Package ‘scDHA’

September 15, 2021

Type Package

Title Single-Cell Decomposition using Hierarchical Autoencoder

Version 1.1.2

Maintainer Duc Tran <duct@nevada.unr.edu>

Description Provides a fast and accurate pipeline for single-cell analyses. The 'scDHA' software package can perform clustering, dimension reduction and visualization, classification, and time-trajectory inference on single-cell data (Tran et.al. (2021) <DOI:10.1038/s41467-021-21312-2>).

License GPL-3

Encoding UTF-8

LazyData true

Depends R (>= 3.4)

Imports matrixStats, foreach, doParallel, igraph, Matrix, uwot, cluster, clusterCrit, Rcpp, RcppParallel, RcppAnnoy, methods, torch (>= 0.3.0), RhpcBLASctl

LinkingTo Rcpp, RcppArmadillo, RcppParallel, RcppAnnoy

RoxygenNote 7.1.0

Suggests testthat, knitr, mclust

NeedsCompilation yes

VignetteBuilder knitr

URL https://github.com/duct317/scDHA

BugReports https://github.com/duct317/scDHA/issues

Author Duc Tran [aut, cre],
    Tin Nguyen [fnd]

Repository CRAN

Date/Publication 2021-09-15 20:20:02 UTC
R topics documented:

- **Goolam** .................................................. 2
- **Goolam_result** ........................................... 2
- **scDHA** ....................................................... 3
- **scDHA.class** ................................................ 4
- **scDHA.pt** ..................................................... 5
- **scDHA.vis** ................................................... 6
- **scDHA.w** ..................................................... 7

Index ................................. 9

---

**Description**
Goolam dataset in list format, include scRNA-seq data and cell type information.

**Usage**
Goolam

**Format**
An object of class list of length 2.

---

**Description**
Result of processing Goolam dataset using 'scDHA' function.

**Usage**
Goolam_result

**Format**
An object of class list of length 4.
scDHA

Description

The main function to perform dimension deduction and clustering.

Usage

scDHA(
  data = data,
  k = NULL,
  method = "scDHA",
  sparse = FALSE,
  n = 5000,
  ncores = 10L,
  gen_fil = TRUE,
  do.clus = TRUE,
  sample.prob = NULL,
  seed = NULL
)

Arguments

data Gene expression matrix, with rows represent samples and columns represent genes.
k Number of clusters, leave as default for auto detection. Has no effect when do.clus = False.
method Method used for clustering. It can be "scDHA" or "louvain". The default setting is "scDHA".
sparse Boolean variable indicating whether data is a sparse matrix. The input must be a non negative sparse matrix.
n Number of genes to keep after feature selection step.
ncores Number of processor cores to use.
gen_fil Boolean variable indicating whether to perform scDHA gene filtering before performing dimension deduction and clustering.
do.clus Boolean variable indicating whether to perform scDHA clustering. If do.clus = False, only dimension deduction is performed.
sample.prob Probability used for classification application only. Leave this parameter as default, no user input is required.
seed Seed for reproducibility.
Value

List with the following keys:

- cluster - A numeric vector containing cluster assignment for each sample. If do.clus = False, this value is always NULL.
- latent - A matrix representing compressed data from the input data, with rows represent samples and columns represent latent variables.

Examples

```r
library(scDHA)
# Load example data (Goolam dataset)
data('Goolam'); data <- t(Goolam$data); label <- as.character(Goolam$label)
# Log transform the data
data <- log2(data + 1)
if(torch::torch_is_installed())# scDHA need libtorch installed
{
    # Generate clustering result, the input matrix has rows as samples and columns as genes
    result <- scDHA(data, ncores = 2, seed = 1)
    # The clustering result can be found here
    cluster <- result$cluster
}
```

```
scDHA.class

Description

Perform classification of new data based on available data.

Usage

```
scDHA.class(
    train = train,
    train.label = train.label,
    test = test,
    ncores = 10L,
    seed = NULL
)
```

Arguments

- `train` Expression matrix of available data, with rows represent samples and columns represent genes.
- `train.label` A vector containing label for each sample in training data.
test
Expression matrix new data for classification, with rows represent samples and columns represent genes.
ncores
Number of processor cores to use.
seed
Seed for reproducibility.

Value
A vector contain classified labels for new data.

Examples

library(scDHA)
#Load example data (Goolam dataset)
data('Goolam'); data <- t(Goolam$data); label <- as.character(Goolam$label)
#define transform the data
data <- log2(data + 1)
#Split data into training and testing sets
set.seed(1)
idx <- sample.int(nrow(data), size = round(nrow(data)*0.75))
train.x <- data[idx, ]; train.y <- label[idx]
test.x <- data[-idx, ]; test.y <- label[-idx]
if(torch::torch_is_installed()) #scDHA need libtorch installed
{
  #Predict the labels of cells in testing set
  prediction <- scDHA.class(train = train.x, train.label = train.y, test = test.x,
                           ncores = 2, seed = 1)
  #Calculate accuracy of the predictions
  acc <- round(sum(test.y == prediction)/length(test.y), 2)
  print(paste0("Accuracy = ", acc))
}
Value

List with the following keys:

- **pt** - Pseudo-time values for each sample.

Examples

```r
library(scDHA)
# Load example data (Goolam dataset)
data('Goolam'); data <- t(Goolam$data); label <- as.character(Goolam$label)
# Log transform the data
data <- log2(data + 1)
if(torch::torch_is_installed()) { # scDHA need libtorch installed
  result <- scDHA(data, ncores = 2, seed = 1)
  cell.stages <- c("2cell", "4cell", "8cell", "16cell", "blast")
  result <- scDHA.pt(result, start.point = 1, ncores = 2, seed = 1)
  r2 <- round(cor(result$pt, as.numeric(factor(label, levels = cell.stages)))^2, digits = 2)
}
```

---

**scDHA.vis**

*scDHA visualization*

Description

Generating 2D embedded data for visualization.

Usage

```r
scDHA.vis(sc = sc, method = "UMAP", ncores = 10L, seed = NULL)
```

Arguments

- **sc**
  - Embedding object produced by the `scDHA` function.
- **method**
  - Visualization method to use. It can be "UMAP" or "scDHA". The default setting is "UMAP".
- **ncores**
  - Number of processor cores to use.
- **seed**
  - Seed for reproducibility.
**Value**

A list with the following keys:

- **pred** - A matrix representing the 2D projection of single-cell data, where rows represent samples and columns represent latent components.

**Examples**

```r
library(scDHA)
# Load example data (Goolam dataset)
data('Goolam'); data <- t(Goolam$data); label <- as.character(Goolam$label)
# Log transform the data
data <- log2(data + 1)
if(torch::torch_is_installed()) {  # scDHA need libtorch installed
  # Generate clustering result, the input matrix has rows as samples and columns as genes
  result <- scDHA(data, ncores = 2, seed = 1)
  # Generate 2D representation, the input is the output from scDHA function
  result <- scDHA.vis(result, ncores = 2, seed = 1)
  # Plot the representation of the dataset, different colors represent different cell types
  plot(result$pred, col=factor(label), xlab = "scDHA1", ylab = "scDHA2")
}
```

**Description**

This function will plot a graph with normalized weights of all genes so user can select the appropriate number of genes to keep.

**Usage**

```r
scDHA.w(data = data, sparse = FALSE, ncores = 10L, seed = NULL)
```

**Arguments**

- **data** - Gene expression matrix, with rows represent samples and columns represent genes.
- **sparse** - Boolean variable indicating whether data is a sparse matrix. The input must be a non negative sparse matrix.
- **ncores** - Number of processor cores to use.
- **seed** - Seed for reproducibility.

**Value**

A plot with normalized weights of all genes.
Examples

library(scDHA)
# Load example data (Goolam dataset)
data('Goolam'); data <- t(Goolam$data); label <- as.character(Goolam$label)
# Log transform the data
data <- log2(data + 1)
if(torch::torch_is_installed()) # scDHA need libtorch installed
{
  # Generate weight variances for each genes
  weight_variance <- scDHA.w(data, ncores = 2, seed = 1)
  # Plot weight variances for top 5,000 genes
  # plot(weight_variance, xlab = "Genes", ylab = "Normalized Weight Variance", xlim=c(1, 5000))
  # Plot the change of weight variances for top 5,000 genes
  weight_variance_change <- weight_variance[-length(weight_variance)] - weight_variance[-1]
  # plot(weight_variance_change, xlab = "Genes", ylab = "Weight Variance Change", xlim=c(1, 5000))
}
Index

* datasets
  Goolam, 2
  Goolam_result, 2

Goolam, 2
Goolam_result, 2

scDHA, 3
scDHA.class, 4
scDHA.pt, 5
scDHA.vis, 6
scDHA.w, 7