Package 'genekitr'

May 27, 2022

Type Package

Title Gene Analysis Toolkit in R

Version 0.8.5

Maintainer Yunze Liu <jieandze1314@gmail.com>

URL https://github.com/GangLiLab/genekitr

BugReports https://github.com/GangLiLab/genekitr/issues

Description An analysis toolkit focused on genes. It mainly includes five features (search, convert, analysis, plot, and export). The user just needs to input feature id ('entrez', 'symbol', 'ensembl' or 'uniprot') to retrieve feature information and PubMed records, to convert id types, to easily do enrichment analysis and draw publication-level plots of GO, KEGG and GSEA, to plot group interaction and export results as sheets in one excel file to easily share and communicate with others.

License GPL-3

Encoding UTF-8

LazyData true

Depends R (>= 4.0)

Imports clusterProfiler, dplyr, europepmc, fst, ggplot2, ggraph, igraph, magrittr, openxlsx, stringr, stringi, tidyr, VennDiagram, rlang

Suggests AnnotationDbi, BiocManager, cowplot, DOSE, data.table, easyPubMed, forcats, fgsea, futile.logger, ggplotify, ggsci, ggupset, ggrepel, ggridges, ggnewscale, GOplot, GOSemSim, msigdbr, RColorBrewer, rappdirs, rentrez, reshape2, rio, rrvgo, scales, stats, testthat (>= 3.0.0), tibble, knitr, rmarkdown

RoxygenNote 7.2.0

NeedsCompilation no

Author Yunze Liu [aut, cre] (<https://orcid.org/0000-0002-7414-8556>)

Repository CRAN

Date/Publication 2022-05-27 10:00:08 UTC

R topics documented:

`as.enrichdat` *Modify dataframe for enrichment plot*

Description

To make sure colname contains Description, Count, FoldEnrich/GeneRatio, pvalue/qvalue/p.adjust

Usage

```
as.enrichdat(enrich_df)
```

Arguments

`enrich_df`	Enrichment analysis ‘data.frame‘ result.

Value

‘data.frame‘

Datasets *Datasets geneList entrez gene list with decreasing fold change value*

Description

Datasets geneList entrez gene list with decreasing fold change value

Datasets Differential expression analysis result of GSE42872

Datasets msig_species contains msigdb species information

Datasets msig_category contains msigdb category information

Datasets biocOrg_name contains organism name of bioconductor

Datasets keggOrg_name contains organism name of KEGG https://www.genome.jp/kegg/catalog/org_list.html

Datasets ensOrg_name contains organism name of ensembl

expoSheet *Export list of datasets into different Excel sheets*

Description

Export list of datasets into different Excel sheets

Usage

```
expoSheet(
  data_list,
  name_list,
  filename = NULL,
  dir = tempdir(),
  overwrite = TRUE
)
```

Arguments

data_list	List of datasets.
name_list	List of data names.
filename	A character string naming an xlsx file.
dir	A character string naming output directory.
overwrite	If TRUE, overwrite any existing file.

Value

An Excel file.

Examples

```
## Not run:
library(openxlsx)
expoSheet(
  data_list = list(mtcars, ToothGrowth),
  name_list = list("mtcars", "tooth"),
  filename = "test.xlsx", dir = tempdir()
)

## End(Not run)
```

---

genGO *Gene GO enrichment analysis*

Description

Gene GO enrichment analysis

Usage

```
genGO(
  id,
  group_list = NULL,
  org,
  ont,
  pAdjustMethod = "BH",
  pvalueCutoff = 0.05,
  qvalueCutoff = 0.1,
  minGSSize = 10,
  maxGSSize = 500,
  universe,
  ...
)
```

Arguments

id	A vector of gene id which can be entrezid, ensembl or symbol.
group_list	A list of gene id groups, default is NULL.
org	Organism name from 'biocOrg_name'.
ont	One of "bp", "mf", and "cc" subontologies, or "all" for all three.
pAdjustMethod	One of "holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", "none".
pvalueCutoff	Adjusted pvalue cutoff, default is 0.05.
qvalueCutoff	Adjusted pvalue cutoff, default is 0.1.
minGSSize	Minimal size of each gene set for analyzing, default is 10.
maxGSSize	Maximal size of each gene set for analyzing, default is 500.

universe	Background genes. If missing, then all gene list in orgdb will be used as background.
...	other argument to ‘enrichGO‘ function

Value

A ‘data.frame‘ contains gene ratio and fold enrichment.

Examples

```
## Not run:
data(geneList, package = "genekitr")

# only gene ids
id <- names(geneList)[abs(geneList) > 2]
ego <- genGO(id,
  org = "human", ont = "bp", pvalueCutoff = 0.01,
  qvalueCutoff = 0.01)

# gene id with groups
id <- c(head(names(geneList),50),tail(names(geneList),50))
group <- list(group1  = c(rep('up',50),rep('down',50)),
  group2 = c(rep('A',40),rep('B',60)))

gego <- genGO(id, group_list = group,
  org = "human", ont = "bp", pvalueCutoff = 0.1,
  qvalueCutoff = 1)


## End(Not run)
```

---

genGSEA *GSEA for a gene list with decreasing logFC value*

Description

GSEA for a gene list with decreasing logFC value

Usage

```
genGSEA(
  genelist,
  org,
  category = c("C1", "C2", "C3", "C4", "C5", "C6", "C7", "C8", "H"),
  subcategory = NULL,
  use_symbol = TRUE,
  minGSSize = 10,
  maxGSSize = 500,
  pvalueCutoff = 0.05,
```

```
  ...
)
```

Arguments

genelist	Order ranked genelist in decreasing order, gene can be entrez, ensembl or symbol.
org	Organism name from 'msig_org'.
category	MSigDB collection abbreviation, one of C1','C2','C3', 'C4','C5','C6','C7','C8','H'.
subcategory	MSigDB sub-collection abbreviation, choose from 'msig_category'.
use_symbol	Logical to set result gene id as gene symbol, default is TRUE.
minGSSize	Minimal size of each geneSet for analyzing, default is 10.
maxGSSize	Maximal size of each geneSet for analyzing, default is 500.
pvalueCutoff	Adjusted pvalue cutoff, default is 0.05.
...	Other argument to 'GSEA' function

Value

GSEA list

Examples

```
data(geneList, package = "genekitr")
gse <- genGSEA(genelist = geneList, org = "human",
  category = "H",use_symbol = TRUE)
```

---

genInfo *Get gene related information*

Description

Get gene related information

Usage

```
genInfo(id = NULL, org = "hs", unique = FALSE)
```

Arguments

id	Gene id (symbol, ensembl or entrez id) or uniprot id. If this argument is NULL, return all gene info.
org	Latin organism shortname from 'ensOrg_name'. Default is human.
unique	Logical, if one-to-many mapping occurs, only keep one record with fewest NA. Default is FALSE.

Value

A ‘data.frame‘.

Examples

```
# example1: input list with fake id and one-to-many mapping id
x = genInfo(id = c(
  "MCM10", "CDC20", "S100A9", "MMP1", "BCC7",
  "FAKEID", "TP53", "HBD", "NUDT10"))

# example2: statistics of human gene biotypes
genInfo(org = 'hs') %>% {table(.$gene_biotype)}
```

---

genKEGG *Gene enrichment of KEGG analysis*

Description

Gene enrichment of KEGG analysis

Usage

```
genKEGG(
  id,
  group_list = NULL,
  org,
  pAdjustMethod = "BH",
  pvalueCutoff = 0.05,
  qvalueCutoff = 0.1,
  minGSSize = 10,
  maxGSSize = 500,
  universe,
  ...
)
```

Arguments

id	A vector of entrez gene.
group_list	A list of gene id groups, default is NULL.
org	KEGG organism name from ‘keggOrg_name‘.
pAdjustMethod	One of "holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", "none".
pvalueCutoff	Numberic of adjusted pvalue cutoff, default is 0.05.
qvalueCutoff	Numberic of adjusted pvalue cutoff, default is 0.1.

minGSSize	Numberic of minimal size of each geneSet for analyzing, default is 10.
maxGSSize	Numberic of maximal size of each geneSet for analyzing, default is 500.
universe	Background genes. If missing, the orgdb all gene list will be used as background.
...	Other argument to 'enrichKEGG' function

Value

A 'data.frame'.

Examples

```
## Not run:
# only gene ids
data(geneList, package = "genekitr")
id <- names(geneList)[abs(geneList) > 1]
keg <- genKEGG(id, org = "human")

# gene id with groups
id <- c(head(names(geneList),100),tail(names(geneList),100))
group <- list(group1  = c(rep('up',100),rep('down',100)),
              group2 = c(rep('A',130),rep('B',70)))
gkeg <- genKEGG(id, group_list = group,
              org = "human", pvalueCutoff = 0.05,
              qvalueCutoff = 0.05)


## End(Not run)
```

---

getPubmed *Get pubmed paper records by searching abstract*

Description

Get pubmed paper records by searching abstract

Usage

```
getPubmed(term, keys)
```

Arguments

term	query terms e.g. gene id, GO/KEGG term or id
keys	other searching keys

Value

A list of 'tibble' for pubmed records

Examples

```
term <- c("Tp53","Brca1","Tet2")
keys <- c('stem cell','mouse')
l <- getPubmed(term,keys)
# very easy to output
expoSheet(l,name_list = term, filename = 'test.xlsx',dir = tempdir())
```

---

importPanther *Import Panther web result*

Description

Import Panther web result

Usage

```
importPanther(panther_file)
```

Arguments

panther_file	Panther result file.

Value

'data.frame'

---

plotEnrich *Plot for GO and KEGG enrichment analysis*

Description

Plot for GO and KEGG enrichment analysis

Usage

```
plotEnrich(
  enrich_df,
  fold_change = NULL,
 plot_type = c("bar", "wego", "dot", "bubble", "lollipop", "geneheat", "genechord",
    "network", "gomap", "goheat", "gotangram", "wordcloud", "upset"),
  term_metric = c("FoldEnrich", "GeneRatio", "Count", "RichFactor"),
  stats_metric = c("p.adjust", "pvalue", "qvalue"),
  sim_method = c("JC", "Resnik", "Lin", "Rel", "Jiang", "Wang"),
```

```
  up_color = "red",
  down_color = "blue",
  show_gene = "all",
  xlim_left = 0,
  xlim_right = NA,
  wrap_length = NULL,
  scale_ratio = 1,
  org = NULL,
  ont = NULL,
  layout,
  ...
)
```

Arguments

enrich_df	Enrichment analysis ‘data.frame‘ result.
fold_change	Fold change or logFC values with gene IDs as names. Used in "heat" and "chord" plot.
plot_type	Choose from "bar", "wego","bubble","dot", "lollipop","geneheat", "genechord", "network","gomap","goheat","gotangram","wordcloud","upset".
term_metric	Pathway term metric from one of 'GeneRatio','Count','FoldEnrich' and 'RichFactor'.
stats_metric	Statistic metric from one of "pvalue", "p.adjust", "qvalue".
sim_method	Method of calculating the similarity between nodes, one of one of "Resnik", "Lin", "Rel", "Jiang" , "Wang" and "JC" (Jaccard similarity coefficient) methods. Used in "map","goheat","gotangram","wordcloud".
up_color	Color of stronger statistics (e.g. Pvalue 0.01) or higher logFC, default is "red".
down_color	Color of weaker statistics (e.g. Pvalue 1) or lower logFC, default is "blue".
show_gene	Select genes to show. Default is "all". Used in "heat" and "chord" plot.
xlim_left	X-axis left limit, default is 0.
xlim_right	X-axis right limit, default is NA.
wrap_length	Numeric, wrap text if longer than this length. Default is NULL.
scale_ratio	Numeric, scale of node and line size. Default is 1. Used in "network" and "gomap".
org	Organism name from ‘biocOrg_name‘.
ont	One of "BP", "MF", and "CC".
layout	Grapgh layout in "map" plot, e,g, "circle", "dh", "drl", "fr","graphopt", "grid", "lgl", "kk", "mds", "nicely" (default),"randomly", "star".
...	other arguments from ‘plot_theme‘ function

Value

A ggplot object

Examples

```
## Not run:
## example data
library(ggplot2)
library(igraph)
library(ggraph)
data(geneList, package = "genekitr")
id <- names(geneList)[abs(geneList) > 1.5]
logfc <- geneList[id]

ego <- genGO(id,
  org = "human", ont = "bp", pvalueCutoff = 0.01,
  qvalueCutoff = 0.01)
ego <- ego[1:10,]
all_ego <- genGO(id,
  org = "human", ont = "all", pvalueCutoff = 0.01,
  qvalueCutoff = 0.01)

## example plots
plotEnrich(ego,plot_type = "dot")

plotEnrich(ego,plot_type = "bubble",scale_ratio = 0.4)

plotEnrich(ego,plot_type = "bar")

plotEnrich(all_ego,plot_type = 'wego')

plotEnrich(ego,plot_type = "lollipop",
down_color = "#325CAC", up_color = "#E69056",
wrap_length = 25, scale_ratio = 0.4)

plotEnrich(ego,plot_type = "geneheat")

show_gene = c('BRCA2','CDK1','JUN','MCM8','TIPIN')
plotEnrich(ego,plot_type = "geneheat",show_gene = show_gene)
plotEnrich(ego,fold_change = logfc, plot_type = "geneheat",show_gene = show_gene)

plotEnrich(ego,fold_change = logfc,plot_type = "genechord",show_gene = show_gene)

plotEnrich(ego,plot_type = "network", scale_ratio = 0.5)

plotEnrich(ego,plot_type = "gomap", wrap_length = 25)

plotEnrich(ego,plot_type = "goheat",sim_method = 'Rel')

plotEnrich(ego,plot_type = "gotangram",sim_method = 'Rel')

plotEnrich(ego,plot_type = "wordcloud",sim_method = 'Rel')

plotEnrich(ego,plot_type = "upset")

## End(Not run)
```

plotGSEA *GSEA plot*

Description

GSEA plot

Usage

```
plotGSEA(
  gsea_list,
  plot_type = c("volcano", "classic", "fgsea", "ridge", "bar"),
  stats_metric = c("p.adjust", "pvalue", "qvalue"),
  show_pathway = 3,
  show_gene = NULL,
  colors = NULL,
  ...
)
```

Arguments

gsea_list	GSEA result from ‘genGSEA‘ function
plot_type	GSEA plot type, one of ’volcano’, ’classic’, ’fgsea’, ’ridge’ or ’bar’.
stats_metric	Statistic metric from one of "pvalue", "p.adjust", "qvalue".
show_pathway	Select plotting pathways by specifying number (will choose top N pathways) or pathway name.
show_gene	Select genes to show. Default is "all". Used in "classic" plot.
colors	Color vector. Deafault is NULL.
...	other arguments transfer to ‘plot_theme‘ function

Value

A ggplot object

Examples

```
## Not run:
library(ggplot2)
## get GSEA result
data(geneList, package = "genekitr")
gse <- genGSEA(genelist = geneList, org = "human",
               category = "H",use_symbol = TRUE, pvalueCutoff = 0.05)
## volcano plot
# get top3 of up and down pathways
```

```
plotGSEA(gse, plot_type = 'volcano', show_pathway = 3)
# choose pathway by character
pathways = c("HALLMARK_P53_PATHWAY","HALLMARK_GLYCOLYSIS","HALLMARK_DNA_REPAIR")
plotGSEA(gse, plot_type = 'volcano',show_pathway = pathways)

## classic pathway plot
genes = c("MET", "TP53", "PMM2")
plotGSEA(gse, plot_type = 'classic', show_pathway = pathways, show_gene = genes)

## fgsea for multiple pathway
plotGSEA(gse, plot_type = 'fgsea', show_pathway = 3)

## ridgeplot
plotGSEA(gse, plot_type = 'ridge',
  show_pathway = 10, stats_metric = 'p.adjust')

## two-side barplot
plotGSEA(gse, plot_type = 'bar',main_text_size = 8,
colors = c('navyblue','orange'))


## End(Not run)
```

---

plotVenn *Venn plot for groups of genes*

Description

If gene group over 4, plot will be visulized using UpSet plot.

Usage

```
plotVenn(
  venn_list,
  use_venn = TRUE,
  color = NULL,
  alpha_degree = 0.3,
  text_size = 1,
  border_thick = 1,
  remove_grid = FALSE,
  ...
)
```

Arguments

venn_list	A list of gene id.
use_venn	Logical, use venn to plot, default is 'TRUE', the other option is upsetplot for large list.

color	Colors for gene lists, default is NULL.
alpha_degree	Alpha transparency of each circle's area, default is 0.3.
text_size	Text size, default is 1.
border_thick	Numeric, border thickness, default is 1.
remove_grid	Logical, remove circle or grid lines, default is 'FALSE'.
...	other arguments transfer to 'plot_theme' function

Value

A ggplot object

Examples

```
library(ggplot2)
set1 <- paste0(rep("gene", 100), sample(c(1:1000), 100))
set2 <- paste0(rep("gene", 100), sample(c(1:1000), 100))
set3 <- paste0(rep("gene", 100), sample(c(1:1000), 100))
set4 <- paste0(rep("gene", 100), sample(c(1:1000), 100))
set5 <- paste0(rep("gene", 100), sample(c(1:1000), 100))
sm_gene_list <- list(gset1 = set1, gset2 = set2, gset3 = set3)
la_gene_list <- list(
  gset1 = set1, gset2 = set2, gset3 = set3,
  gset4 = set4, gset5 = set5
)
plotVenn(sm_gene_list,
  text_size = 1.5, alpha_degree = 1,
  remove_grid = TRUE, color = ggsci::pal_lancet()(3))
plotVenn(la_gene_list,
  text_size = 15, alpha_degree = 0.2, border_thick = 2,
  remove_grid = TRUE, use_venn = FALSE)
```

---

plotVolcano *Volcano plot for differential expression analysis*

Description

Volcano plot for differential expression analysis

Usage

```
plotVolcano(
  deg_df,
  stat_metric = c("p.adjust", "pvalue"),
  stat_cutoff = 0.05,
  logFC_cutoff = 1,
  up_color = "red",
  down_color = "blue",
```

```
  show_gene = NULL,
  dot_size = 1.75,
  ...
)
```

Arguments

deg_df	DEG dataframe with gene id, logFC and stat(e.g. pvalue/qvalue).
stat_metric	Statistic metric from "pvalue" or "p.adjust".
stat_cutoff	Statistic cutoff, default is 0.05.
logFC_cutoff	Log2 fold change cutoff, default is 1 which is actually 2 fold change.
up_color	Color of up-regulated genes, default is "red".
down_color	Color of down-regulated genes, default is "blue".
show_gene	Select genes to show, default is no genes to show.
dot_size	Volcano dot size, default is 1.75.
...	other arguments from ‘plot_theme‘ function

Value

A ggplot object

Examples

```
## Not run:
library(ggplot2)
data(deg, package = "genekitr")
plotVolcano(deg,'p.adjust',remove_legend = T,dot_size = 3)

# show some genes
plotVolcano(deg,'p.adjust',remove_legend = T,
show_gene = c("CD36", "DUSP6", "NUPR1", "IER3"))

## End(Not run)
```

---

plot_theme *Themes for all plots*

Description

Change ggplot text, font, legend and border

Usage

```
plot_theme(
  main_text_size = 8,
  legend_text_size = 6,
  font_type = "sans",
  border_thick = 1.5,
  remove_grid = TRUE,
  remove_border = FALSE,
  remove_main_text = FALSE,
  remove_legend_text = FALSE,
  remove_legend = FALSE
)
```

Arguments

main_text_size	Numeric, main text size
legend_text_size	Numeric, legend text size
font_type	Character, specify the plot text font family, default is "sans".
border_thick	Numeric, border thickness, default is 1. If set 0, remove both border and ticks.
remove_grid	Logical, remove background grid lines, default is FALSE.
remove_border	Logical, remove border line, default is FALSE.
remove_main_text	Logical, remove all axis text, default is FALSE.
remove_legend_text	Logical, remove all legend text, default is FALSE.
remove_legend	Logical, remove entire legend, default is FALSE.

Value

ggplot theme

Examples

```
library(ggplot2)
ggplot(mtcars, aes(x=wt, y=mpg))+ geom_point()+
  plot_theme(font_type = 'Times', border_thick = 2)
```

---

transId *Transform gene id among symbol, entrezid, ensembl and uniprot.*

Description

Transform gene id among symbol, entrezid, ensembl and uniprot.

Usage

```
transId(id, transTo, org = "hs", unique = FALSE, keepNA = FALSE)
```

Arguments

id	Gene ids.
transTo	Transform to what type. User could select one or more from "symbol", "entrez", "ensembl" or "uniprot."
org	Latin organism shortname from ‘ensOrg_name‘. Default is human.
unique	Logical, if one-to-many mapping occurs, only keep one record with fewest NA. Default is FALSE.
keepNA	If some id has no match at all, keep it or not. Default is FALSE.

Value

A two-column data frame, first is input id and second is transformed id.

Examples

```
## Not run:
# example1:
transId(
  id = c("Cyp2c23", "Fhit", "Gal3st2b", "Trp53", "Tp53"),
  transTo = "ensembl", org = "mouse", keepNA = FALSE
)

## example2: input id with one-to-many mapping and fake one
transId(
  id = c("MMD2", "HBD", "RNR1", "TEC", "BCC7", "FAKEID", "TP53"),
  transTo = c("entrez","ensembl"), keepNA = TRUE
)

# example3: auto-recognize ensembl version number
transId('ENSG00000141510.11','symbol')

## End(Not run)
```

Index