Package 'ssMutPA'

October 16, 2024

Type Package

Title Single-Sample Mutation-Based Pathway Analysis

Version 0.1.2

Description A systematic bioinformatics tool to perform single-sample mutation-based pathway analysis by integrating somatic mutation data with the Protein-Protein Interaction (PPI) network. In this method, we use local and global weighted strategies to evaluate the effects of network genes from mutations according to the network topology and then calculate the mutation-based pathway enrichment score (ssMutPES) to reflect the accumulated effect of mutations of each pathway. Subsequently, the ssMutPES profiles are used for unsupervised spectral clustering to identify cancer subtypes.

License GPL (>= 2)

Encoding UTF-8

LazyData true

RoxygenNote 7.3.1

Imports ggplot2, ggridges, grDevices, igraph, kernlab, maftools, Matrix, NbClust, parallel, pheatmap, RColorBrewer, stats, survival, utils

Suggests knitr, qpdf, rmarkdown

Depends R (>= 4.0.0)

VignetteBuilder knitr

NeedsCompilation no

Author Junwei Han [aut, cre, cph], Yalan He [aut], Qian Wang [aut]

Maintainer Junwei Han <hanjunwei1981@163.com>

Repository CRAN

Date/Publication 2024-10-15 22:00:02 UTC

Contents

`cox_data` *cox_data Univariate cox proportional hazards regression data.*

Description

A data frame for implementing univariate cox proportional hazards regression.

Usage

```
cox_data
```

Format

An object of class `data.frame` with 20 rows and 293 columns.

dotplot *Plotting the Dot plot.*

Description

The function is used to draw a graph to reflect the univariate HRs and P-values of the pathways in different cancer types.

Usage

```
dotplot(data, low_col = "#6ADD26", high_col = "#AB2513", cut_point = 5)
```

Arguments

data	A pathway activity score matrix, which rows represent the pathways and the columns are samples.
low_col, high_col	Colours for low and high ends of the gradient.
cut_point	The threshold of HRs,when HR is greater than the cut_point,HR is assigned cut_point.

Value

A Dot plot

Examples

```
#load the data
data(dot_data)
#perform the function `dotplot`.
dotplot(dot_data)
```

dot_data *dot_data A data frame.*

Description

The data frame was used to plot a dot plot.

Usage

```
dot_data
```

Format

An object of class `data.frame` with 350 rows and 4 columns.

FastSEAscore *Calculate the enrichment score of the pathways .*

Description

The function 'FastSEAscore' is used to calculate the pathway enrichment score.

Usage

```
FastSEAscore(labels.list, correl_vector)
```

Arguments

labels.list	The position of the genes in the pathway in the ranked gene list .
correl_vector	A ranked list of all genes in the PPI network.

Value

Enrichment scores of pathways based on predefined gene ranked lists.

Examples

```
#load the data
pathway_path<-system.file("extdata","kegg_323_gmt.Rdata",package = "ssMutPA")
load(pathway_path)
pathway_list<-split(kegg_323_gmt[,2],kegg_323_gmt[,1])
data(RWR_res)
gene_list<-sort(RWR_res,decreasing=TRUE)
tag.indicator <- sign(match(names(gene_list), pathway_list[[1]], nomatch = 0))
#perform the function `FastSEAscore`.
Path_ES<-FastSEAscore(labels.list=tag.indicator,correl_vector = gene_list)
names(Path_ES)<-names(pathway_list)[1]
```

get_heatmap *Plotting a heatmap with subtype labels.*

Description

The function 'get_heatmap' is used to plot a heatmap with subtype labels.

Usage

```
get_heatmap(
  Path_ES,
  Path_name,
  samp_class,
  scale = "row",
  cluster_rows = TRUE,
  cluster_cols = FALSE,
  show_rownames = TRUE,
  show_colnames = FALSE,
  fontsize = 8,
  annotation_legend = TRUE,
  annotation_names_row = TRUE,
  annotation_names_col = TRUE
)
```

Arguments

Path_ES	Single-sample mutation-based pathway enrichment score profiles.The file can be generated by the function 'get_RWR_ES'.
Path_name	The names of the pathways that you want to show in the heatmap.The Path_name must be included in the row names of the Path_ES .
samp_class	A vector containing information about the subtype labels.The vector can be generated by the function 'get_samp_class'.
scale	character indicating if the values should be centered and scaled in either the row direction or the column direction, or none. Corresponding values are "row", "column" and "none".
cluster_rows	Boolean values determining if rows should be clustered or hclust object.
cluster_cols	Boolean values determining if columns should be clustered or hclust object.
show_rownames	Boolean specifying if row names are be shown.
show_colnames	Boolean specifying if column names are be shown.
fontsize	base font size for the plot.
annotation_legend	Boolean value showing if the legend for annotation tracks should be drawn .
annotation_names_row	Boolean value showing if the names for row annotation tracks should be drawn.
annotation_names_col	Boolean value showing if the names for column annotation tracks should be drawn.

Value

A heatmap

Examples

```
#load the data
data(Path_ES,sample_class,Path_Name)
#perform the function `get_heatmap`.
get_heatmap(Path_ES,Path_name=Path_Name,samp_class=sample_class)
```

`get_indicate` *Indicator function.*

Description

Indicator function.

Usage

```
get_indicate(vector, cutpoint)
```

Arguments

`vector`	A numerical value.
`cutpoint`	The threshold of the indicator function.

Value

An integer with the value 1 or 0.

`get_mut_status` *Converts MAF file or data in other formats into mutation matrix.*

Description

The function 'get_mut_status' is used to convert MAF file or data in other formats into a binary mutation matrix.

Usage

```
get_mut_status(maf_data, nonsynonymous = TRUE, TCGA = TRUE, mut_rate = 0)
```

Arguments

`maf_data`	The patients' somatic mutation data, which in MAF format or others.
`nonsynonymous`	Logical. Determine if extract the non-silent somatic mutations .
`TCGA`	Logical. Determine whether the file is in MAF format .
`mut_rate`	Used to filter genes with target mutation rate .

Value

A binary mutations matrix, in which 1 represents that a particular gene has mutated in a particular sample, and 0 represents that gene is wild type.

Examples

```
#load the data
mut_path <- system.file("extdata","mutation_data.Rdata",package = "ssMutPA")
load(mut_path)
#perform the function `get_mut_status`.
mut_status<-get_mut_status(mutation_data,nonsynonymous=TRUE,TCGA=TRUE,mut_rate=0)
```

`get_RWR_ES` *Calculate the single-sample mutation-based pathway enrichment score.*

Description

The function 'get_RWR_ES' is used to calculate the single-sample mutation-based pathway enrichment score. Using somatic mutation data,PPI network and pathway data.

Usage

```
get_RWR_ES(
  mut_status,
  min_sample = 0,
  max_sample = dim(mut_status)[1],
  net_data,
  pathway_data,
  r = 0.7,
  Numcore = 2,
  BC_Num = length(V(net_data)$name),
  cut_point = 0
)
```

Arguments

`mut_status`	A binary mutation matrix.The file can be generated by the function 'get_mut_status'.
`min_sample`	The minimum number of mutated genes contained in a sample,default to 0.
`max_sample`	The maximum number of mutated genes contained in a sample.
`net_data`	A list of the PPI network information, including nodes and edges.
`pathway_data`	A data frame containing the pathways and their corresponding genes. The first column is the names of pathways and the second column is the genes included in the pathways.
`r`	A numeric value between 0 and 1. r is a certain probability of continuing the random walk or restarting from the restart set. Default to 0.7.

Numcore	The number of threads when running programs with multiple threads,default to 2 .
BC_Num	Number of background genes required to calculate seed node weight.
cut_point	The threshold of indicator function .

Value

A single-sample mutation-based pathway enrichment score profiles, where each element represents the enrichment score of a pathway in a sample.

Examples

```
#load the data
data(mut_status)
net_path <- system.file("extdata","ppi_network.Rdata",package = "ssMutPA")
load(net_path)
pathway_path<-system.file("extdata","kegg_323_gmt.Rdata",package = "ssMutPA")
load(pathway_path)
samp_name<-c("TCGA-06-0881-01A","TCGA-76-4934-01A")
examp_data<-mut_status[,samp_name]
#perform the function `get_RWR_ES`.
Path_ES<-get_RWR_ES(examp_data,net_data=ppi_network,pathway_data=kegg_323_gmt,BC_Num=12436)
```

---

get_samp_class *Determining cancer subtypes using unsupervised spectral clustering algorithm.*

Description

The function ‘get_samp_class‘ using spectral clustering algorithm to obtain cancer subtype labels.

Usage

```
get_samp_class(
  Path_ES,
  sur,
  seed_num = 50,
  cox_pval = 0.05,
  min.nc = 2,
  max.nc = 5
)
```

Arguments

Path_ES	Single-sample mutation-based pathway enrichment score(ssMutPES) profiles.The file can be generated by the function ‘get_RWR_ES‘.
sur	A matrix containing the samples’ survival time and survival states.

seed_num	The number of seeds for iterating to select the optimal clustering result.
cox_pval	A custom threshold to filter characteristic pathways, default to 0.05.
min.nc	Minimal number of clusters.
max.nc	maximal number of clusters,greater or equal to min.nc.

Value

A list containing the filtered pathways,the best seed for clustering,and cancer subtyoe labels .

Examples

```
#load the data.
surv_path <- system.file("extdata","sur.Rdata",package = "ssMutPA")
load(surv_path)
data(Path_ES)
#perform function `get_samp_class`.
res<-get_samp_class(Path_ES,sur,seed_num=5,cox_pval=0.05,min.nc = 2,max.nc =5)
```

---

get_seeds_score *Calculate the local wight of seed nodes.*

Description

The function ‘get_seeds_score‘ is used to calculate the local wight of seed nodes in a single sample.

Usage

```
get_seeds_score(net_data, seed, mut_gene, BC_Num, cut_point = 0)
```

Arguments

net_data	A list of the PPI network information, including nodes and edges.
seed	A vector containing the gene symbols of the seed nodes.
mut_gene	A vector containing the gene symbols of the mutated genes in a single sample.
BC_Num	Number of background genes.
cut_point	The threshold of indicator function.

Value

A data frame containing the weight of seed nodes.

Examples

```
#load the data
net_path <- system.file("extdata","ppi_network.Rdata",package = "ssMutPA")
load(net_path)
data(mut_status)
seed<-intersect(names(mut_status[,1])[which(mut_status[,1]!=0)],igraph::V(ppi_network)$name)
mut_gene<-intersect(names(mut_status[,1])[which(mut_status[,1]!=0)],igraph::V(ppi_network)$name)
#perform the function `get_seeds_score`.
Seeds_Score<-get_seeds_score(net_data=ppi_network,seed=seed,mut_gene,BC_Num=12436,cut_point=0)
```

---

get_univarCox_result *Perform the univariate Cox proportional hazards regression analysis.*

Description

The function 'get_univarCox_result' is used to perform the univariate Cox proportional hazards regression analysis.

Usage

```
get_univarCox_result(DE_path_sur)
```

Arguments

DE_path_sur	A matrix containing the activity values of all pathways in each sample, along with the survival time and survival status of the samples.Note that the column names of survival time and survival status must be "survival" and "event".

Value

A data frame containing the pathways' coefficient, HR, confidence interval, and survival related difference p-value .

Examples

```
#get path of the mutation annotation file.
data(cox_data)
#perform function `get_univarCox_result`.
res<-get_univarCox_result(cox_data)
```

mountain_plot *Plotting the density ridges plot.*

Description

The function 'mountain_plot' is used to draw a graph to reflect the distribution of the data.

Usage

```
mountain_plot(data, sample_class, Path_name)
```

Arguments

data	A pathway activity score matrix,which row names represent the pathways and the column names are samples.
sample_class	A vector containing subtype labels of the samples.
Path_name	The names of the pathways that you want to show in the graph.The 'Path_name' must be included in the row names of the data.

Value

Density ridges plot

Examples

```
#load the data
data(Path_ES,sample_class)
#perform the function `mountain_plot`.
mountain_plot(data=Path_ES,sample_class=sample_class,Path_name=rownames(Path_ES)[c(12,20,74)])
```

MRWR *A global propagation algorithm, random walk with restart (RWR), to predict probable influence of nodes in the network by seed nodes.*

Description

The function 'MRWR' is used to predict probable influence of nodes in the network by seed nodes.

Usage

```
MRWR(
  net_AdjMatrNorm,
  Seeds,
  net_data,
  mut_gene,
  r = 0.7,
  BC_Num = length(V(net_data)$name),
  cut_point = 0
)
```

Arguments

net_AdjMatrNorm	Row normalized network adjacency matrix.
Seeds	A vector containing the gene symbols of the seed nodes.
net_data	A list of the PPI network information,including nodes and edges .
mut_gene	A vector containing the gene symbols of the mutated genes in a sample.
r	A numeric value between 0 and 1. r is a certain probability of continuing the random walk or restarting from the restart set. Default to 0.7.
BC_Num	Number of background genes required to calculate seed node weight.
cut_point	The threshold of indicator function .

Value

An matrix of global weight, where the row names are genes in the network and the column names are samples.

Examples

```
#load the data
net_path <- system.file("extdata","ppi_network.Rdata",package = "ssMutPA")
load(net_path)
net_AdjMatr<-as.matrix(igraph::get.adjacency(ppi_network))
net_AdjMatrNorm <- t(t(net_AdjMatr)/(Matrix::colSums(net_AdjMatr, na.rm = FALSE, dims = 1)))
data(mut_status)
mut_gene<-intersect(names(mut_status[,1])[which(mut_status[,1]!=0)],igraph::V(ppi_network)$name)
seed<-intersect(names(mut_status[,1])[which(mut_status[,1]!=0)],igraph::V(ppi_network)$name)
#perform the function `MRWR`.
RWR_res<-MRWR(net_AdjMatrNorm,Seeds=seed,net_data=ppi_network,mut_gene,BC_Num = 12436)
```

mut_onco *mut_onco A binary mutation matrix.*

Description

mut_onco is used to calculate the mutation frequency of genes..

Usage

```
mut_onco
```

Format

An object of class `matrix` (inherits from `array`) with 11968 rows and 44 columns.

mut_status *mut_status A binary mutation matrix.*

Description

The mut_status is a binary mutation matrix,in which 1 represents mutation and 0 represents wild type.

Usage

```
mut_status
```

Format

An object of class `matrix` (inherits from `array`) with 12802 rows and 50 columns.

Oncoplot *Drawing a waterfall plot of a particular pathway.*

Description

Load the data in MAF format and draw a waterfall plot.

Usage

```
Oncoplot(
  maf,
  samp_class,
  sur,
  mut_status,
  pathway,
  pathway_name,
  isTCGA = FALSE,
  top = 20,
  clinicalFeatures = c("sample_group", "event"),
  class_col = c("#00468B", "#ED0000"),
  event_col = c("#B3DE69", "#BC80BD"),
  sortByAnnotation = TRUE,
  gene_mar = 7,
  removeNonMutated = FALSE,
  drawRowBar = TRUE,
  drawColBar = TRUE,
  leftBarData = NULL,
  leftBarLims = NULL,
  rightBarData = NULL,
  rightBarLims = NULL,
  topBarData = NULL,
  logColBar = FALSE,
  draw_titv = FALSE,
  showTumorSampleBarcodes = FALSE,
  fill = TRUE,
  showTitle = TRUE,
  titleText = NULL,
  vc_cols = NULL
)
```

Arguments

maf	A data of MAF format.
samp_class	A vector containing subtype labels of the samples.
sur	A matrix containing the samples' survival time and survival status.
mut_status	A binary mutations matrix.The file can be generated by the function 'get_mut_status'.
pathway	A list containing pathway information .
pathway_name	The names of the pathways that you want to visualize.For example "JAK-STAT signaling pathway".
isTCGA	Is input MAF file from TCGA source? If TRUE uses only first 12 characters from Tumor_Sample_Barcode.
top	How many top genes to be drawn,genes are arranged from high to low depending on the frequency of mutations. defaults to 20.

clinicalFeatures
: Columns names from 'clinical.data' slot of MAF to be drawn in the plot.

class_col
: The color of sample class .

event_col
: The color of survival status .

sortByAnnotation
: Logical sort oncomatrix (samples) by provided 'clinicalFeatures'. Sorts based on first 'clinicalFeatures'. Defaults to TRUE. column-sort.

gene_mar
: Margin width for gene names.

removeNonMutated
: Logical. If TRUE removes samples with no mutations in the GenePathwayOncoplots for better visualization. Default FALSE.

drawRowBar
: Logical. Plots righ barplot for each gene. Default TRUE.

drawColBar
: Logical plots top barplot for each sample. Default TRUE.

leftBarData
: Data for leftside barplot. Must be a data.frame with two columns containing gene names and values. Default 'NULL'.

leftBarLims
: Limits for 'leftBarData'. Default 'NULL'.

rightBarData
: Data for rightside barplot. Must be a data.frame with two columns containing to gene names and values. Default 'NULL' which draws distibution by variant classification. This option is applicable when only 'drawRowBar' is TRUE.

rightBarLims
: Limits for 'rightBarData'. Default 'NULL'.

topBarData
: Default 'NULL' which draws absolute number of mutation load for each sample. Can be overridden by choosing one clinical indicator(Numeric) or by providing a two column data.frame contaning sample names and values for each sample. This option is applicable when only 'drawColBar' is TRUE.

logColBar
: Plot top bar plot on log10 scale. Default FALSE.

draw_titv
: Logical Includes TiTv plot. Default FALSE

showTumorSampleBarcodes
: Logical to include sample names.

fill
: Logical. If TRUE draws genes and samples as blank grids even when they are not altered.

showTitle
: Default TRUE.

titleText
: Custom title. Default 'NULL'.

vc_cols
: named vector of colors for each Variant_Classification.

Value

A waterfall plot

Examples

```
#load the data
mut_path <- system.file("extdata","maffile.txt",package = "ssMutPA")
maf<-maftools::read.maf(mut_path ,isTCGA = FALSE)
pathway_path <- system.file("extdata","kegg_323_gmt.Rdata",package = "ssMutPA")
```

```
load(pathway_path)
data(samp_class_onco,mut_onco,sur_onco)
samples <- names(samp_class_onco)
samp_class_onco <- paste0("class_",samp_class_onco)
names(samp_class_onco) <- samples
sur_onco$event <- ifelse(sur_onco$event%in%1,"Dead","Alive")
col <- c("#8DD3C7", "#FFFFB3", "#BEBADA", "#FB8072", "#80B1D3")
##draw a waterfall plot
#win.graph()
Oncoplot(maf,samp_class_onco,sur_onco,mut_onco,kegg_323_gmt,"IL-17 signaling pathway",vc_cols=col)
```

`Path_ES` *Path_ES Single-sample mutation-based pathway enrichment score profiles.*

Description

The Path_ES is used for clustering and other analysis .

Usage

```
Path_ES
```

Format

An object of class `matrix` (inherits from `array`) with 215 rows and 882 columns.

`Path_Name` *Path_Name*

Description

The pathway names that the user wants to display.

Usage

```
Path_Name
```

Format

An object of class `character` of length 50.

RWR_res *RWR_res Random walk results based on mutated genes in a single sample.*

Description

The RWR_res is random walk results,used for enrichment analysis .

Usage

```
RWR_res
```

Format

An object of class `numeric` of length 12436.

sample_class *sample_class The subtype labels of samples.*

Description

The subtype labels of samples.

Usage

```
sample_class
```

Format

An object of class `integer` of length 882.

samp_class_onco *samp_class_onco*

Description

Sample subtype labels.

Usage

```
samp_class_onco
```

Format

An object of class `integer` of length 44.

Seeds_Score *Seeds_Score The local weight of seed nodes.*

Description

Seeds_Score

Usage

```
Seeds_Score
```

Format

An object of class `data.frame` with 35 rows and 2 columns.

sur_onco *sur_onco*

Description

Patient survival-related information.

Usage

```
sur_onco
```

Format

An object of class `data.frame` with 44 rows and 2 columns.

Index