Reproducibility in the Application to the Mouse Cerebellum Slide-seq Data

This file contains instructions for reproducing the results and figures in the application to the mouse cerebellum Slide-seq data. The codes are in "code" folder. Please set the working directory to the source file location ("code" folder).

The following code files can be directly run as their input data are in the "input\_data" and "result\_data" folders: Step2\_BACT.R, Step7\_Figure4.R and Step8\_DEanalysis.R.

**Data**

The mouse cerebellum Slide-seq data is collected by Rodriques et al. (2019). This dataset contains 24,847 cells and 18,906 genes. The barcode file “Puck\_180430\_1.tar.gz.” is downloaded from the Broad institute’s single-cell repository

<https://singlecell.broadinstitute.org/single_cell/study/SCP354/slide-seq-study#study-download>.

The dataset contains the csv file “BeadLocationsForR.csv” which provides the spatial coordinates of all cells, and the csv file “MappedDGEForR.csv” which is the ST gene raw count matrix.

In the data preprocessing procedure, we randomly chose 8000 cells, log-normalized the count data, picked 5000 top HVGs and selected 50 top principal components.

The preprocessed data includes:

* "gene\_raw\_counts.csv": The raw ST count data submatrix of the 8000 chosen cells.
* "coordinates.csv": The spatial coordinates of the 8000 chosen cells.
* "processed\_gene\_data\_5000HVGs.csv": The processed data of the 5000 top HVGs derived from the log-normalized gene data of the 8000 chosen cells.
* "coord\_and\_pc.RData": The spatial coordinates and preprocessed gene data matrix.

**Code**

**Step 1: "Step1\_preprocessing.R"**

Preprocess the ST raw count data matrix. The output of this step is "gene\_raw\_counts.csv," "coordinates.csv," "processed\_gene\_data\_5000HVGs.csv," and "coord\_and\_pc.RData."

**Step 2: "Step2\_BACT.R"**

Implement BACT on the data from Step 1, and conduct the MCMC posterior inference. The output of this step is "slideseq\_result\_BACT.csv" which saves the estimated cluster labels for all the cells.

**Step 3: "Step3\_SpaGCN.R"**

Implement SpaGCN model. The output of this step is "slideseq\_result\_SpaGCN.csv".

In the "slideseq\_result\_SpaGCN.csv":

* "coord\_x": first dimension coordinate.
* "coord\_y": second dimension coordinate.
* "pred": cell typing labels.

**Step 4: "Step4\_STAGATE.py"**

Implement STAGATE model. The output of this step is "slideseq\_result\_STAGATE.csv".

In the "slideseq\_result\_STAGATE.csv":

* "coord\_x": first dimension coordinate.
* "coord\_y": second dimension coordinate.
* "mclust": cell typing labels obtained by R package mclust.

**Step 5: "Step5\_BANKSY.py"**

Implement BANKSY model. The output of this step is "slideseq\_result\_BANKSY.csv".

In the "slideseq\_result\_BANKSY.csv":

* "coord\_x": first dimension coordinate.
* "coord\_y": second dimension coordinate.
* "mclust": cell typing labels obtained by R package mclust.

**Step 6: "Step6\_BASS.R"**

Implement BASS model. The output of this step is "slideseq\_result\_BASS.csv".

In the "slideseq\_result\_BASS.csv":

* "x": first dimension coordinate.
* "y": second dimension coordinate.
* "c": estimated cell typing labels obtained via the MCMC samples.
* "z": estimated region labels obtained via the MCMC samples.

**Step 7: "Step7\_Figure4.R"**

Draw Figure 4 in the manuscript using the cell typing results of all methods.

**Step 8: "Step8\_DEanalysis.R"**

Conduct pathway analysis based on the differentially expressed genes selected from the cell typing result of BACT.