Package 'TinderMIX'

July 20, 2020

Type Package

Title TinderMIX: An R package to cluster gene expression by contour plots

Version 0.1.0

Author Angela Serra

Maintainer Angela Serra <angela.serra@tuni.fi>

Description The TinderMIX package allows to analyse toxicogenomics data with multiple dose levels and time-points. It allows to identify the expression patterns with respect to both variables and to cluster molecular features accordingly. It also identify enriched pathways/go terms that are associated to each cluster.

Depends R (>= 3.4),
stats,
utils,
AnnotationDbi,
gProfileR,
gtools,
reshape,
plotly,
clv,
gplots,
org.Hs.eg.db,
org.Mm.eg.db,
org.Rn.eg.db,
xlsx,
pracma,
raster,
wordcloud,
ggplot2,
fmsb,
biomaRt

License GPL (>= 3)

Encoding UTF-8

LazyData true

RoxygenNote 6.1.1

Suggests knitr,
rmarkdown,
testthat

VignetteBuilder knitr

R topics documented:

`build_items_list` *This function computes the venn diagram of the genes associated to time, dose or their interaction*

Description

This function computes the venn diagram of the genes associated to time, dose or their interaction

Usage

```
build_items_list(PvalMat, p.val.th = 0.01)
```

Arguments

`PvalMat`	matrix with pvalue associated to the dose, timepoint and the dose*timepoint effect that is the output of the compute_anova_dose_time function
`p.val.th`	is the threshold at which p.values are considered significant. Default = 0.01

Value

a list containing the genes in each position of the venn diagram

`compute_anova_dose_time` *This function computes a two way anova between dose and time for the expression value of every genes*

Description

This function computes a two way anova between dose and time for the expression value of every genes

Usage

```
compute_anova_dose_time(exp_data, pheno_data, dose_index, time_point_index,
  adj.method = "none")
```

Arguments

`exp_data`	is the expression matrix with genes on the rows and samples on the columns
`pheno_data`	is a dataframe with phenodata informations. Samples are on the rows. The columns should include the dose and time point information.
`dose_index`	numeric value specifing the column of the phenodata table containing the doses
`time_point_index`	numeric value specifing the column of the phenodata table containing the time points
`adj.method`	a string specifying the adjustement method for the pvalue

Value

a matrix with pvalue associated to the dose, timepoint and the dose*timepoint effect

`compute_BMD_IC50` *This function identify the BMD area and the IC50 value in the time and dose maps*

Usage

```
compute_BMD_IC50(immy, coord, geneName, activity_threshold = 0.1,
  BMD_response_threshold = 0.95, nDoseInt = 3, nTimeInt = 3,
  doseLabels = c("Late", "Middle", "Early"),
  timeLabels = c("Sensitive", "Intermediate", "Resilient"),
  toPlot = TRUE, addLegend = TRUE, tosave = FALSE, path = ".",
  mode = "cumulative")
```

Arguments

immy	z-maps of the fitted 3D model, with doses on the columns and time points on the rows
coord	matrix with x and y coordinate. The first column contain the doses, while the second one the time points
geneName	is a character string containing the gene name
activity_threshold	threshold defining the responsive gene area. Eg. if the immy maps contains genes logFC, then an activity_threhdold = 0.58 means that the active area will be the one with an effect of 1.5 bigger or smaller than the controls
BMD_response_threshold	a threshold to define the portion of dose-response area to be identified as labels for the gene.
nDoseInt	number of dose related breaks in the gene label's table. default is 3
nTimeInt	number of time related breaks in the gene label's table. default is 3
doseLabels	vector of colnames (doses) for the gene label's table. default is c("Sensitive","Intermediate","Resilient
timeLabels	vector of rownames (time points) for the gene label's table. default c("Late","Middle","Early")
toPlot	it true the gene map is displayed
addLegend	if true the legend will be added to the plot
tosave	if true a png of the gene map is saved in path
path	path of the folder where to save the gene map
mode	is a character specifying when an area is called active. values can be "cumulative" or "presence". If presence, an area is called active if at least one of its pixel is on the BMD curve. If cumulative, the number of region needed to reach the th
	an object of class TinderMIX containing the fitted BMD object, the IC50 value. The function plot the map showing the responsive region.
	This function identify the BMD area and the IC50 value in the time and dose maps

---

compute_enrichment_for_clusters

This function perform enchment of the genes in each cluster

Description

This function perform enchment of the genes in each cluster

Usage

```
compute_enrichment_for_clusters(optimal_clustering, corrType = "fdr",
  type_enrich = "KEGG", org_enrich = "rnorvegicus", pth = 0.05,
  sig = FALSE, mis = 0, only_annotated = FALSE)
```

Arguments

optimal_clustering	vector of final clustering
corrType	string specifing the algorithm used for determining the significance threshold, one of gSCS, fdr, bonferroni. Default: fdr
type_enrich	string specifying the enrichment type. Default = KEGG
org_enrich	string specifying the organism. Default = rnorvegicus
pth	numeric value specifyint the pvalue threshold. Default = 0.05
sig	whether all or only statistically significant results should be returned
mis	minimum size of functional category, smaller categories are excluded
only_annotated	statistical domain size, one of "annotated", "known"

Value

a list with the enriched pathways for each cluster of genes

---

compute_fc *This function starts from a phenodata and gene expression data matrix and compute all the possible pairwise foldchange values*

Description

This function starts from a phenodata and gene expression data matrix and compute all the possible pairwise foldchange values

Usage

```
compute_fc(exp_data, pheno_data, dose_index, time_index)
```

Arguments

exp_data	is the expression matrix with genes on the rows and samples on the columns
pheno_data	is a dataframe with phenodata informations. Samples are on the rows. The columns should include the dose and time point information. Doses of controls need to be indicated as 0
dose_index	index of the column containiing the dose
time_index	index of the column containing the time

Value

a list containing two new matrices

fc_data	a matrix with all the possible fold_changes
pdata	the new phenodata table

`compute_pathways` *This function perform enchment of a set of genes*

Description

This function perform enchment of a set of genes

Usage

```
compute_pathways(geneList = rownames(res$Mat), corrType = "fdr",
  type_enrich = "KEGG", annType = "SYMBOL", org_enrich = "hsapiens",
  pth = 0.05, sig = FALSE, mis = 0, only_annotated = FALSE)
```

Arguments

`geneList`	vector of gene identifiers
`corrType`	string specifing the algorithm used for determining the significance threshold, one of gSCS, fdr, bonferroni. Default: fdr
`type_enrich`	string specifying the enrichment type. Default = KEGG
`annType`	gene annotation type. default = SYMBOL
`org_enrich`	string specifying the organism. Default = rnorvegicus
`pth`	numeric value specifyint the pvalue threshold. Default = 0.05
`sig`	whether all or only statistically significant results should be returned
`mis`	minimum size of functional category, smaller categories are excluded
`only_annotated`	statistical domain size, one of "annotated", "known"

Value

a list with the enriched pathways for each cluster of genes

`convert_genes` *This function convert genes identifiers*

Description

This function convert genes identifiers

Usage

```
convert_genes(organism = "hsapiens", GList, annType = "SYMBOL")
```

Arguments

`organism`	a string specifying the organism under analysis
`GList`	a list of genes identifier
`annType`	string specifying the wanted gene identifier

Value

a list with the converted genes identifiers

`create_contour` *This function fits a 3D regression model for every gene in the dataset and creates an N x N contour plot*

Description

This function fits a 3D regression model for every gene in the dataset and creates an N x N contour plot

Usage

```
create_contour(exp_data, pheno_data, responsive_genes, dose_index,
  time_point_index, gridSize = 50, pvalFitting.adj.method = "fdr",
  pvalFitting = 0.05, logScale = FALSE, modelSelection = c(1, 2))
```

Arguments

`exp_data`	is the expression matrix with genes on the rows and samples on the columns
`pheno_data`	is a dataframe with phenodata informations. Samples are on the rows. The columns should include the dose and time point information.
`responsive_genes`	responsive_genes character vector with the genes statistically significant for the two-way anova
`dose_index`	numeric value specifing the column of the phenodata table containing the doses
`time_point_index`	numeric value specifing the column of the phenodata table containing the time points
`gridSize`	numeric value specifing size of the z-grid
`logScale`	boolean specifying if the fitting is performed by using the dose and time in log or linear scale
`modelSelection`	is a vector of indices specifying which model to fit. 1:linear 2: poly2, 3: poly3

Value

a list with list with estimated contour objects, 3D fitted objects, fitting statistics and feature values for time and dose

`GenesMap`	a matrix with the z-maps computed for each gene
`RPGenes`	a list with the 3D fitted objects
`Statis`	a matrix witht the fitting statistics: PValue,Adj.R.Square,RMSE
`DFList`	a list with the data used for the fitting
`ModList`	a list with the fitted models

`create_gene_table` *This function create a table with the information on the dynamic-dose-dependent genes*

Description

This function create a table with the information on the dynamic-dose-dependent genes

Usage

```
create_gene_table(DDRGene, contour_res, nTimeInt, nDoseInt,
  biomart_dataset = "rnorvegicus_gene_ensembl")
```

Arguments

`DDRGene`	is the results of the run_all_BMD_IC50 function
`contour_res`	is the result of the create_contour function
`nTimeInt`	number of time points
`nDoseInt`	number of dose levels
`biomart_dataset`	is a string specifying the dataset to use in the useEnsembl function. e.g rnorvegicus_gene_ensembl

Value

a data frame

`create_pathway_prototypes` *This function create prototypes for a list of pathways*

Description

This function create prototypes for a list of pathways

Usage

```
create_pathway_prototypes(enrichedPath = enrichedPath, annIDs,
  contour_res = contour_res, nPerm = 100)
```

Arguments

`enrichedPath`	dataframe of enriched pathways coming from the compute_pathways function
`annIDs`	vector with the pathways IDs for which the prototype will be computed
`contour_res`	a list with the contours object returned in output by the create_contour function
`nPerm`	number of permutation to run to compute the pvalue associated to pathway correaltion

Value

a list with the prototypes of the pathways, their genes correlation and a pvalue

`create_tic_tac_toe_wordcloud` *This function perform enrichment for each gene label and compute a pathway wordcloud for every label*

Description

This function perform enrichment for each gene label and compute a pathway wordcloud for every label

Usage

```
create_tic_tac_toe_wordcloud(Enriched_list = Enriched_list,
  max.words = 200, scale = c(0.8, 2.5), random.order = FALSE,
  min.freq = 0, toplot = TRUE)
```

Arguments

`max.words`	max number of words in wordcluds
`scale`	a vector of length 2 indicating the range of the size of the words.
`random.order`	plot words in random order. If false, they will be plotted in decreasing frequency
`min.freq`	words with frequency below min.freq will not be plotted
`toplot`	boolean specifying if plotting wordcloud
`Mat`	matrix of gene labels

Value

a list with the enriched pathways for each cluster of genes

`enrich` *This function perform enchment of the genes in each cluster*

Description

This function perform enchment of the genes in each cluster

Usage

```
enrich(x, type, org, pval, adjust_method, sig = FALSE, mis = 0,
  only_annotated = TRUE)
```

Arguments

x	a dataframe with the gene names on the first column
type	string specifying the enrichment type. Default = KEGG
org	string specifying the organism. Default = rnorvegicus
pval	numeric value specifyint the pvalue threshold. Default = 0.05
adjust_method	string specifing the algorithm used for determining the significance threshold, one of gSCS, fdr, bonferroni. Default: fdr
sig	whether all or only statistically significant results should be returned
mis	minimum size of functional category, smaller categories are excluded
only_annotated	statistical domain size, one of "annotated", "known"

Value

a list with the enriched pathways for each cluster of genes

---

fisher_test *This function construct confusion matrix between patient classes and the obtained clustering*

Description

This function construct confusion matrix between patient classes and the obtained clustering

Usage

```
fisher_test(classes, clustering, matrixRownames, nCluster)
```

Arguments

classes	is a vector of patient labels
clustering	is a vector of clustering results
matrixRownames	is a vector of names to assign as rownames of the confusion matrix
nCluster	is the number of obtained clusters

Value

the confusion matrix

label2DMap *This function assigns a label to contour map*

Usage

```
label2DMap(map, BMD, coord, myContour, th = 0.95, mode = "mix",
  nDoseInt = 3, nTimeInt = 3, doseLabels = c("Late", "Middle",
  "Early"), timeLabels = c("Sensitive", "Intermediate", "Resilient"),
  toplot = FALSE)
```

Arguments

map	matrix containing the z-map for a specific gene or cluster prototype
BMD	matrix containing the dose-response area
coord	matrix with x and y coordinate. The first column contain the doses, while the second one the time points
myContour	matrix with coordinate of bmd area border
th	a threshold to define the portion of dose-response area to be identified as labels for the gene.
mode	is a character specifying when an area is called active. values can be "cumulative" or "presence". If presence, an area is called active if at least one of its pixel is on the BMD curve. If cumulative, the number of region needed to reach the th
	\itemnDoseIntnumber of dose related breaks in the gene label's table. default is 3
	\itemnTimeIntnumber of time related breaks in the gene label's table. default is 3
	\itemdoseLabelsvector of colnames (doses) for the gene label's table. default is c("Sensitive","Intermediate","Resilient")
	\itemtimeLabelsvector of rownames (time points) for the gene label's table. default c("Late","Middle","Early")
	\itemnDoseIntnumber of dose related breaks in the gene label's table. default is 3
	\itemnTimeIntnumber of time related breaks in the gene label's table. default is 3
	\itemdoseLabelsvector of colnames (doses) for the gene label's table. default is c("Sensitive","Intermediate","Resilient")
	\itemtimeLabelsvector of rownames (time points) for the gene label's table. default c("Late","Middle","Early")
	a list with 9x9 matrices specifying if the gene is active at low, mid or high time points and dose levels
	This function assigns a label to contour map

`plot3d` *This function plots the fitted 3d surface for the expression value of a gene*

Description

This function plots the fitted 3d surface for the expression value of a gene

Usage

```
plot3d(toPlot = list(x, y, z), DF, logScale = FALSE)
```

Arguments

`toPlot`	is a list containing the predicted value for the x, y and z axis
`DF`	is the data frame containing the information for the samples used in the fitting process

Value

a plotly object

`plot_cake_diagrams_time_dose_effect`
this function takes in input the result of the function run_all_BMD_IC50 and plot a 3x3 multiplot with the number of dose responsive genes fir tge 12 segment of time and dose interaction the letters d and t (independently if they are capital or small) stand for dose and time +/- indicate if the gene fc is increasing or decreasing with respect of dose and time capital letters are used to indicate which between dose and time has a stronger effect

Description

this function takes in input the result of the function run_all_BMD_IC50 and plot a 3x3 multiplot with the number of dose responsive genes fir tge 12 segment of time and dose interaction the letters d and t (independently if they are capital or small) stand for dose and time +/- indicate if the gene fc is increasing or decreasing with respect of dose and time capital letters are used to indicate which between dose and time has a stronger effect

Usage

```
plot_cake_diagrams_time_dose_effect(res, timeLabels, doseLabels)
```

Arguments

`res`	is the result object from the run_all_BMD_IC50 function
`timeLabels`	is the vector with time labels predefined by the user
`doseLabels`	is the vector with dose labels predefined by the user

Value

a ggplot object

`plot_dynamic_dose_responsive_map` *This function takes in input the result of the function create_contour and plot the dynamic dose responsive activation map of a specific gene*

Description

This function takes in input the result of the function create_contour and plot the dynamic dose responsive activation map of a specific gene

Usage

```
plot_dynamic_dose_responsive_map(contour_res, geneName, activity_threshold,
  BMD_response_threshold, mode, nTimeInt, nDoseInt, timeLabels, doseLabels)
```

Arguments

`contour_res`	is the result object from the create_contour function
`geneName`	is the name of the gene

Value

a ggplot object

`plot_kegg_radar_chart` *this function takes in input the pathways enriched and retur a radar chart for each one of the gene label category*

Description

this function takes in input the pathways enriched and retur a radar chart for each one of the gene label category

Usage

```
plot_kegg_radar_chart(Enriched_list, n = 5, vlcex = 1.5,
  kegg_level = 1, mar = c(2, 1, 1, 1))
```

Arguments

`Enriched_list`	is the list of dataframe resulting from enrichment for each gene category
`n`	is the max number of pathways to plot in each radar plot
`vlcex`	is the size of the labels
`kegg_level`	is the level of the kegg hierarchy to be considered in the plotting
`mar`	are the margin settings for the plot
`caxislabels`	is a character vector for center axis labels, overwriting values specified in axistype option. If NULL, the values specified by axistype option are used. Default is NULL.

Value

a ggplot object

`plot_number_genes_labels` *This function takes in input the result of the function run_all_BMD_IC50 and plot a 3x3 heatmap with the number of dose responsive genes for each label*

Description

This function takes in input the result of the function run_all_BMD_IC50 and plot a 3x3 heatmap with the number of dose responsive genes for each label

Usage

```
plot_number_genes_labels(res, drugName, timeLabels, doseLabels)
```

Arguments

`res`	is the result object from the run_all_BMD_IC50 function
`drugName`	is the name of the drug that will be used in the title
`timeLabels`	is the vector with time labels predefined by the user
`doseLabels`	is the vector with dose labels predefined by the user

Value

a ggplot object

`read_excel_allsheets` *read excel file as a list of dataframe*

Description

read excel file as a list of dataframe

Usage

```
read_excel_allsheets(filename, tibble = FALSE)
```

Arguments

`filename`	is the path to the file
`tibble`	boolean specifying if the content of each sheet should be read as tibble or dataframe

Value

a data frame contained in the excel file

`run_all_BMD_IC50` *This function run the compute_BMD_IC50 for all genes and return a matrix with label associated to every gene*

Usage

```
run_all_BMD_IC50(contour_res, activity_threshold = 0.1,
  BMD_response_threshold = 0.95, nDoseInt = 3, nTimeInt = 3,
  doseLabels = c("Late", "Middle", "Early"),
  timeLabels = c("Sensitive", "Intermediate", "Resilient"),
  tosave = FALSE, toPlot = FALSE, addLegend = FALSE, path = ".",
  relGenes, mode = "cumulative")
```

Arguments

`contour_res`	object resulting from the create_contour function
`activity_threshold`	threshold defining the responsive gene area. Eg. if the immy maps contains genes logFC, then an activity_threhdold = 0.58 means that the active area will be the one with an effect of 1.5 bigger or smaller than the controls
`BMD_response_threshold`	a threshold to define the portion of dose-response area to be identified as labels for the gene.
`nDoseInt`	number of dose related breaks in the gene label's table. default is 3
`nTimeInt`	number of time related breaks in the gene label's table. default is 3
`doseLabels`	vector of colnames (doses) for the gene label's table. default is c("Sensitive","Intermediate","Resilient
`timeLabels`	vector of rownames (time points) for the gene label's table. default c("Late","Middle","Early")
`tosave`	if true a png of the gene map is saved in path
`path`	path of the folder where to save the gene map
`relGenes`	vector of genes with signifincant pvalues from the fitting
`mode`	is a character specifying when an area is called active. values can be "cumulative" or "presence". If presence, an area is called active if at least one of its pixel is on the BMD curve. If cumulative, the number of region needed to reach the th \itemcoordmatrix with x and y coordinate. The first column contain the doses, while the second one the time points \itemgeneNameis a character string containing the gene name a list with two object: Mat is a matrix with genes on the rows and labels on the columns. GeneRes is a list of results from the compute_BMD_IC50 function, one for every gene This function run the compute_BMD_IC50 for all genes and return a matrix with label associated to every gene

`write_xlsx_for_funmappone` *This function create an excel file with the same format of the input need by the FunMappOne tool*

Description

This function create an excel file with the same format of the input need by the FunMappOne tool

Usage

```
write_xlsx_for_funmappone(optimal_clustering,
  filePath = "../contour_clustering/gene_clustering.xlsx")
```

Arguments

`optimal_clustering`	is a numeric vector with the clustering result for every gene
`filePath`	is a string specifying the path of the xlsx file

Index