Reproducibility in the Application to the Human Dorsolateral Prefrontal Cortex Data

This file contains instructions for reproducing the results and figures in the application to the human dorsolateral prefrontal cortex 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: Step3\_BACT.R and Step8\_FigureS7\_S8.R.

**Data**

The human dorsolateral prefrontal cortex data (DLPFC) consist of 12 brain sections with manual annotation collected from three subjects Maynard et al. (2021). The section 151507 is used for detailed analysis, whose raw count matrix contains 4,226 spots and 33,538 genes. The data file “151507.h5ad” is downloaded from the link

<http://sdmbench.drai.cn/tcm/download/?file_path=/mnt/JINGD/data/file/sdmbench/db/151507.h5ad>.

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

The preprocessed data includes:

* "gene\_matrix\_raw\_151507.csv": The raw ST data matrix saved from the h5ad file.
* "coordinates.csv\_151507": The spatial coordinates saved from the h5ad file.
* "coord\_and\_pc\_151507.RData": The spatial coordinates and preprocessed gene data matrix.

**Code**

**Step 1: "Step1\_read\_data.py"**

Read data from the h5ad file. The output of this step is "gene\_matrix\_raw\_151507.csv" and "coordinates\_151507.csv."

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

Preprocess the ST raw count data matrix using the function DataPreprocess() in the R package BACT. The output of this step is "coord\_and\_pc\_151507.RData."

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

Implement BACT on the data from Step 2, and conduct the MCMC posterior inference. The output of this step is "dlpfc151507\_result\_BACT.csv" which saves the estimated cluster labels for all the cells.

**Step 4: "Step4\_SpaGCN.R"**

Implement SpaGCN model. The output of this step is "dlpfc151507\_result\_SpaGCN.csv".

In the "dlpfc151507\_result\_SpaGCN.csv":

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

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

Implement STAGATE model. The output of this step is "dlpfc151507\_result\_STAGATE.csv".

In the "dlpfc151507\_result\_STAGATE.csv":

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

**Step 6: "Step6\_BANKSY.py"**

Implement BANKSY model. The output of this step is "dlpfc151507\_result\_BANKSY.csv".

In the "dlpfc151507\_result\_BANKSY.csv":

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

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

Implement BASS model. The output of this step is "dlpfc151507\_result\_BASS.csv".

In the "dlpfc151507\_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 8: "Step8\_FigureS7\_S8.R"**

Draw Figures S7 and S8 in the supplementary materials using the cell typing results of all methods.