Package ‘scAnnotate’

July 1, 2022

Type Package

Title An Automated Cell Type Annotation Tool for Single-Cell RNA-Sequencing Data

Version 0.0.1

Description An entirely data-driven cell type annotation tools, which requires training data to learn the classifier, but not biological knowledge to make subjective decisions. It consists of three steps: preprocessing training and test data, model fitting on training data, and cell classification on test data. See Xiangling Ji, Danielle Tsao, Kailun Bai, Min Tsao, Xuekui Zhang (2022) <doi:10.1101/2022.02.19.481159> for more details.

Depends R(>= 4.0.0)

License GPL-3

URL https://doi.org/10.1101/2022.02.19.481159

Encoding UTF-8

LazyData true

RoxygenNote 7.1.2

Suggests knitr, testthat (>= 3.0.0), rmarkdown

VignetteBuilder knitr

Imports glmnet, stats, MTPS, Seurat (>= 4.0.5), harmony

Config/testthat/edition 3

NeedsCompilation no

Author Xiangling Ji [aut], Danielle Tsao [aut], Kailun Bai [ctb], Min Tsao [aut], Xuekui Zhang [aut, cre]

Maintainer Xuekui Zhang <xuekui@uvic.ca>

Repository CRAN

Date/Publication 2022-07-01 11:00:02 UTC
Description

calculate the F1 score of each cell population, mean of F1 score and overall accuracy

Usage

eva_cal(prediction, cell_label)

Arguments

  prediction    A vector of annotate cell type labels
  cell_label    A vector of original cell type labels

Value

  A matrix contain the F1 score of each cell population, mean of F1 score and overall accuracy

Examples

data(predict_label)
data(pbmc2)
eva_cal(prediction = predict_label, cell_label = pbmc2[,1])
**Description**

A subset of human Peripheral Blood Mononuclear Cells (PBMC) scRNA-seq data that was sequenced using Drop-seq platform. The Seurat(version 4.0.5) package was used for normalized using the NormalizeData function with the "LogNormalize" method and a scale factor of 10,000. After modeling the mean-variance relationship with the FindVariableFeautre function within "vst" methods, we selected the top 2,000 highly variable genes and only used this selection going forward. The dataframe of the cell type label and a gene expression matrix of 598 cells in the row and 2,000 genes in the column.

**Usage**

data(pbmc1, package="scAnnotate")

**Format**
a data frame

**References**


**Description**

A subset of human PBMC scRNA-seq data that was sequenced using inDrops platform. The Seurat(version 4.0.5) package was used for normalized using the NormalizeData function with the "LogNormalize" method and a scale factor of 10,000. After modeling the mean-variance relationship with the FindVariableFeautre function within "vst" methods, we selected the top 2,000 highly variable genes and only used this selection going forward. The dataframe of the cell type label and a gene expression matrix of 644 cells in the row and 2,000 genes in the column.

**Usage**

data(pbmc2, package="scAnnotate")

**Format**
a data frame

**References**

**predict_label**

**predict_label**

**Description**

Cell type annotation of pbmc2 data that training from pbmc1 data by 'scAnnotate'.

**Usage**

```r
data(predict_label, package="scAnnotate")
```

**Format**

a data frame

---

**scAnnotate**

**scAnnotate**

**Description**

Annotate cell type labels of test data using a trained mixture model from training data

**Usage**

```r
scAnnotate(
    train,
    test,
    distribution = c("normal", "dep"),
    correction = c("auto", "harmony", "seurat"),
    screening = c("wilcox", "t.test"),
    threshold = 0
)
```

**Arguments**

- **train**: A data frame of cell type label in the first column and a gene expression matrix where each row is a cell and each column is a gene from training data
- **test**: A data matrix where each row is a cell and each column is a gene from test data
- **distribution**: A character string indicate the distribution assumption on positive gene expression, should be one of "normal"(default), or "dep"
- **correction**: A character string indicate the batch effect removal, should be one of "auto"(default), "seurat", or "harmony". "auto" will automatically select the batch effect removal follow our suggestion using Seurat for big data (all cell population sample size greater than 100 except one cell population less than 100 and greater than 20) and using harmony for small data.
scAnnotate

screening  A character string indicate the gene screening methods, should be one of "wilcox" (default) or "t.test".

threshold  A numeric number indicate the threshold used for probabilities to classify cells into classes, should be number from "0" (default) to "1". If there's no probability higher than the threshold associated to a cell type, the cell will be labeled as "unassigned"

Value

A vector contain annotate cell type labels for test data

Examples

data(pbmc1)
data(pbmc2)
predict_label=scAnnotate(train=pbmc1,
test=pbmc2[,-1],
distribution="normal",
correction="harmony",
screening ="wilcox",
threshold=0)
Index

* datasets
  pbmc1, 3
  pbmc2, 3
  predict_label, 4

eva_cal, 2
pbmc1, 3
pbmc2, 3
predict_label, 4
scAnnotate, 4