Package 'SpaTopic'

April 23, 2024

Type Package

Title Topic Inference to Identify Tissue Architecture in Multiplexed Images

Version 1.1.0

Date 2024-04-22

Description A novel spatial topic model to integrate both cell type and spatial information to identify the complex spatial tissue architecture on multiplexed tissue images without human intervention. The Package implements a collapsed Gibbs sampling algorithm for inference. 'SpaTopic' is scalable to large-scale image datasets without extracting neighborhood information for every single cell. For more details on the methodology, see <https://xiyupeng.github.io/SpaTopic/>.

License GPL (>= 3)

Depends R (>= 3.5.0),

Imports Rcpp (>= 0.12.0), RANN (>= 2.6.0), sf (>= 1.0-12), methods (>= 3.4), foreach (>= 1.5.0), iterators (>= 1.0),

LinkingTo Rcpp, RcppArmadillo, RcppProgress,

Suggests knitr, rmarkdown, SeuratObject (>= 4.9.9.9086), doParallel (>= 1.0),

VignetteBuilder knitr

RoxygenNote 7.2.3

Encoding UTF-8

LazyData true

URL https://github.com/xiyupeng/SpaTopic

BugReports https://github.com/xiyupeng/SpaTopic/issues

NeedsCompilation yes

Author Xiyu Peng [aut, cre] (<https://orcid.org/0000-0003-4232-0910>)

Maintainer Xiyu Peng <pansypeng124@gmail.com>

Repository CRAN

Date/Publication 2024-04-22 23:50:11 UTC

R topics documented:

`gibbs.res-class` *A class of the output from 'SpaTopic'*

Description

Outputs from function `SpaTopic_inference`. A `list` contains the following members:

- `$Perplexity`. The perplexity is for the training data. Let N be the total number of cells across all images. $Perplexity = exp(-loglikelihood/N)$
- `$Deviance`. $Deviance = -2loglikelihood$.
- `$loglikelihood`. The model log-likelihood.
- `$loglike.trace`. The log-likelihood for every collected posterior sample. NULL if `trace` = FALSE.
- `$Beta`. Topic content matrix with rows as celltypes and columns as topics
- `$Theta`. Topic prevalent matrix with rows as regions and columns as topics
- `$Ndk`. Number of cells per topic (col) per region (row).
- `$Nwk`. Number of cells per topic (col) per celltype (row).
- `$Z.trace`. Number of times cell being assigned to each topic across all posterior samples. We can further compute the posterior distributions of `Z` (topic assignment) for individual cells.
- `$doc.trace`. `Ndk` for every collected posterior sample. NULL if `trace` = FALSE.
- `$word.trace`. `Nwk` for every collected posterior sample. NULL if `trace` = FALSE.

See Also

`SpaTopic_inference`

lung5 *Example input data for 'SpaTopic'*

Description

multiplexed image data on tumor tissue sample from non small cell lung cancer patient

Usage

```
lung5
```

Format

'lung5' A data frame with 100149 rows and 4 columns:

image Image ID

X X coordinate of the cell

Y Y coordinate of the cell

type cell type

Source

<https://nanostring.com/products/cosmx-spatial-molecular-imager/ffpe-dataset/nsclc-ffpe-dataset/>

See Also

SpaTopic_inference,Seurat5obj_to_SpaTopic

Seurat5obj_to_SpaTopic *Convert a Seurat v5 object as the input of 'SpaTopic'*

Description

Prepare 'SpaTopic' input from one Seurat v5 object

Usage

```
Seurat5obj_to_SpaTopic(object, group.by, image = "image1")
```

Arguments

object	Seurat v5 object
group.by	character. The name of the column that contains celltype information in the Seurat object.
image	character. The name of the image. Default is "image1".

Value

Return a data frame as the input of 'SpaTopic'

See Also

lung5

Examples

```
## nano.obj is a Seurat v5 object
#dataset<-Seurat5obj_to_SpaTopic(object = nano.obj,
#                 group.by = "predicted.annotation.l1",image = "image1")
## Expect output
data("lung5")
```

SpaTopic-Package *'SpaTopic' R package*

Description

The 'SpaTopic' R package is centered around the 'SpaTopic' algorithm to infer the spatial tissue architectures from multiplexed images.

Details

The package implements a Collapsed Gibbs sampling algorithm to infer topics, corresponding to distinct tissue microenvironments across multiple tissue images. Without obtaining the cell neighborhood info of every single cell, 'SpaTopic' runs much faster than other KNN-based methods on large-scale images.

The main functions in the 'SpaTopic' package

- Prepare input Seurat5obj_to_SpaTopic
- Model Inference SpaTopic_inference

Author(s)

Xiyu Peng <pansypeng124@gmail.com>

SpaTopic_inference *'SpaTopic': fast topic inference to identify tissue architecture in multiplexed images*

Description

This is the main function of 'SpaTopic', implementing a Collapsed Gibbs Sampling algorithm to learn topics, which referred to different tissue microenvironments, across multiple multiplexed tissue images. The function takes cell labels and coordinates on tissue images as input, and returns the inferred topic labels for every cell, as well as topic contents, a distribution over celltypes. The function recovers spatial tissue architectures across images, as well as indicating cell-cell interactions in each domain.

Usage

```
SpaTopic_inference(
  tissue,
  ntopics,
  sigma = 50,
  region_radius = 400,
  kneigh = 5,
  npoints_selected = 1,
  ini_LDA = TRUE,
  ninit = 10,
  niter_init = 100,
  beta = 0.05,
  alpha = 0.01,
  trace = FALSE,
  seed = 123,
  thin = 20,
  burnin = 1000,
  niter = 200,
  display_progress = TRUE,
  do.parallel = FALSE,
  n.cores = 1,
  axis = "2D"
)
```

Arguments

tissue	(Required). A data frame or a list of data frames. One for each image. Each row represent a cell with its image ID, X, Y coordinates on the image, celltype, with column names (image, X, Y, type), respectively. You may add another column Y2 for 3D tissue image.
ntopics	(Required). Number of topics. Topics will be obtained as distributions of cell types.

sigma	Default is 50. The lengthscale of the Nearest-neighbor Exponential Kernel. Sigma controls the strength of decay of correlation with distance in the kernel function. Please check the paper for more information. Need to be adjusted based on the image resolution
region_radius	Default is 400. The radius for each grid square when sampling region centers for each image. Need to be adjusted based on the image resolution and pattern complexity.
kneigh	Default is 5. Only consider the top 5 closest region centers for each cell.
npoints_selected	Default is 1. Number of points sampled for each grid square when sampling region centers for each image. Used with region_radius.
ini_LDA	Default is TRUE. Use warm start strategy for initialization and choose the best one to continue. If 0, it simply uses the first initialization.
ninit	Default is 10. Number of initialization. Only retain the initialization with the highest log likelihood (perplexity).
niter_init	Default is 100. Warm start with 100 iterations in the Gibbs sampling during initialization.
beta	Default is 0.05. A hyperparameter to control the sparsity of topic content (topic-celltype) matrix Beta. A smaller value introduces more sparse in Beta.
alpha	Default is 0.01. A hyperparameter to control the sparsity of document (region) content (region-topic) matrix Theta. For our application, we keep it very small for the sparsity in Theta.
trace	Default is FALSE. Compute and save log likelihood, Ndk, Nwk for every posterior samples. Useful when you want to use DIC to select number of topics, but it is time consuming to compute the likelihood for every posterior samples.
seed	Default is 123. Random seed.
thin	Default is 20. Key parameter in Gibbs sampling. Collect a posterior sample for every thin=20 iterations.
burnin	Default is 1000. Key parameter in Gibbs sampling. Start to collect posterior samples after 1000 iterations. You may increase the number of iterations for burn-in for highly complex tissue images.
niter	Default is 200. Key parameter in Gibbs sampling. Number of posterior samples collected for model inference.
display_progress	Default is TRUE. Display the progress bar.
do.parallel	Default is FALSE. Use parallel computing through R package foreach.
n.cores	Default is 1. Number of cores used in parallel computing.
axis	Default is "2D". You may switch to "3D" for 3D tissue images. However, the model inference for 3D tissue is still under test.

Value

Return a gibbs.res-class object. A list of outputs from Gibbs sampling.

See Also

gibbs.res-class

Examples

```
## tissue is a data frame containing cellular information from one image or
## multiple data frames from multiple images.

data("lung5")
## NOT RUN, it takes about 90s
library(sf)
#gibbs.res<-SpaTopic_inference(lung5, ntopics = 7,
#                              sigma = 50, region_radius = 400)


## generate a fake image 2 and make an example for multiple images
## NOT RUN
#lung6<-lung5
#lung6$image<-"image2"  ## The image ID of two images should be different
#gibbs.res<-SpaTopic_inference(list(A = lung5, B = lung6),
#                 ntopics = 7, sigma = 50, region_radius = 400)
```

---

stratified_sampling_sf

Spatially stratified random sample points from an image.

Description

Spatially stratified random sample points from an image by R package sf

Usage

```
stratified_sampling_sf(
  points,
  cellsize = c(600, 600),
  num_samples_per_stratum = 1
)
```

Arguments

points	a data frame contains all points in a image with X, Y coordinates.
cellsize	a vector of length 2 contains the size of each grid square. Default c(600,600).
num_samples_per_stratum	number of point selected from each grid square. Default 1.

Value

Return a vector contains index of sampled points.

Examples

```
data("lung5")
pt_idx<-stratified_sampling_sf(lung5, cellsize = c(600,600))
```

Index