README document for manuscript “BACT: nonparametric Bayesian cell typing for single-cell spatial transcriptomics data”

1. **Data**
   1. Abstract

The mouse visual cortex STARmap\* data is collected by Wang et al. (2018). STARmap is an imaging-based spatial transcriptomics technique that enables the single-cell resolution detection of gene expressions in mouse visual cortex tissue samples, and STARmap\* is an extension of STARmap to include a larger set of target genes and cover a larger spatial domain. STARmap\* dataset comprises an ST count matrix, a cell coordinate matrix, and annotations for all cells. The ST count matrix has 1,207 cells and 1,020 genes.

The mouse hypothalamic preoptic region MERFISH data consist of five tissue sections from the mouse hypothalamic preoptic region collected by Chen et al. (2015). Each of the five datasets consists of around 5500 cells and 155 genes. Notice that the raw data have been normalized. The MERFISH\_0.19 is used for visualizing the cell typing performances of all methods, whose ST data matrix contains 5,803 cells and 155 genes.

The mouse cerebellum Slide-seq data is collected by Rodriques et al. (2019). The authors applied Slide-seq to profile gene expressions at the single-cell resolution in mouse cerebellum tissue samples. This dataset contains 24,847 cells and 18,906 genes.

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.

* 1. Availability

The mouse visual cortex STARmap\* data are publicly available for download via the website provided by Yuan et al. (2024) at

<http://sdmbench.drai.cn/tcm/download/?file_path=/mnt/JINGD/data/file/sdmbench/db/STARmap_20180505_BY3_1k.h5ad>.

No registration is required. The h5 file “STARmap.h5ad” contains the gene expression data and the spatial coordinates of all cells. The cell type annotation is available in the tutorial of STAGATE (Dong and Zhang, 2022):

<https://stagate.readthedocs.io/en/latest/T9_STARmap.html>,

which is provided at <https://drive.google.com/drive/folders/1I1nxheWlc2RXSdiv24dex3YRaEh780my?usp=sharing>.

The mouse hypothalamic preoptic region MERFISH data are publicly available for download via the website provided by Yuan et al. (2024).

MERFISH\_0.04: <http://sdmbench.drai.cn/tcm/download/?file_path=/mnt/JINGD/data/file/sdmbench/db/MERFISH_0.04.h5ad>.

MERFISH\_0.09:

<http://sdmbench.drai.cn/tcm/download/?file_path=/mnt/JINGD/data/file/sdmbench/db/MERFISH_0.09.h5ad>.

MERFISH\_0.14:

<http://sdmbench.drai.cn/tcm/download/?file_path=/mnt/JINGD/data/file/sdmbench/db/MERFISH_0.14.h5ad>.

MERFISH\_0.19:

<http://sdmbench.drai.cn/tcm/download/?file_path=/mnt/JINGD/data/file/sdmbench/db/MERFISH_0.19.h5ad>.

MERFISH\_0.24:

<http://sdmbench.drai.cn/tcm/download/?file_path=/mnt/JINGD/data/file/sdmbench/db/MERFISH_0.24.h5ad>.

No registration is required. The h5 file “MERFISH\_0.xx.h5ad” contains the gene expression data, the spatial coordinates of all cells, and the cell type annotations.

The mouse cerebellum Slide-seq data are publicly available for download from the

Broad institute’s single-cell repository

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

and download the barcode file “Puck\_180430\_1.tar.gz.” 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.

The DLPFC section 151507 data is publicly available for download via the website provided by Yuan et al. (2024) at

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

No registration is required. The h5 file “151507.h5ad” contains the gene expression data, the spatial coordinates of all cells, and the cell type annotations.

* 1. Description

**Mouse visual cortex STARmap\* data**

Citation: Wang X., W. E. Allen, M. A. Wright, E. L. Sylwestrak, N. Samusik, S. Vesuna, K. Evans, C. Liu, C. Ramakrishnan, J. Liu, G. P. Nolan, F.-A. Bava, and K. Deisseroth. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science, 361(6400): eaat5691, 2018.

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

**Mouse hypothalamic preoptic region MERFISH data**

Citation: Chen K. H., A. N. Boettiger, J. R. Moffitt, S. Wang, and X. Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. Science, 348(6233):aaa6090, 2015.

In the data preprocessing procedure, since the raw data have been normalized, we directly took log and selected 50 top principal components.

**Mouse cerebellum Slide-seq data**

Citation: Rodriques S. G., R. R. Stickels, A. Goeva, C. A. Martin, E. Murray, C. R. Vanderburg, J. Welch, L. M. Chen, F. Chen, and E. Z. Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science, 363(6434):1463–1467, 2019.

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.

**Human dorsolateral prefrontal cortex data**

Citation: Kristen R. Maynard, Leonardo Collado-Torres, Lukas M. Weber, Cedric Uytingco, Brianna K. Barry, Stephen R. Williams, Joseph L. Catallini, Matthew N. Tran, Zachary Besich, Madhavi Tippani, Jennifer Chew, Yifeng Yin, Joel E. Kleinman, Thomas M. Hyde, Nikhil Rao, Stephanie C. Hicks, Martinowich Keri, and Jaffe Andrew E. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nature Neuroscience, 24(3):425–436, 2021.

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

1. **Code**
   1. Abstract

All of the data preprocessing and analysis in this paper were completed using R and Python. The code is provided to conduct preprocessing on the raw data, implement BACT via Markov chain Monte Carlo method, compare against competing methods SpaGCN, STAGATE and BANKSY, and generate descriptive plots.

* 1. Description

All of the R and Python scripts are available as the supplementary code.

R license information: GPL (>= 2).

Python license information: PSF.

For R and R packages, we use R version 4.1.3 (2022-03-10, “One Push-Up”). Please note that for Windows system users, the version of Rtools needs to be compatible with the version of R. The used R packages are: (System\_preparation.R)

* aricode, version="1.0.0"
* dplyr, version="1.0.8"
* RColorBrewer, version="1.1-3"
* tibble, version="3.1.6"
* tidyr, version="1.2.0"
* ggplot2, version="3.3.5"
* mclust, version="5.4.9"
* scales, version="1.1.1"
* pheatmap, version="1.0.12"
* latex2exp, version="0.9.4"
* ggbreak, version="0.1.0"
* patchwork, version="1.1.2"
* mvtnorm, version="1.1.3"
* MCMCpack, version="1.6.3"
* rBeta2009, version="1.0"
* truncnorm, version="1.0.8"
* BiocManager, version="1.30.16"
* SingleCellExperiment, version="1.16.0"
* scran, version="1.22.1"
* scater, version="1.22.0"
* BiocSingular, version="1.10.0"
* edgeR, versioin="3.36.0"

For Python and Python packages, we use Python version 3.8.5. The used Python packages are: (execute the code ''pip install -r requirements.txt'' in the command)

* scanpy, version 1.8.2
* anndata, version 0.8.0
* umap, version 0.5.3
* numpy, version 1.22.1
* scipy, version 1.4.1
* pandas, version 1.3.4
* scikit-learn, version 1.0.2
* statsmodels, version 0.13.2
* python-igraph, version 0.9.11
* louvain, version 0.7.1
* pynndescent, version 0.5.7
* squidpy, version 1.2.2
* tqdm, version 4.61.1

For SpaGCN,

(execute the code ''pip install -r requirements\_spagcn.txt'' in the command)

* Python >= 3.7.0
* torch, version 1.10.2

For STAGATE,

(execute the code ''pip install -r requirements\_stagate.txt'' in the command)

* tensorflow, version 1.15

For BANKSY,

* Python >= 3.8
* tensorflow, version 1.15

For BASS,

* R >= 4.0.3
* GIGrvg (>= 0.5),
* Matrix (>= 1.3.4),
* harmony (>= 1.2.1),
* label.switching (>= 1.8),
* mclust (>= 5.4.7),
* Rcpp (>= 1.0.8),
* RcppArmadillo (>= 0.10.8.1.0),
* RcppDist (>= 0.1.1),
* SPARK (>= 1.1.1),
* scran (>= 1.20.1),
* scater (>= 1.16.2)

A MacBook Pro was used for the real application analyses in this paper. The details of the computer are:

* Operating system: MacOS Catalina 10.15.5
* CPU: Intel Core i5 2GHz
* RAM: 16GB

The computing platform was used for the real application analyses in this paper. The details of the computing platform are:

* Operating system: CentOS 7.8.2003
* CPU: Intel Gold 5218 (16 cores, 32 threads) 2.3GHz
* RAM: 192GB
  1. Packages installation

**BACT**

Install BINRES package by conducting the following code in R environment.

> install.packages("PKG\_PATH/BACT\_1.0.tar.gz", repos = NULL, type="source")

**SpaGCN**

SpaGCN is a graph-convolutional-network-based spatial cell typing method (Hu et al., 2021). The Python code of SpaGCN is publicly available on GitHub <https://github.com/jianhuupenn/SpaGCN>. The following code of installation is provided in the tutorial code file “SpaGCN/tutorial/[tutorial.ipynb](http://localhost:8888/notebooks/github/[Hu2021] SpaGCN/SpaGCN/tutorial/tutorial.ipynb" \t "/Users/yyq/Documents\\x/_blank).”

cd SpaGCN/SpaGCN\_package/

python3 setup.py install --user

**STAGATE**

STAGATE is another spatial cell typing method based on a graph attention auto-encoder to obtain low-dimensional latent embeddings of cells (Dong and Zhang, 2022). The Python code of STAGATE is publicly available on GitHub <https://github.com/zhanglabtools/STAGATE>. The following code of installation is provided in the README.md file.

cd STAGATE-main

python setup.py build

python setup.py install

**BANKSY**

BANSKY unifies cell type and tissue domain detection through the cell embedding in a product space of local neighborhood transcriptomics (Singhal et al., 2024). The Python code of BANKSY is publicly available on GitHub <https://github.com/prabhakarlab/Banksy_py>. The following code of installation is provided in the README.md file.

cd Banksy\_py

pip install -r requirements.txt

**BASS**

BANSK BASS is able to perform cell type identification and domain detection simultaneously with multiple tissue sections (Li and Zhou, 2022). The R code of BASS is publicly available on GitHub <https://github.com/zhengli09/BASS>. The following code of installation is provided in the README.md file.

if(!require(devtools))

install.packages(devtools)

devtools::install\_github("zhengli09/BASS")

* 1. Instructions for Use

All data preprocessing and analysis as well as Figure 1 in the manuscript and Tables Sxxx, Sxxx in the supplementary materials can be reproduced.

Detailed workflow information is contained in the "README.docx" in "Real\_application" directory. One should firstly check and install the R and Python by conducting the code files "System\_preparation.R," "requirements.txt," "requirements\_spagcn.txt," and "requirements\_stagate.txt," respectively.

The general steps in the first real application (mouse visual cortex STARmap\* data) are:

1. Read date from h5ad file.

2. Conduct data preprocessing.

3. Apply BACT to the preprocessed data. There are 6,000 iterations in the MCMC with 3,000 burn-in steps. Total execution time is about 15 minutes on a MacBook Pro with Intel Core i5 CPU at 2GHz and 16GB of RAM.

4. Implement the SpaGCN, STAGATE, BANKSY, and BASS.

5. Generate Figure 2b-h in the manuscript.

The general steps in the second real application (mouse hypothalamic preoptic region MERFISH data) are:

1. Read date from h5ad files. MERFISH\_0.19 is used for visualization.

2. Conduct data preprocessing.

3. Apply BACT to the preprocessed data. There are 4,000 iterations in the MCMC with 2,000 burn-in steps. For MERFISH\_0.19, the total execution time is about 30 minutes on a MacBook Pro with Intel Core i5 CPU at 2GHz and 16GB of RAM.

4. Implement the SpaGCN, STAGATE, BANKSY, and BASS.

5. Generate Figures 3 in the manuscript, and Supplementary Figure S1 in the supplementary materials.

The general steps in the third real application (mouse cerebellum Slide-seq data) are:

1. Conduct data preprocessing.

2. Apply BACT to the preprocessed data. There are 8,000 iterations in the MCMC with 4,000 burn-in steps. Total execution time is about 1.5 hours on a MacBook Pro with Intel Core i5 CPU at 2GHz and 16GB of RAM.

3. Implement the SpaGCN, STAGATE, BANKSY, and BASS.

4. Generate Figure 4b-f in the manuscript, and Supplementary Figure S2 in the supplementary materials.

5. Carry out pathway analysis based on the differentially expressed genes.

The general steps in the fourth real application (human dorsolateral prefrontal cortex data) are:

1. Read date from h5ad file.

2. Conduct data preprocessing.

3. Apply BACT to the preprocessed data. There are 6,000 iterations in the MCMC with 3,000 burn-in steps. Total execution time is about 1.2 hours on a MacBook Pro with Intel Core i5 CPU at 2GHz and 16GB of RAM.

4. Implement the SpaGCN, STAGATE, BANKSY, and BASS.

5. Generate Supplementary Figures S7 and S8 in the supplementary materials.

* 1. Remark

When installing the R package BACT, if users encounter an error related to the C++ compiler that RStudio overwrites part of "Makevars" and "Makevars.win" in a way that bypasses a default compiler, then add the following command to "Makevars" and "Makevars.win" to avoid this issue.

"CXX = clang++ -mmacosx-version-min=10.13 -std=gnu++11"

The discussion about this issue can be found in the GitHub link

<https://github.com/RcppCore/RcppArmadillo/issues/324>.