Package ‘CIDER’

November 19, 2021

Type Package

Title Meta-Clustering for Single-Cell Data Integration and Evaluation

Version 0.99.0

Maintainer Zhiyuan Hu <zhiyuan.cheryl.hu@gmail.com>

Description A workflow of (a) meta-clustering based on inter-group similarity measures and (b) a ground-truth-free test metric to assess the biological correctness of integration in real datasets. See Hu Z, Ahmed A, Yau C (2021) <doi:10.1101/2021.03.29.437525> for more details.


BugReports https://github.com/zhiyhu/CIDER/issues

Imports limma (>= 3.42.0), edgeR (>= 3.28.0), stats (>= 3.6.2), foreach (>= 1.4.7), Seurat (>= 3.1.0), utils (>= 3.6.2), pheatmap (>= 1.0.0), dbscan (>= 1.0-5), kernlab (>= 0.9-29), doParallel, igraph, parallel, graphics, ggplot2, viridis

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.1.2

LazyData true

Suggests knitr, rmarkdown, testthat, statmod (>= 1.2.2), cowplot

Depends R (>= 3.5.0)

NeedsCompilation no

Author Zhiyuan Hu [aut, cre] (<https://orcid.org/0000-0002-1688-6032>), Christopher Yau [aut] (<https://orcid.org/0000-0001-7615-8523>)

Repository CRAN

Date/Publication 2021-11-19 14:40:08 UTC
**R topics documented:**

- `calculateDistMatOneModel` ......................................................... 2
- `cosineSimilarityR` ..................................................................... 3
- `downsampling` ............................................................................. 3
- `estimateProb` ............................................................................... 4
- `finalClustering` .............................................................................. 5
- `gatherInitialClusters` .................................................................. 6
- `getDistMat` ...................................................................................... 6
- `getGroupFit` ..................................................................................... 8
- `getIDEr` ............................................................................................ 8
- `hdbscan.seurat` ............................................................................... 9
- `initialClustering` ........................................................................... 10
- `measureSimilarity` ......................................................................... 11
- `mergeInitialClusters` ..................................................................... 12
- `pancreas` .......................................................................................... 13
- `plotDistMat` ....................................................................................... 13
- `plotHeatmap` ....................................................................................... 14
- `plotNetwork` ....................................................................................... 14
- `scatterPlot` ......................................................................................... 15

**Index** 17

---

**`calculateDistMatOneModel`**

*Calculate distance matrix with in one model*

**Description**

This function is called by `getDistMat`.

**Usage**

```r
calculateDistMatOneModel(
  matrix,
  metadata,
  verbose = TRUE,
  method = "voom",
  additional.variate = NULL
)
```

**Arguments**

- **matrix**: The count matrix. Rows are genes/features and columns are samples/cells.
- **metadata**: Data frame. Its rows should correspond to columns of the `matrix` input.
- **verbose**: Print the message and progress bar (default: TRUE)
- **method**: Methods for DE analysis. Options: "voom" or "trend" (default)
- **additional.variate**: additional variate to include into the linear model to regress out
cosineSimilarityR

Value
A similarity matrix

Author(s)
Zhiyuan Hu

See Also
This function is called by getDistMat

cosineSimilarityR  cosine similarity in R

description

Description
cosine similarity in R

Usage
cosineSimilarityR(x)

Arguments
x  a matrix

Value
a similarity matrix among all rows of the input matrix

downsampling  Downsampling cells

description

Description
Downsampling cells from each group for IDER-based similarity calculation.

Usage
downsampling(
  metadata,
  n.size = 35,
  seed = 12345,
  include = FALSE,
  replace = FALSE,
  lower.cutoff = 3
)

Arguments

metadata  Data frame. It includes at least 2 columns, label and batch. Each row corresponds to one cell. Required.
n.size  Numeric. The number of cells used in each group. (Default: 35)
seed  Numeric. Seed used to sample. (Default: 12345)
include  Boolean. Using ‘include = TRUE’ to include the group smaller than required size. (Default: FALSE)
replace  Boolean. Using ‘replace = TRUE’ if the group is smaller than required size and some cells will be repeatedly used. (Default: FALSE)
lower.cutoff  Numeric. The minimum size of groups to keep. (Default: 3)

Value

A numeric list of which cells will be kept for downstream computation.

Description

Estimate the empirical probability of whether two set of cells from distinct batches belong to the same population

Usage

estimateProb(seu, ider, n_size = 40, n.perm = 5, verbose = FALSE)

Arguments

seu  A Seurat object
ider  The output list of function ‘getIDEr’.
n.size  Number of cells per group used to compute the similarity. Default: 40
n.perm  Numeric. Time of permutations.
verbose  Boolean. Print out progress or not. (Default: FALSE)

Value

A Seurat object with IDER-based similarity and empirical probability of rejection

See Also

Usage of this function should be after hdbscan.seurat and getIDEr
finalClustering

**Final clustering step for meta-clustering**

### Description

Merge initial clusters into final clusters based on the matrix of IDEr.

### Usage

```r
finalClustering(
  seu,  # Seurat S4 object after the step of 'getIDEr'. Required.
  dist,  # A list. Output of 'getIDEr'. Required.
  cutree.by = "h",  # Character. Cut the tree by which parameter, height ("h") or number of clusters ("k"). (Default: h)
  cutree.h = 0.45,  # Numeric between 0 and 1. The height used to cut the tree. Ignored if 'cutree.by = 'k'. (Default: 0.45)
  cutree.k = 3,  # Numeric/integer. Used to cut the tree. Ignored if 'cutree.by = 'h'. (Default: 3)
  hc.method = "complete"  # Character. Used to choose the hierarchical clustering method.
)
```

### Arguments

- **seu**: Seurat S4 object after the step of `getIDEr`. Required.
- **dist**: A list. Output of `getIDEr`. Required.
- **cutree.by**: Character. Cut the tree by which parameter, height ("h") or number of clusters ("k"). (Default: h)
- **cutree.h**: Numeric between 0 and 1. The height used to cut the tree. Ignored if `cutree.by = 'k'`. (Default: 0.45)
- **cutree.k**: Numeric/integer. Used to cut the tree. Ignored if `cutree.by = 'h'`. (Default: 3)
- **hc.method**: Character. Used to choose the hierarchical clustering method.

### Value

Seurat S4 object with final clustering results in `CIDER_clusters` of meta.data.

### See Also

- `getIDEr`

### Examples

```r
library(CIDER)
data("pancreas")
ider <- getIDEr(pancreas, downsampling.size = 30)
seu <- finalClustering(pancreas, ider)
head(seu$CIDER_cluster)
```
gatherInitialClusters  *Gather initial cluster names*

**Description**
Gather initial cluster names

**Usage**
gatherInitialClusters(seu_list, seu)

**Arguments**
- **seu_list**: A list containing Seurat objects. Required.
- **seu**: A Seurat object

**Value**
A Seurat object containing initial clustering results in ‘seu$initial_cluster’.

**Functions**
- gatherInitialClusters: initial clustering results from a Seurat object list to one Seurat object. Follows the function ‘mergeInitialClusters’.

**See Also**
mergeInitialClusters

---

getDistMat  *Calculate the Similarity Matrix*

**Description**
Compute the IDER-based similarity matrix for a list of Seurat objects. This function does not regress out batch effects and is designed to be used at the initial clustering step.
getDistMat

Usage

getAddress(
  seu_list,
  verbose = TRUE,
  tmp.initial.clusters = "seurat_clusters",
  method = "trend",
  additional.variate = NULL,
  downsampling.size = 35,
  downsampling.include = TRUE,
  downsampling.replace = TRUE
)

Arguments

seu_list A list containing Seurat objects. Required.
verbose Print the message and progress bar (default: TRUE)
tmp.initial.clusters
  One of the colnames from ‘Seurat@meta.data’. Used as the group. Default: "seurat_clusters"
method Methods for DE analysis. Options: "voom" or "trend" (default)
additional.variate
  additional variate to include into the linear model to regress out
downsampling.size
  Number of cells used per group. Default: 35
downsampling.include
  Whether to include the group of size smaller than ‘downsampling.size’. Default: TRUE
downsampling.replace
  Whether to use ‘replace’ in sampling for group of size smaller than ‘downsampling.size’ if they are kept. Default: TRUE

Value

A list of similarity matrices

Author(s)

Zhiyuan Hu

See Also

calculateDistMatOneModel
**getGroupFit**  
*Calculate IDER-based similarity between two groups*

**Description**
Calculate IDER-based similarity between two groups

**Usage**
getGroupFit(logCPM, design, contrast_m)

**Arguments**
- **logCPM**: logCPM
- **design**: design
- **contrast_m**: contrast matrix

**Value**
Numeric. The IDER-based similarity between two groups.

---

**getIDEr**  
*Compute IDER-based similarity*

**Description**
Calculate the similarity matrix based on the metrics of Inter-group Differential ExpRession (IDER) with the selected batch effects regressed out.

**Usage**
getIDEr(
  seu,
  group.by.var = "initial_cluster",
  batch.by.var = "Batch",
  verbose = TRUE,
  use.parallel = FALSE,
  n.cores = 1,
  downsampling.size = 40,
  downsampling.include = TRUE,
  downsampling.replace = TRUE
)

---
Arguments

- **seu**: Seurat S4 object with the column of ‘initial_cluster’ in its meta.data. Required.
- **group.by.var**: Initial clusters (batch-specific groups) variable. Needs to be one of the ‘colnames(seu@meta.data)’. Default: “initial_cluster”.
- **batch.by.var**: Batch variable. Needs to be one of the ‘colnames(seu@meta.data)’. Default: “Batch”.
- **verbose**: Boolean. Print the message and progress bar. (Default: TRUE)
- **use.parallel**: Boolean. Use parallel computation, which requires doParallel; no progress bar will be printed out. Run time will be 1/n.cores compared to the situation when no parallelisation is used. (Default: FALSE)
- **n.cores**: Numeric. Number of cores used for parallel computing (default: 1).
- **downsampling.size**: Numeric. The number of cells representing each group. (Default: 40)
- **downsampling.include**: Boolean. Using ‘include = TRUE’ to include the group smaller than required size. (Default: FALSE)
- **downsampling.replace**: Boolean. Using ‘replace = TRUE’ if the group is smaller than required size and some cells will be repeatedly used. (Default: FALSE)

Value

A list of four objects: a similarity matrix, a numeric vector recording cells used and the data frame of combinations included.

See Also

- `plotNetwork`
- `finalClustering`

Examples

```r
library(CIDER)
data(“pancreas”)ider <- getIDEr(pancreas, downsampling.size = 30)
head(ider)
```

Description

This function applies HDBSCAN, a density-based clustering method, on the corrected dimension reduction.
Usage

hdbscan.seurat(seu, reduction = "pca", dims = seq_len(15), minPts = 25)

Arguments

seu a Seurat object containing integrated or batch corrected PCA.
reduction Character. Name of the dimension reduction after integration or batch correction. (Default: PCA)
dims Numeric vector. Dimensions used for initial clustering. (Default: 1:15)
minPts Integer. Minimum size of clusters. Will be passed to the ‘hdbscan‘ function. (Default: 25)

Value

A Seurat object having two additional columns in its meta.data: dbscan_cluster and initial_cluster.

See Also

Usage of this function should be followed by getIDEr and estimateProb.

initialClustering Initial clustering

Description

Perform batch-specific initial clustering.

Usage

initialClustering(
  seu,
  batch.var = "Batch",
  cut.height = 0.4,
  nfeatures = 2000,
  additional.vars.to.regress = NULL,
  dims = seq_len(14),
  resolution = 0.6,
  downsampling.size = 50,
  verbose = FALSE
)
**measureSimilarity**

**Arguments**

- **seu**: Seurat S4 object. Required.
- **batch.var**: Character. One of the column names of ‘seu@meta.data’. It is used to partition the Seurat object into smaller ones. Default: "Batch"
- **cut.height**: Numeric. Height used to cut hierarchical trees. Default: 0.4
- **nfeatures**: Number of high variance genes used. Default: 2000
- **additional.vars.to.regress**: Additional variables to regress out. Needs to among column names of ‘seu@meta.data’. Default: ‘NULL’
- **dims**: Number of dimension used for clustering. Passed to Seurat. Default: ‘1:14’
- **resolution**: Resolution for clustering. Passed to Seurat. Default: 0.6
- **downsampling.size**: Numeric. The number of cells representing each group. (Default: 40)
- **verbose**: Print the progress bar or not. Default: FALSE

**Value**

Seurat S4 object with initial cluster information in ‘initial_cluster’ of meta.data.

**See Also**

- `getIDEr`
- `finalClustering`

---

**measureSimilarity**

*Measure similarity between two vectors*

**Description**

Measure similarity between two vectors

**Usage**

`measureSimilarity(x1, x2, method = "pearson")`

**Arguments**

- **x1**: x1
- **x2**: x2
- **method**: method

**Value**

Similarity matrix
mergeInitialClusters  Merge Initial Clusters

Description

Merge Initial Clusters

Usage

mergeInitialClusters(
  seu_list,  
  dist_list,  
  use = "coef",  
  method = "hc",  
  hc.method = "average",  
  cutree.by = "h",  
  cutree.h = 0.6,  
  cutree.k = 3
)

Arguments

  seu_list  
  A list containing Seurat objects. Required.
  dist_list  
  A list containing similarity matrices. The output of 'getDistMat ()'
  use  
  Default: "coef". No other option available currently.
  method  
  method = "hc"
  hc.method  
  Passed to the 'method' parameter of 'hclust()'. Default: "average"
  cutree.by  
  Cut trees by height ("h", default) or number of clusters ("k")
  cutree.h  
  Height used to cut the tree. Default: 0.6.
  cutree.k  
  Number of clusters used to cut the tree. Default: 3.

Value

  a list of Seurat objects containing the updated initial clustering information in 'seu_list[[seu_itor]]$inicluster'. The original initial cluster information is stored in 'seu_list[[seu_itor]]$inicluster_tmp'.

See Also

  hclust cutree gatherInitialClusters initialClustering
pancreas  

Pancreatic scRNA-Seq data.

Description
Toy data to test functions. It contains 12474 genes and only 222 cells. The count matrix and sample information were downloaded from NCBI GEO accession GSE84133.

Usage
pancreas

Format
A Seurat object.

Source

Examples
data("pancreas")

plotDistMat  

Plot Similarity Matrix with pheatmap

Description
Plot Similarity Matrix with pheatmap

Usage
plotDistMat(dist.list, use = "coef")

Arguments
dist.list Output of function 'getDistMat()'. Required.
use Default: "coef". No other option currently that can be used.

Value
A pheatmap showing the similarity matrix

See Also
getDistMat
plotHeatmap

Plot Heatmap for the IDER-based similarity matrix

Description

Plot Heatmap for the IDER-based similarity matrix

Usage

plotHeatmap(seu, ider)

Arguments

seu          An Seurat object.
ider         Output of function ‘getIDEr‘.

Value

A heatmap shows the similarity between shared groups in two batches

See Also

getIDEr

plotNetwork

Plot Network Graph

Description

Network visualisation for an IDER-based similarity matrix. The vertexes are initial clusters, and the edge width denotes the similarity between two initial clusters.

Usage

plotNetwork(
  seu,
  ider,
  colour.by = NULL,
  weight.factor = 6.5,
  col.vector = NULL,
  vertex.size = 1
)
Arguments

seu Seurat S4 object after the step of 'getIDER', containing 'initial_cluster' and 'Batch' in its meta.data. Required.
ider A list. Output of 'getIDER'. Required.
colour.by Character. It should be one of the colnames of Seurat object meta.data. It is used to colour the vertex of the network graph. (Default: NULL)
weight.factor Numerical. Adjust the thickness of the edges. (Default: 6.5)
col.vector A vector of Hex colour codes. If no value is given (default), a vector of 74 colours will be used.
vertex.size Numerical. Adjust the size of vertexes. (Default: 1)

Value

An igraph object

See Also

getIDEr graph_from_data_frame

Description

Scatterplot of a Seurat object based on dimension reduction.

Usage

scatterPlot(
  seu,
  reduction,
  colour.by,
  colvec = NULL,
  title = NULL,
  sort.by.numbers = TRUE,
  viridis_option = "B"
)

Arguments

seu Seurat S4 object after the step of 'getIDER'. Required.
reduction Character. The dimension reduction used to plot. Common options: "pca", "tsne", "umap". The availability of dimension reduction can be checked by 'Reductions(seu)'.

Scatterplot by a selected feature
scaterPlot

colour.by character. One of the column names of 'seu@meta.data'. Can be either discreet or continuous variables.
colvec A vector of Hex colour codes. If no value is given (default), a vector of 74 colours will be used.
title character. Title of the figure.
sort.by.numbers boolean. Whether to sort the groups by the number of cells. (default: True)
viridis_option viridis_option. (default: B)

Value

A scatter plot
Index

* datasets
  pancreas, 13

calculateDistMatOneModel, 2, 7
cosineSimilarityR, 3
cutree, 12
downsampling, 3
estimateProb, 4, 10
finalClustering, 5, 9, 11
gatherInitialClusters, 6, 12
getDistMat, 3, 6, 13
getGroupFit, 8
getIDEr, 4, 5, 8, 10, 11, 14, 15
graph_from_data_frame, 15

hclust, 12
hdbscan.seurat, 4, 9

initialClustering, 10, 12

measureSimilarity, 11
mergeInitialClusters, 6, 12

pancreas, 13
plotDistMat, 13
plotHeatmap, 14
plotNetwork, 9, 14

scatterPlot, 15