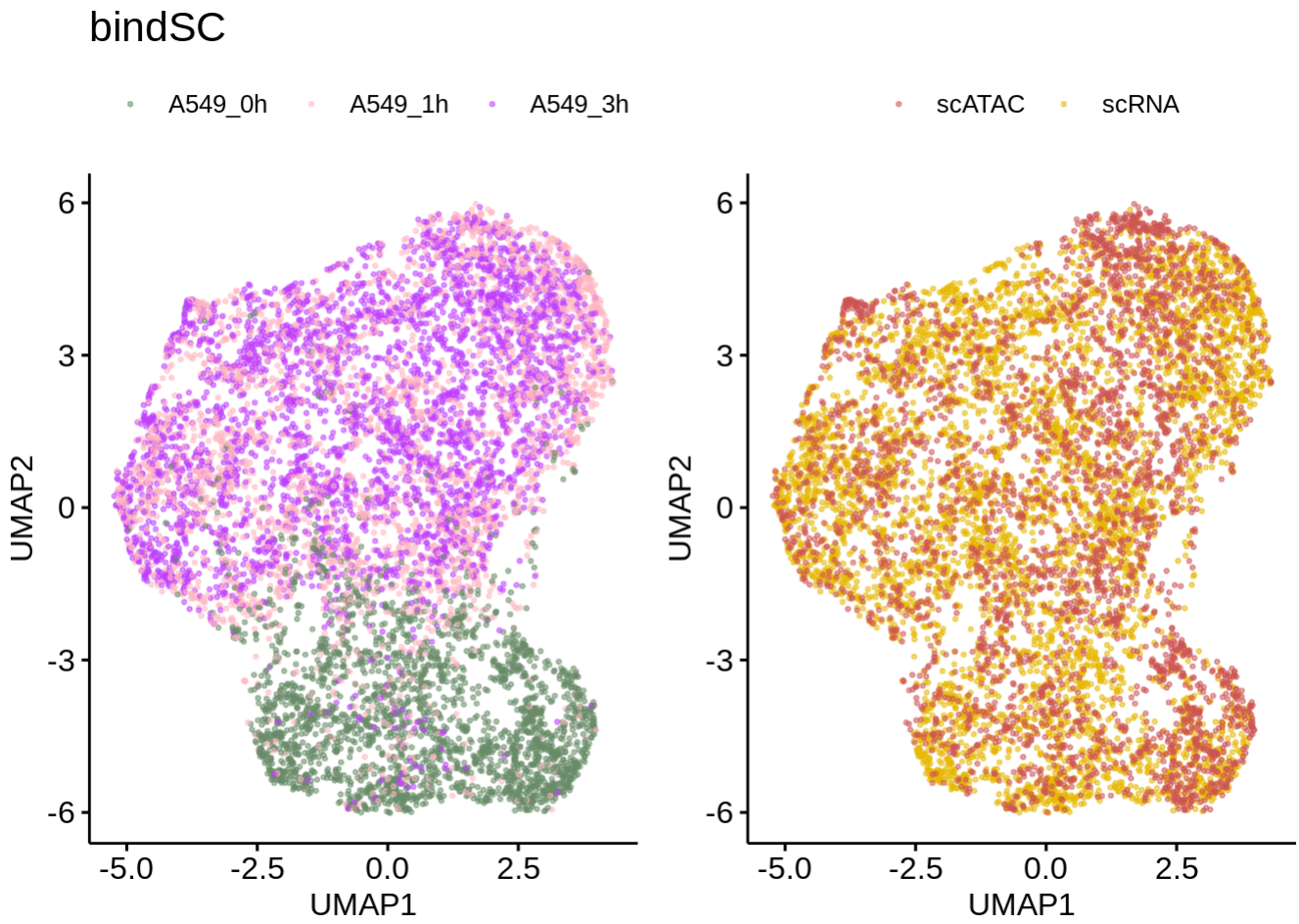
- `u` : contains the canonical correlation vectors for cells from scRNA-seq data (`K` cells by `num.X` factor)
- `r` : contains the canonical correlation vectors for cells from scATAC-seq data (`L` cells by `num.X` factor)
- `s` : contains the canonical correlation vectors for features from scRNA-seq data (`M` genes by `num.Y` factor)
- `v` : contains the canonical correlation vectors for features from scATAC-seq data (`N` loci by `num.Y` factor)
- `Z` : contains the estimated transition matrix (`M` genes by `L` cells)

bindSC also returns the relative change for `Z` ,

- `delta` : relative change ratio for `Z` during iteration

Cell coordinates for two datasets in latent space are often sufficient to define joint clusters that correpond across datasets. We used UMAP to visualize the cells in latent space colored by cell type (left figure) and by technology (right figure).

```
source("./eval_plot.r")
cell_type <- c(A549_RNA$RNA_meta$cell_type, A549_ATAC$ATAC_meta$group)
data_type <- c(rep("scRNA", dim(out$u)[1]), rep("scATAC",dim(out$r)[1]))
result <- umap_plot(out, cell_type, data_type)
result$plt
```



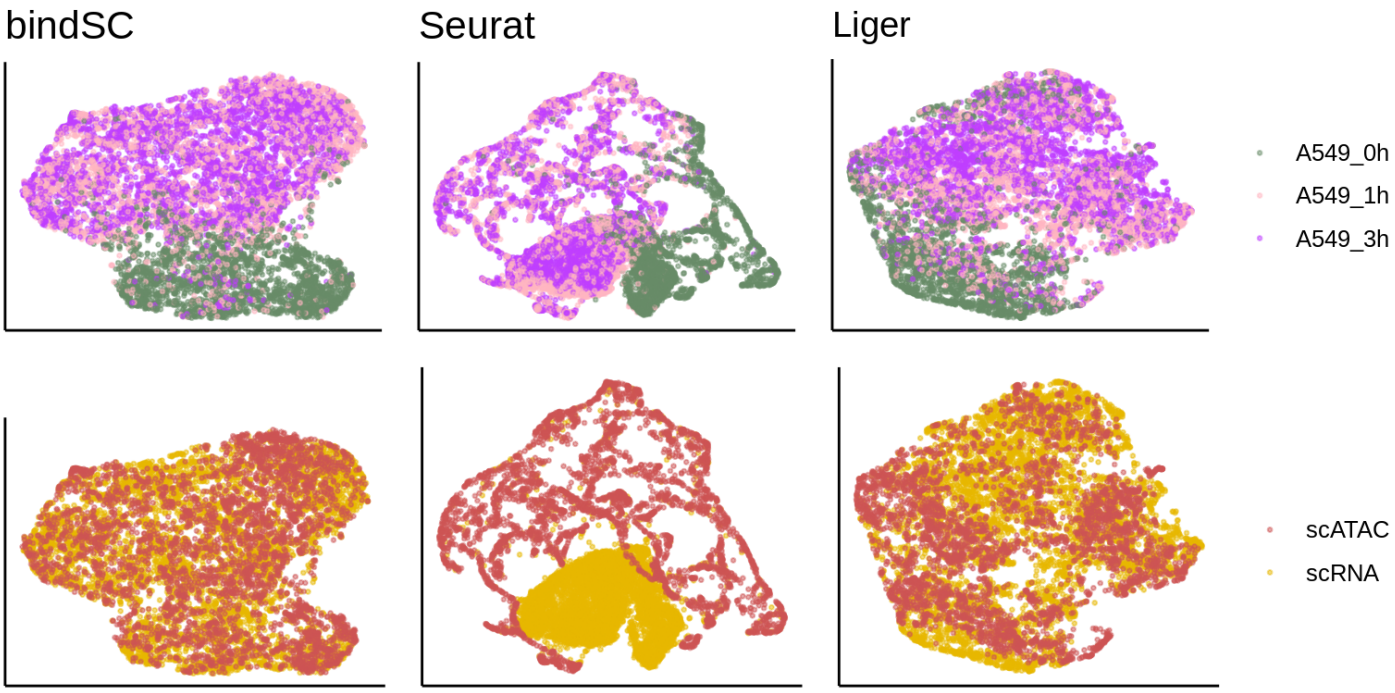
Stage 4: Comparsion among bindSC, Seurat, and Liger methods

Three metrics are used to measure method performance since we have the cell correspondence as the gold standard. **For convenience, we have prepared the pre-processed results for Seurat and Liger which are ready to use.**

- `Silhoutte coefficient` : High value means cell-type architectures is well preserved
- `Alignment score` : High value means uniformity of mixing for two datasets in the latent space
- `Anchor accuracy` : High value means cell correspondence can be found in cell's k-neighbor size (k ranges from 5 o 200)

```
eval <-  method_compare(result$plt_dt, "A549")
```

```
eval$coembeding
```




```
# Gray curve in anchor accuracy metric denotes results from random guess.
eval$eval
```