Package 'SCIntRuler'

July 12, 2024

Type Package

Title Guiding the Integration of Multiple Single-Cell RNA-Seq Datasets

Version 0.99.6

Maintainer Yue Lyu <yuelyu0521@gmail.com>

Description The accumulation of single-cell RNA-seq ('scRNA-seq') studies highlights the potential benefits of integrating multiple datasets. By augmenting sample sizes and enhancing analytical robustness, integration can lead to more insightful biological conclusions. However, challenges arise due to the inherent diversity and batch discrepancies within and across studies. 'SCIntRuler', a novel R package, addresses these challenges by guiding the integration of multiple 'scRNA-seq' datasets.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.3.0

Imports Rcpp, Matrix, batchelor, base, Seurat, SeuratObject, MatrixGenerics, SingleCellExperiment, SummarizedExperiment, dplyr, coin, harmony, ggplot2, gridExtra, cowplot, magrittr, stats

LinkingTo Rcpp

URL https://github.com/yuelyu21/SCIntRuler, https://yuelyu21.github.io/SCIntRuler/

BugReports https://github.com/yuelyu21/SCIntRuler/issues

Suggests BiocStyle, knitr, rmarkdown, testthat (>= 3.0.0)

VignetteBuilder knitr

Depends R (>= 4.3.0)

LazyData true

LazyDataCompression xz

Config/testthat/edition 3

NeedsCompilation yes

Author Yue Lyu [aut, cre] (<https://orcid.org/0000-0002-8912-6624>)

Repository CRAN

Date/Publication 2024-07-12 15:20:08 UTC

Contents

`SCIntRuler-package` *SCIntRuler: Integration of Single-Cell RNA-seq Datasets*

Description

The SCIntRuler package addresses the challenges of integrating multiple single-cell RNA-seq (scRNA-seq) datasets. It provides tools to enhance analytical robustness by augmenting sample sizes and reducing batch discrepancies. Developed using the Seurat framework, SCIntRuler includes both existing and novel workflows for single-cell analysis.

Value

This is the main page for SCIntRuler package.

Why SCIntRuler? Integrating scRNA-seq datasets can be complex due to various factors such as batch effects and sample diversity. SCIntRuler provides a statistical metric to aid in crucial decisions regarding dataset integration, ensuring more robust and accurate analyses.

NA

Features

- Informed Decision Making: Helps researchers decide on the necessity of data integration and the most suitable method.
- Flexibility: Suitable for various scenarios, accommodating different levels of data heterogeneity.
- Robustness: Enhances analytical robustness in joint analyses of merged or integrated scRNA-seq datasets.
- User-Friendly: Streamlines decision-making processes, simplifying the complexities involved in scRNA-seq data integration.

Getting Started

Refer to the "Getting Started with SCIntRuler" article in the package vignettes for detailed user instructions.

Author(s)

Yue Lyu

`CalcuSCIR` *Calculate SCIntRuler*

Description

Calculate SCIntRuler

Usage

```
CalcuSCIR(fullcluster, seuratlist, testres, p = 0.1)
```

Arguments

`fullcluster`	A list of clusters that generated by the function GetCluster()
`seuratlist`	A list of Seurat objects, usually can be got by SplitObject().
`testres`	Result from function PermTest()
`p`	P-value that will be used as the cut-off, default value is 0.1

Value

SCIntRuler

Examples

```
data(sim_result)
data(sim_data_sce)
sim_data <- SCEtoSeurat(sim_data_sce)
seuratlist <- Seurat::SplitObject(sim_data, split.by = "Study")
CalcuSCIR(sim_result[[1]], seuratlist, sim_result[[4]])
```

crossdist *Cross-Distance Matrix Calculation*

Description

Computes the pairwise Euclidean distance between rows of two matrices.

Usage

```
crossdist(m1, m2)
```

Arguments

m1	Numeric matrix.
m2	Numeric matrix.

Value

Numeric matrix of distances.

Examples

```
mat1 <- matrix(1:4, ncol = 2)
mat2 <- matrix(5:8, ncol = 2)
dist_matrix <- crossdist(mat1, mat2)
```

FindCell *Find cells indicating shared biological features across conditions*

Description

Find cells indicating shared biological features across conditions

Usage

```
FindCell(seuratobj, seuratlist, fullcluster, distmat, firstn = 15)
```

Arguments

seuratobj	The Seurat object that all samples/subjects were merged together.
seuratlist	A list of Seurat objects, usually can be got by SplitObject().
fullcluster	A list of clusters that generated by the function GetCluster().
distmat	A list of distance vectors generated by the function FindNNDist().
firstn	The number of nearest cells were detected that you want to include in the permutation test. Default to be 15.

Value

A list of two vectors: one is for which cluster of which sample will be highlighted and the second one is which cells will be selected.

Examples

```
data(sim_data_sce)
data(sim_result)
sim_data <- SCEtoSeurat(sim_data_sce)
seuratlist <- Seurat::SplitObject(sim_data, split.by = "Study")
FindCell(sim_data, seuratlist, sim_result[[1]], sim_result[[3]], 15)
```

FindNNDist *Find the nearest neighbors*

Description

Find the nearest neighbors

Usage

```
FindNNDist(fullcluster, normCount, meaningn = 20)
```

Arguments

fullcluster	A list of clusters that generated by the function GetCluster().
normCount	A list of normalized gene count matrix generated by the function NormData().
meaningn	default to be 20

Value

A list of distance vectors

Examples

```
data(sim_result)
meaningn <- 20
FindNNDist(sim_result[[1]], sim_result[[2]], meaningn = meaningn)
```

FindNNDistC *Find the nearest neighbors*

Description

Find the nearest neighbors

Usage

```
FindNNDistC(fullcluster, normCount, meaningn = 20)
```

Arguments

fullcluster	A list of clusters that generated by the function GetCluster().
normCount	A list of normalized gene count matrix generated by the function NormData().
meaningn	default to be 20

Value

A list of distance vectors

Examples

```
data(sim_result)
meaningn <- 20
FindNNDistC(sim_result[[1]], sim_result[[2]], meaningn = meaningn)
```

GetCluster *Get broad and fine clusters*

Description

Get broad and fine clusters

Usage

```
GetCluster(seuratlist, n1 = 50, n2 = 200)
```

Arguments

seuratlist	A list of Seurat objects, usually can be got by SplitObject(). We also accept the SingleCellExperiment object input.
n1	If the number of cells was smaller than n1, then the cluster will remain unchanged called rare cluster. The default value of n1 is 50.
n2	If the count of cells within a broad cluster is more than n2, the cluster is subdivided randomly into three fine clusters. If the cell count falls within the range of n1 to n2, two fine clusters are generated randomly. Default value is 200.

Value

A list of data frames.

Examples

```
data(sim_data_sce)
sim_data <- SCEtoSeurat(sim_data_sce)
seuratlist <- Seurat::SplitObject(sim_data, split.by = "Study")
fullcluster <- GetCluster(seuratlist)
```

NormData *Normalized RNA data matrix*

Description

Normalized RNA data matrix

Usage

```
NormData(seuratlist)
```

Arguments

seuratlist	A list of Seurat objects, usually can be got by SplitObject().

Value

A list of matrix.

Examples

```
data(sim_data_sce)
sim_data <- SCEtoSeurat(sim_data_sce)
seuratlist <- Seurat::SplitObject(sim_data, split.by = "Study")
normCount <- NormData(seuratlist)
```

PermTest *Permutation Test*

Description

Permutation Test

Usage

```
PermTest(fullcluster, distmat, firstn)
```

Arguments

fullcluster	A list of clusters that generated by the function GetCluster()
distmat	A list of distance vectors generated by the function FindNNDist().
firstn	The number of nearest cells were detected that you want to include in the permutation test.

Value

A list of two lists, one is the relative within-between distance and another is p-value of permutation test. Default to be 15.

Examples

```
data(sim_result)
testres <- PermTest(sim_result[[1]], sim_result[[3]],15)
```

PlotSCIR *Plot SCIntRuler*

Description

Plot SCIntRuler

Usage

```
PlotSCIR(fullcluster, seuratlist, testres, legendtitle = NULL, title = NULL)
```

Arguments

fullcluster	A list of clusters that generated by the function GetCluster.
seuratlist	A list of Seurat objects, usually can be got by SplitObject().
testres	Result from function PermTest()
legendtitle	Title of legend, default to be NULL
title	Title of figure, default to be NULL

Value

A ggplot2 object

Examples

```
data(sim_data_sce)
data(sim_result)
sim_data <- SCEtoSeurat(sim_data_sce)
seuratlist <- Seurat::SplitObject(sim_data, split.by = "Study")
PlotSCIR(sim_result[[1]], seuratlist, sim_result[[4]])
```

SCEtoSeurat *Input and Split SingleCellExperiment Data*

Description

This function takes a SingleCellExperiment object and a variable by which to split it, converts it to a Seurat object, and then splits it according to the specified variable.

Usage

```
SCEtoSeurat(sce)
```

Arguments

sce	A SingleCellExperiment object.

Value

A Seurat objects.

Examples

```
data(sim_data_sce)
# seuratlist <- InputData(sim_data_sce,"Study")
seuratobj <- SCEtoSeurat(sim_data_sce)
```

`sim_data_sce` *My Example Dataset*

Description

An example PBMC data with SingleCellExperiment format, including 3000 cells and 800 genes.

Usage

```
sim_data_sce
```

Format

An example PBMC data with SingleCellExperiment format

int_elementMetadata A DataFrame with 3000 rows and 1 column, storing simulated gene information.

int_colData A DataFrame with 800 rows and 3 columns, representing metadata for each cell.

int_metadata A list containing two elements that provide additional global metadata about the experiment.

rowRanges A CompressedGRangesList object providing genomic range data associated with each row/gene.

colData A DataFrame with 800 rows and 8 columns, detailing cell-level metadata.

assays A SimpleAssay object with matrix dimensions 3000x800, representing the gene expression matrix.

elementMetadata A DataFrame linked with assays, providing gene-level metadata.

Details

The "sim_data_sce" object is designed to serve as a teaching and development aid for methods that require complex single-cell expression data. It includes several typical features found in single-cell datasets, such as varied levels of gene expression and metadata describing both cells and genes.

The data within this object are entirely synthetic and should not be used for real analysis. The main use case is for testing and development of single-cell analysis methodologies.

Value

Simulation data to exemplify the usage of the method.

References

The data were generated using a combination of random number generation for expression values and curated sources for metadata to simulate realistic experimental scenarios.

Examples

```
data("sim_data_sce")
```

sim_result *My Example Dataset*

Description

An result example data with results from different functions.

Usage

```
sim_result
```

Format

An result example data

fullcluster A runable example of GetCluster, which is a list of clusters for each study.

normCount A runable example of NormData, which is a list of normalized RNA expression matrixs for each study.

distmat A runable example of FindNNDist, which is a list of distance matrixs for each study.

testres A runable example of CalcuSCIR, which is a list of test results for each study.

Value

Simulation data to examplify the usage of the method.

Examples

```
# Load the data
data("sim_result")
```

SummCluster *Get maximum number of broad clusters*

Description

Get maximum number of broad clusters

Usage

```
SummCluster(fullcluster)
```

Arguments

fullcluster	A list of clusters that generated by the function GetCluster()

Value

A list

Examples

```
data(sim_result)
SCout <- SummCluster(sim_result[[1]])
```

Index