Package ‘EasyCellType’

May 3, 2024

Title Annotate cell types for scRNA-seq data

Version 1.6.0

Description We developed EasyCellType which can automatically examine the input marker lists obtained from existing software such as Seurat over the cell marker databases. Two quantification approaches to annotate cell types are provided: Gene set enrichment analysis (GSEA) and a modified version of Fisher’s exact test. The function presents annotation recommendations in graphical outcomes: bar plots for each cluster showing candidate cell types, as well as a dot plot summarizing the top 5 significant annotations for each cluster.

License Artistic-2.0

RoxygenNote 7.3.1

Encoding UTF-8

Depends R (>= 4.2.0)

biocViews SingleCell, Software, GeneExpression, GeneSetEnrichment

Imports clusterProfiler, dplyr,forcats, ggplot2, magrittr, rlang,

stats, org.Hs.eg.db, org.Mm.eg.db, AnnotationDbi, vctrs (>= 0.6.4), BiocStyle

Suggests knitr, rmarkdown, testthat (>= 3.0.0), Seurat, BiocManager,

devtools, BiocStyle

VignetteBuilder knitr

Config/testthat/edition 3

git_url https://git.bioconductor.org/packages/EasyCellType

git_branch RELEASE_3_19

git_last_commit 4abd5fc

git_last_commit_date 2024-04-30

Repository Bioconductor 3.19

Date/Publication 2024-05-03

Author Ruoxing Li [aut, cre, ctb],

Ziyi Li [ctb]

Maintainer Ruoxing Li <ruoxingli@outlook.com>
Contents

cellmarker_tissue .......................................................... 2
clustermole_tissue .......................................................... 2
coremarkers ................................................................. 3
easyct ............................................................................ 3
gene_pbmc ............................................................... 5
mapsymbol ................................................................. 5
panglao_tissue ............................................................... 6
pbmc_data ................................................................. 6
plot_bar ................................................................. 7
plot_dot ................................................................. 7
process_results ............................................................. 8
summarycelltype ............................................................ 8
test_fisher ................................................................. 9

Index 10

---

cellmarker_tissue  Tissues in CellMarker database.

Description

A list containing 2 elements: Human tissues and Mouse tissues.

Usage

data(cellmarker_tissue)

Format

A list with 2 elements:

**Human**  Human tissue

**Mouse**  Mouse tissue

---

clustermole_tissue  Tissues in ClusterMole database.

Description

A list containing 2 elements: Human tissues and Mouse tissues.

Usage

data(clustermole_tissue)
coremarkers

Format
A list with 2 elements:

- **Human**: Human tissue
- **Mouse**: Mouse tissue

---

<table>
<thead>
<tr>
<th>coremarkers</th>
<th>Title Summarize markers contributing to the cell type annotation</th>
</tr>
</thead>
</table>

Description
Title Summarize markers contributing to the cell type annotation

Usage
coremarkers(test, data, species)

Arguments
- **test**: Test used to annotation cell types: "GSEA" or "fisher"
- **data**: Annotation results.
- **species**: "Human" or "Mouse"

Value
A data frame containing genes contributed to cell annotation

Examples
```r
## core_markers <- coremarkers("GSEA", data)
```

---

easyct  

Annotate cell types for scRNA-seq data

Description
This function is used to run the annotation analysis using either GSEA or a modified Fisher’s exact test. We expect users to input a data frame containing expressed markers, cluster information and the differential score (log fold change). The gene lists in that data frame should be sorted by their differential score.
Usage

easyct(
    data,
    db = "cellmarker",
    genetype = "Entrezid",
    species = "Human",
    tissue = NULL,
    p_cut = 0.5,
    test = "GSEA",
    scoretype = "std"
)

Arguments

data A data frame containing the markers, cluster, and expression scores; Marker
genes should be sorted in each cluster. Order of the columns should be gene,
cluster and expression level score. An example data can be loaded using 'data(gene_pbmc)'.

db Name of the reference database: cellmarker, clustermole or panglaodb;
genetype Indicate the gene type in the input data frame: "Entrezid" or "symbol".
species Human or Mouse. Human in default.
tissue Tissue types can be specified when running the analysis. Length of tissue can be
larger than 1. The possible tissues can be seen using 'data(cellmarker_tissue)',
'data(clustermole_tissue)' and 'data(panglao_tissue)'.
p_cut Cutoff of the P value for GSEA.
test "GSEA" or "fisher"; "GSEA" is used in default.
scoretype Argument used for GSEA. Default value is "std". If all scores are positive, then
scoretype should be "pos".

Value

A list containing the test results for each cluster.

Examples

data(gene_pbmc)
result <- easyct(gene_pbmc, db="cellmarker", species="Human",
tissue=c("Blood", "Peripheral blood", "Blood vessel",
"Umbilical cord blood", "Venous blood"), p_cut=0.3, test="GSEA", scoretype="pos")
**gene_pbmc**

**Differential expressed marker genes in 9 clusters.**

**Description**

A data frame containing marker genes, clusters as well as the average of log 2 fold changes. The original data set is from 10X genomics, and we followed the standard workflow provided by Seurat package to process data, and then format to get the data frame.

**Usage**

```r
data(gene_pbmc)
```

**Format**

A data frame with 727 rows and 3 variables:

- **gene**: Entrez IDs of the marker genes
- **cluster**: Cluster
- **score**: Average of log 2 fold changes getting from the process procedure

**Source**

https://cf.10xgenomics.com/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz

---

**mapsymbol**

**Title Convert gene symbol to Entrez ID**

**Description**

This function is used to convert the gene symbol to Entrez Id. Used in easyct function.

**Usage**

```r
mapsymbol(d, species)
```

**Arguments**

- **d**: A data frame where first column contains gene symbols.
- **species**: "Human" or "Mouse".

**Value**

A data frame containing gene symbols and the corresponding Entrez ID
panglao_tissue  
*Tissues in Panglao database.*

**Description**

A list containing 2 elements: Human tissues and Mouse tissues.

**Usage**

```r
data(panglao_tissue)
```

**Format**

A list with 2 elements:

- **Human**  Human tissue
- **Mouse**  Mouse tissue

---

pbmc_data  
*Peripheral Blood Mononuclear Cells (PBMC) data.*

**Description**

Count matrix of Peripheral Blood Mononuclear Cells (PBMC). The original data set is from 10X genomics.

**Usage**

```r
data(pbmc_data)
```

**Format**

A large dgCMatrix: 32378 * 2700

- **i**  Row index of the non-zero values
- **p**  A vector to refer the column index of the non-zero values
- **Dim**  Dimension of the matrix
- **Dimnames**  A list of length 2 containing the row names and column names of the matrix
- **x**  Vector containing all the non-zero values

**Source**

[https://cf.10xgenomics.com/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz](https://cf.10xgenomics.com/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz)
plot_bar

Create bar plots for each cluster

Description
This function is used to generate set of bar plots presenting up to 10 candidate cell types for each cluster.

Usage
plot_bar(test = "GSEA", data, cluster = NULL)

Arguments
- test: "GSEA" or "fisher"
- data: Annotation results
- cluster: Cluster can be specified to print plots.

Value
Bar plots showing show up to 10 candidate cell types for each cluster.

Examples
```r
data(gene_pbmc)
result <- easyct(gene_pbmc, db="cellmarker", species="Human",
tissue=c("Blood", "Peripheral blood", "Blood vessel",
"Umbilical cord blood", "Venous blood"), p_cut=0.3, test="GSEA", scoretype="pos")
plot_bar("GSEA", result)
```

plot_dot

Create dot plot for annotation results

Description
This function is used to generate a dot plot presenting the top 5 candidate cell types for each cluster.

Usage
plot_dot(test = "GSEA", data)

Arguments
- test: Test used to annotate cell types: "GSEA" or "fisher"
- data: Annotation results
Value
A dot plot showing the top 5 significant cell types for each cluster.

Examples
```r
data(gene_pbmc)
result <- easyct(gene_pbmc, db="cellmarker", species="Human",
tissue=c("Blood", "Peripheral blood", "Blood vessel",
"Umbilical cord blood", "Venous blood"), p_cut=0.3, test="GSEA", scoretype="pos")
plot_dot("GSEA", result)
```

| process_results          | Title Annotate cell types for single cell RNA data |

Description
This function is used to process the annotation test results. Processed data will be used to generate plots.

Usage
```r
process_results(test, data)
```

Arguments
- **test**: Test used to annotation cell types: "GSEA" or "fisher"
- **data**: Annotation results.

Value
A data frame used to generate plots.

| summarycelltype           | Print test results |

Description
This function is used to print summary table of annotation results for a specific cluster.

Usage
```r
summarycelltype(test, results, cluster)
```
Arguments

test  "GSEA" or "fisher".
results Annotation results.
cluster Cluster of interest.

Value
A summary table of a annotation results. "core_enrichment" contains markers contributing on the annotation.

Examples

data(gene_pbmc)
result <- easyct(gene_pbmc, db="cellmarker", species="Human",
tissue=c("Blood", "Peripheral blood", "Blood vessel",
"Umbilical cord blood", "Venous blood"), p_cut=0.3, test="GSEA", scoretype="pos")
summarycelltype(test="GSEA", results=result, cluster=0)

test_fisher  Fisher exact test used in function 'easyct'

Description
This function is used to conduct the modified Fisher’s exact test.

Usage
test_fisher(testgenes, ref, cols)

Arguments

testgenes A data frame containing query genes and the expression scores.
ref The reference data base.
cols Column names of the input data frame

Value
A data frame containing the results of fisher’s exact test.
Index

* datasets
  - cellmarker_tissue, 2
  - clustermole_tissue, 2
  - gene_pbmc, 5
  - panglao_tissue, 6
  - pbmc_data, 6

  cellmarker_tissue, 2
  clustermole_tissue, 2
  coremarkers, 3

  easyct, 3
  gene_pbmc, 5

  mapsymbol, 5

  panglao_tissue, 6
  pbmc_data, 6
  plot_bar, 7
  plot_dot, 7
  process_results, 8
  summarycelltype, 8
  test_fisher, 9