Testing unmatched feature integration using simulated data

Jinzhuang Dou

July 7, 2020

This guide will demonstrate the usage of bindSC to intergrate two datasets with unmatched features from simulation dataset.

Stage 1: Load data

For convenience, we have prepared the pre-processed data which are ready to use. User can refer to `sim.R` for the details of running the pre-processing workflow (It will take 10 mins).

```
library(bindSC)

dim(sim$X)
```

```
## [1] 400 300
```

```
dim(sim$Y)
```

```
## [1] 400 300
```

```
dim(sim$Z0)
```

```
## [1] 400 300
```

Usage of BiCCA

```
?BiCCA
```

In this example, dataset `X` is the same with `Y`. The inital transiton matrix `Z0` is generated by permutating rows of `X` with 10% probablity the same with `X`. Seurat and Liger use (`X`, `Z0`) as input and BiCCA uses (`X`, `Z0`, `Y`) as input.

The option `tolereance` is usually set from 0.01 to 0.05 to reduce unnecessary iteration when data size is large. Here I set it to 0.0001 for small sample size.

```
out <- BiCCA(X=sim$X, Z0=sim$Z0, Y=sim$Y,
       num.X = 5, num.Y = 5,
       num.iteration = 100,
       temp.path = "./tp",
       tolerance = 0.0001,
       save = FALSE,
       bigMemory = TRUE,
       block.size = 1000,
       ncore = 1)
```

```
## 2020-07-08 13:59:36  Started!
```

```
## 2020-07-08 13:59:36  Dimension Check: X[400x300] Y[400x300] Z0[400x300]
```

```
##
```

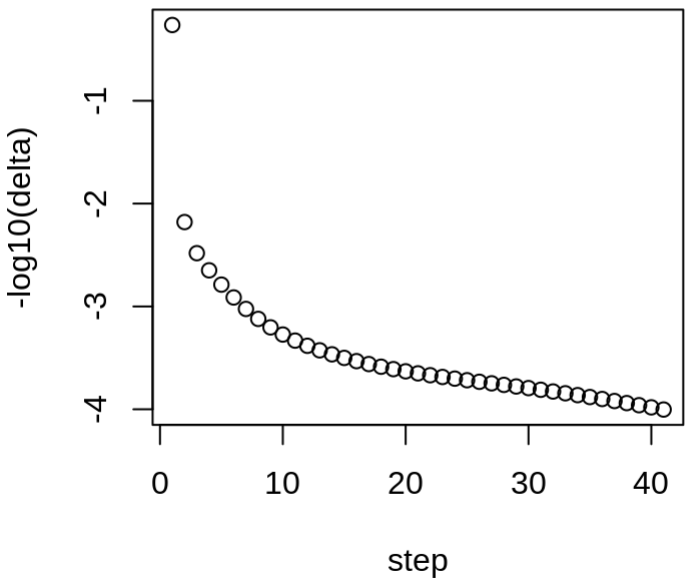
```
## 2020-07-08 14:00:05  Done! The decomposition is converged.
```

```
summary(out)
```

```
##       Length Class  Mode
## u       1500 -none- numeric
## r       1500 -none- numeric
## s       2000 -none- numeric
## v       2000 -none- numeric
## Z     120000 -none- numeric
## delta     41 -none- numeric
```

Show the iteration index `delta` (The iteration will stop in the second step if we set it to be 0.05)

```
plot(log10(out$delta), xlab="step", ylab="-log10(delta)")
```



Stage 2: 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 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 neighbor given fixed neighbor size

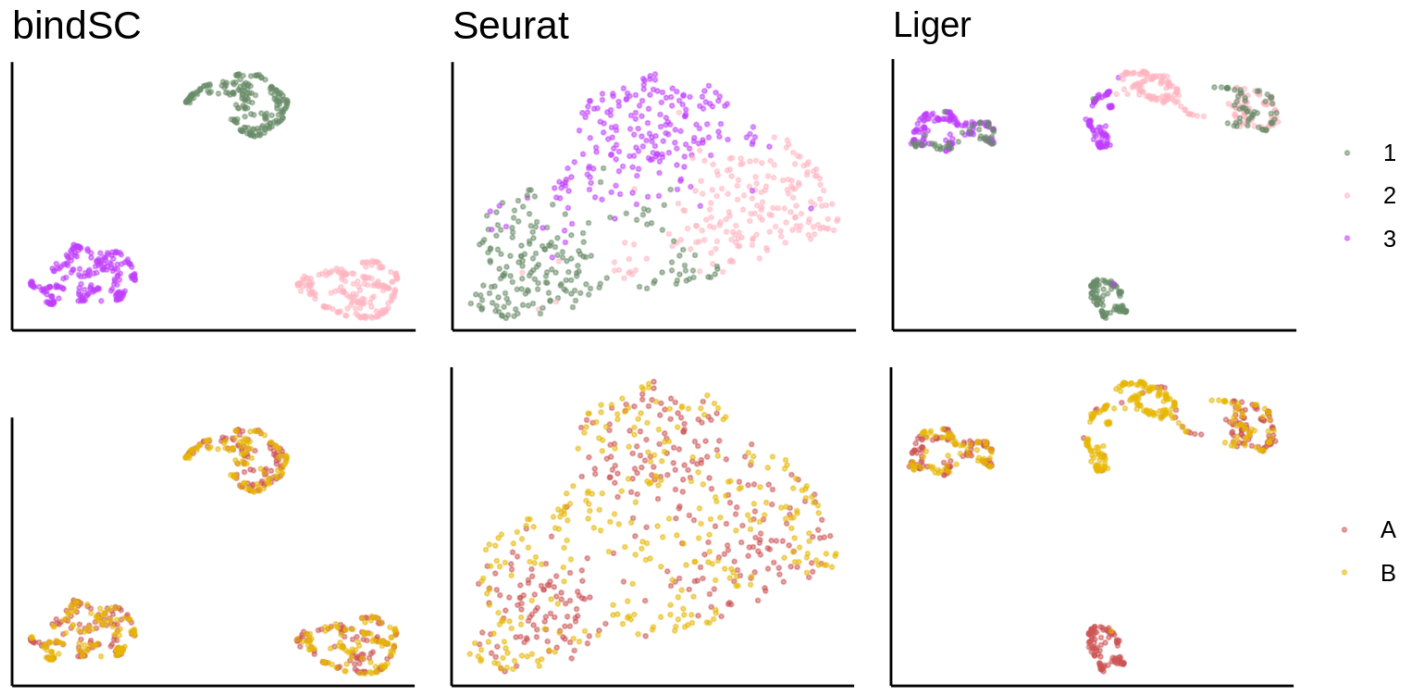
```
source("../A549/eval_plot.r")
```

```
## Loading required package: ggplot2
```

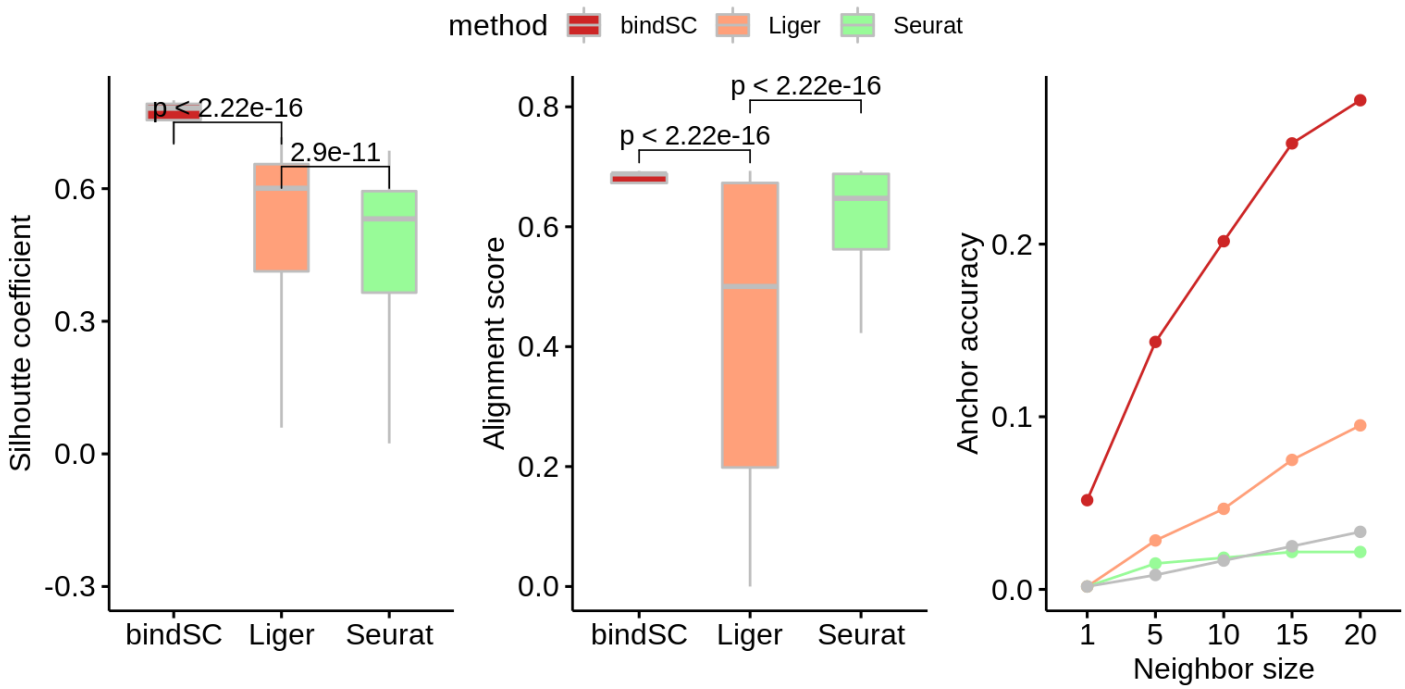
```
## Loading required package: magrittr
```

```
cell_type <- c(sim$X_meta$Group, sim$Y_meta$Group)
data_type <- c(rep("A", dim(out$u)[1]), rep("B",dim(out$r)[1]))
result <- umap_plot(out, cell_type, data_type)
eval <-  method_compare(result$plt_dt, "Sim")
```

```
eval$coembeding
```


bindSC
Seurat
Liger
1
2
3
A
B

```
eval$eval
```


method
bindSC
Liger
Seurat
p < 2.22e-16
2.9e-11
Silhoutte coefficient
0.6
0.3
0.0
-0.3
bindSC
Liger
Seurat
p < 2.22e-16
p < 2.22e-16
Alignment score
0.8
0.6
0.4
0.2
0.0
bindSC
Liger
Seurat
Anchor accuracy
0.2
0.1
0.0
1
5
10
15
20
Neighbor size