Package 'scROSHI'

January 10, 2023

Title Robust Supervised Hierarchical Identification of Single Cells

Version 1.0.0.0

Maintainer Lars Bosshard <bosshard@nexus.ethz.ch>

Description Identifying cell types based on expression profiles is a pillar of single cell analysis. 'scROSHI' identifies cell types based on expression profiles of single cell analysis by utilizing previously obtained cell type specific gene sets. It takes into account the hierarchical nature of cell type relationship and does not require training or annotated data. A detailed description of the method can be found at: Prummer, Bertolini, Bosshard, Barkmann, Yates, Boeva, The Tumor Profiler Consortium, Stekhoven, and Singer (2022) <doi:10.1101/2022.04.05.487176>.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

RoxygenNote 7.2.1

Imports limma, S4Vectors, SingleCellExperiment, stats, SummarizedExperiment, utils, uwot

Depends R (>= 3.60)

biocViews

LazyDataCompression xz

NeedsCompilation no

Author Lars Bosshard [aut, cre] (<https://orcid.org/0000-0002-4550-4777>), Michael Prummer [aut] (<https://orcid.org/0000-0001-9896-3929>)

Repository CRAN

Date/Publication 2023-01-10 13:50:02 UTC

R topics documented:

`config` *Test config file*

Description

Config file for the test_sce_data set

Usage

```
config
```

Format

A data frame with 2 columns and 7 rows

Details

Config file to define major cell types and hierarchical subtypes. It should be provided as a two-column data.frame where the first column are the major cell types and the second column are the subtypes. If several subtypes exists, they should be separated by comma.

Source

scROSHI - robust supervised hierarchical identification of single cells https://www.biorxiv.org/content/10.1101/2022.04.05.487176v1

`f_annot_ctgenes` *Choose best matching cell type*

Description

Choose best matching cell type

Usage

```
f_annot_ctgenes(m.cts, unknown, uncertain)
```

Arguments

m.cts	Matrix containing the cell type scores. The rows represent the cell types, whereas the columns represent the samples.
unknown	If none of the probabilities is above this threshold, the cell type label is assigned to the class unknown.
uncertain	If the ratio between the largest and the second largest probability is below this threshold, the cell type label is assigned to the class uncertain for the major cell types.

Value

Data frame containing the best matching cell type for each sample.

Examples

```
m.cts <- matrix(c(0.2, 0.001, 0.002, 0.1), nrow=2)
colnames(m.cts) <- c("sample1", "sample2")
rownames(m.cts) <- c("cell_type1", "cell_type2")
f_annot_ctgenes(m.cts, 0.05, 0.1)
```

---

f_my_correlation_test *Spearman correlation test*

Description

Calculate spearman correlation p-value or return preset values.

Usage

```
f_my_correlation_test(x, y, min_genes = 5)
```

Arguments

x	A numeric vector of values.
y	A numeric vector of values.
min_genes	A numeric value defining the threshold for the minimum number of genes.

Value

A numeric value representing the p value of the Spearman correlation test.

Examples

```
f_my_correlation_test(rnorm(10,1,2),rnorm(10,5,2))
```

f_my_wilcox_test *Wilcox test*

Description

Calculate Wilcox p-value or return preset values

Usage

```
f_my_wilcox_test(x, y, min_genes = 5)
```

Arguments

x	A numeric vector of values.
y	A numeric vector of values.
min_genes	A numeric value defining the threshold for the minimum number of genes.

Value

A numeric value representing the p value of the Wilcox test.

Examples

```
f_my_wilcox_test(rnorm(10, 1, 2), rnorm(10, 5, 2))
```

f_score_ctgenes_U *Calculate cell type score*

Description

Calculate cell type score

Usage

```
f_score_ctgenes_U(
  sce,
  gset,
  count_data = "normcounts",
  gene_symbol = "SYMBOL",
  min_genes = 5,
  verbose = 0
)
```

Arguments

sce	A SingleCellExperiment object containing the expression profiles of the single cell analysis.
gset	Marker gene list for all cell types.
count_data	Assay name in the SingleCellExperiment object containing the count data.
gene_symbol	Variable name in the row data of the SingleCellExperiment object containing the gene names.
min_genes	Minimum number of genes.
verbose	Level of verbosity. Zero means silent, one makes a verbose output.

Value

Matrix containing the cell type scores. The rows represent the cell types, whereas the columns represent the samples.

Examples

```
data("test_sce_data")
gset <- list(cell_type1 = c("CD79A", "TCL1A", "VPREB3"),
cell_type2 = c("FCER1A", "CLEC10A", "ENHO"))
f_score_ctgenes_U(test_sce_data[,1:3], gset, count_data = "normcounts",
gene_symbol = "SYMBOL", min_genes = 3, verbose = 0)
```

---

f_score_profile_cor *Calculate cell type score: U-test*

Description

Calculate cell type score: U-test

Usage

```
f_score_profile_cor(sce, lprof, min_genes = 5, verbose = 0)
```

Arguments

sce	A SingleCellExperiment object containing the expression profiles of the single cell analysis.
lprof	List of profiles (named expression vectors).
min_genes	Minimum number of genes.
verbose	Level of verbosity. Zero means silent, one makes a verbose output.

Value

Matrix containing the cell type scores.

Examples

```
data("test_sce_data")
set.seed(123)
prof1 <- rpois(n = 20, lambda = 3)
names(prof1) <- rownames(test_sce_data)[51:70]
prof2 <- rpois(n = 20, lambda = 5)
names(prof2) <-  rownames(test_sce_data)[71:90]
lprof <- list(prof1 = prof1, prof2 = prof2)
f_score_profile_cor(test_sce_data[,1:3], lprof, min_genes = 5, verbose = 0)
```

---

marker_list *Marker gene list for the test SCE data Set*

Description

Marker gene list for the test_sce_data set

Usage

```
marker_list
```

Format

Marker gene list for all cell types as a list of genes with cell types as names

Source

scROSHI - robust supervised hierarchical identification of single cells https://www.biorxiv.org/content/10.1101/2022.04.05.487176v1

---

scROSHI *Robust Supervised Hierarchical Identification of Single Cells*

Description

Identifying cell types based on expression profiles is a pillar of single cell analysis.

scROSHI identifies cell types based on expression profiles of single cell analysis by utilizing previously obtained cell type specific gene sets. It takes into account the hierarchical nature of cell type relationship and does not require training or annotated data.

Usage

```
scROSHI(
  sce_data,
  celltype_lists,
  type_config,
  count_data = "normcounts",
  gene_symbol = "SYMBOL",
  cell_scores = FALSE,
  min_genes = 5,
  min_var = 1.5,
  n_top_genes = 2000,
  n_nn = 5,
  thresh_unknown = 0.05,
  thresh_uncert = 0.1,
  thresh_uncert_second = 0.8,
  verbose = 0,
  output = "sce"
)
```

Arguments

sce_data	A SingleCellExperiment object containing the expression profiles of the single cell analysis.
celltype_lists	Marker gene list for all cell types. It can be provided as a list of genes with cell types as names or as a path to a file containing the marker genes. Supported file formats are .gmt or .gmx files.
type_config	Config file to define major cell types and hierarchical subtypes. It should be provided as a two-column data.frame where the first column are the major cell types and the second column are the subtypes. If several subtypes exists, they should be separated by comma.
count_data	Assay name in the SingleCellExperiment object containing the count data.
gene_symbol	Variable name in the row data of the SingleCellExperiment object containing the gene names.
cell_scores	Boolean value determining if the scores should be saved.
min_genes	scROSHI filters out non-unique genes as long as more than min_genes are left. If there is a cell type that has less than min_genes genes, it will be replaced with the cell type list BEFORE filtering for unique genes (default 5).
min_var	Minimum variance for highly variable genes (default 1.5).
n_top_genes	Maximum number of highly variable genes (default 2000).
n_nn	Number of nearest neighbors for umap for assignment of cell types (default 5).
thresh_unknown	If none of the probabilities is above this threshold, the cell type label is assigned to the class unknown (default 0.05).
thresh_uncert	If the ratio between the largest and the second largest probability is below this threshold, the cell type label is assigned to the class uncertain for the major cell types (default 0.1).

thresh_uncert_second
: If the ratio between the largest and the second largest probability is below this threshold, the cell type label is assigned to the class uncertain for the subtypes (default 0.8).

verbose
: Level of verbosity. Zero means silent, one makes a verbose output.

output
: Defines the output. sce: The output is a SingleCellExperiment object with the cell types appended to the meta data. df: The output id a data.frame with two columns. The first column contains the barcode of the cell and the second column contains the cell type labels.

Examples

```
data("test_sce_data")
data("config")
data("marker_list")

results <- scROSHI(sce_data = test_sce_data,
                   celltype_lists = marker_list,
                   type_config = config)
table(results$celltype_final)
```

---

test_sce_data *Test SCE Data Set*

Description

Data from a peripheral blood mononuclear cell experiments from an adult human.

Usage

```
test_sce_data
```

Format

A SingleCellExperiment object with 3368 genes and 1316 cells

Source

scROSHI - robust supervised hierarchical identification of single cells https://www.biorxiv.org/content/10.1101/2022.04.05.487176v1

Index