Package ‘OncoSubtype’

January 20, 2025

Type Package

Title Predict Cancer Subtypes Based on TCGA Data using Machine Learning Method

Version 1.0.0

Author Dadong Zhang <`dadong.zhang.shared@gmail.com`>

Maintainer Dadong Zhang <`dadong.zhang.shared@gmail.com`>

Description Provide functionality for cancer subtyping using nearest centroids or machine learning methods based on TCGA data.

License GPL-3

Encoding UTF-8

LazyData true

RoxygenNote 7.3.1

URL `https://github.com/DadongZ/OncoSubtype`

BugReports `https://github.com/DadongZ/OncoSubtype/issues`

Suggests knitr, rmarkdown, testthat (>= 3.0.0)

VignetteBuilder knitr

LazyDataCompression xz

Imports caret, randomForest, methods, e1071, pheatmap, tibble, dplyr, limma, rlang, Rdpack

RdMacros Rdpack

Depends SummarizedExperiment, R (>= 3.63),

Config/testthat/edition 3

NeedsCompilation no

Repository CRAN

Date/Publication 2024-03-22 17:30:10 UTC

Contents

`centroids_subtype` *Predict the subtypes of selected cancer type based published papers*

Description

Predict the subtypes of selected cancer type based published papers

Usage

```
centroids_subtype(data, disease = "LUSC")
```

Arguments

`data`	data set to predict the subtypes which is a numeric matrix with row names of features and column names of samples
`disease`	character string of the disease to predict subtypes, currently support ’LUSC’, ’LUAD’

Value

an object of class "SubtypeClass" with four slots: genes used for predictiong, predicted subtypes of samples, a matrix of predicting scores, and the method.

Examples

```
## Not run:
library(OncoSubtype)
data <- get_median_centered(example_fpkm)
data <- assays(data)$centered
rownames(data) <- rowData(example_fpkm)$external_gene_name
centroids_subtype(data, disease = 'HNSC')

## End(Not run)
```

`example_fpkm` *example FPKM data*

Description

example FPKM data

Usage

```
example_fpkm
```

Format

SummarizedExperiment object

`get_hvg` *select highly variable genes from a expression matrix*

Description

select highly variable genes from a expression matrix

Usage

```
get_hvg(data, top = 1000)
```

Arguments

`data`	a (normalized) matrix with rownames of features and colnames of samples
`top`	number of top highly variable genes to output

Value

subset with top ranked genes by the variances

Examples

```
## Not run:
library(OncoSubtype)
data <- get_median_centered(example_fpkm)
data <- assays(data)$centered
get_hvg(data)

## End(Not run)
```

`get_median_centered` *convert expression matrix to median-centered*

Description

convert expression matrix to median-centered

Usage

```
get_median_centered(data, log2 = TRUE)
```

Arguments

`data`	a numeric matrix or ‘S4‘ object
`log2`	logical, if ‘TRUE‘ $log2(x+1)$ will be applied before median centering. Defaut ‘TRUE‘.

Value

median-centered express matrix or an object with new slot "centered"

Examples

```
## Not run:
get_median_centered(example_fpkm)

## End(Not run)
```

`get_rf_pred` *Predict the subtypes of selected cancer type*

Description

Predict the subtypes of selected cancer type

Usage

```
get_rf_pred(train_set, test_set, method = "rf", seed = NULL)
```

Arguments

`train_set`	training set with rownames of samples, first column named ’mRNA_subtype’ and the rest of features and expression values.
`test_set`	test set with rownames of features and colnames of samples.
`method`	character string of the method to use currently support ’rf’.
`seed`	integer seed to use.

Value

a matrix with column names of subtypes and predicted probabilities.

`hnsc_centroids` *HNSC predictor centroids*

Description

HNSC predictor centroids from `https://www.nature.com/articles/nature14129`

Usage

```
hnsc_centroids
```

Format

A `tibble` with 728 features and four subtypes.

`load_dataset_from_github` *Load Dataset from GitHub Repository*

Description

Downloads a specified dataset from a GitHub repository if it is not already present in the specified local directory, then loads the dataset into the global environment. This function is designed to help manage package size by storing data externally and loading it on-demand.

Usage

```
load_dataset_from_github(disease, local_dir = path.expand(getwd()))
```

Arguments

`disease`	A character string specifying the disease, which corresponds to the name of the dataset to be loaded (e.g., "LUSC"). The function constructs the filename as `tolower(disease)_tcga.rda` and attempts to load this dataset.
`local_dir`	An optional character string specifying the path to the directory where datasets should be stored locally. If not provided, defaults to a subdirectory named `your_package_name_data` within the user's home directory. Users can specify their own directory path if they prefer to store data in a different location.

Value

Invisible NULL. The function is primarily used for its side effect of loading a dataset into the global environment. However, the function itself does not return the dataset directly.

Examples

```
## Not run:
  load_dataset_from_github("LUSC")

## End(Not run)
```

`luad_centroids` *LUAD predictor centroids*

Description

LUAD predictor centroids from Wilkerson (2012)

Usage

```
luad_centroids
```

Format

A `tibble` with 506 features and three subtypes bronchioid, magnoid, and squamoid.

`lusc_centroids` *LUSC predictor centroids*

Description

LUSC predictor centroids from Wilkerson (2010)

Usage

```
lusc_centroids
```

Format

A `tibble` with 208 features and four subtypes: primitive, classical, secretory, and basal.

`ml_subtype` *Predict the subtypes of selected cancer type using machine learning*

Description

Predict the subtypes of selected cancer type using machine learning

Usage

```
ml_subtype(
  data,
  disease = "LUSC",
  method = "rf",
  removeBatch = TRUE,
  seed = NULL
)
```

Arguments

`data`	data set to predict the subtypes which is a numeric matrix with row names of features and column names of samples
`disease`	character string of the disease to predict subtypes, currently support 'LUSC', 'LUAD', and 'BLCA'.
`method`	character string of the method to use currently support 'rf'.
`removeBatch`	whether do batch effect correction using `limma::removeBatchEffect`, default TRUE.
`seed`	integer seed to use.

Value

An object of class "SubtypeClass" with four slots: genes used for predictiong, predicted subtypes of samples, a matrix of predicting scores, and the method.

References

1. Wilkerson MJ, Yin X, Hayes D, et al. (2010). "Lung Squamous Cell Carcinoma mRNA Expression Subtypes Are Reproducible, Clinically Important, and Correspond to Normal Cell Types." *Clin Cancer Res*, **16**(19), 4864-4875.
2. Wilkerson MJ, Yin X, Hayes D, et al. (2012). "Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability, and methylation." *Plos One*, **7**(5), e36530.
3. Network TCGA (2015). "Comprehensive genomic characterization of head and neck squamous cell carcinomas." *Nature*, **517**, e36530.

Examples

```
## Not run:
library(OncoSubtype)
data <- get_median_centered(example_fpkm)
data <- assays(data)$centered
rownames(data) <- rowData(example_fpkm)$external_gene_name
ml_subtype(data, disease = 'LUAD', method = 'rf', seed = 123)

## End(Not run)
```

`PlotHeat` *Plot heatmap of the train set or test set*

Description

Plot heatmap of the train set or test set

Usage

```
PlotHeat(object, set = "test", ...)
```

Arguments

`object`	a SubtypeClass object
`set`	options could be ’test’, ’train’ or ’both’. Default ’test’.
`...`	Parameters passed to `pheatmap`.

Value

a pheatmap object

Examples

```
## Not run:
library(OncoSubtype)
data <- get_median_centered(example_fpkm)
data <- assays(data)$centered
rownames(data) <- rowData(example_fpkm)$external_gene_name
object <- MLSubtype(data, disease = 'LUSC')
PlotHeat(object, set = 'both', fontsize = 10, show_rownames = FALSE, show_colnames = FALSE)

## End(Not run)
```

`SubtypeClass-class` *Set the SubtypeClass*

Description

Set the SubtypeClass

Value

an object of SubtypeClass with three empty solts

Index