Package 'xmsPANDA'

March 22, 2021

Type Package

Title R Package for Biomarker Discovery, Network, and Data Exploratory
    Analysis

Version 1.3.1

Date 2021-03-19

Author Karan Uppal

Maintainer Karan Uppal <kuppal2@emory.edu>

Description Includes functions to perform feature selection for classification and regression, net-
    work analysis, data exploratory, normalization, and functional analysis.

License GPL2.0

LazyLoad yes

Depends R (>= 3.5.0)

R topics documented:

---

xmsPANDA-package   *xmsPANDA*

Description

R pacakge for biomarker discovery, supervised and unsupervised learning, and network analysis.

Details

Package:	xmsPANDA
Type:	Package
Version:	1.3
Date:	2021-03-10
License:	gpl2.0
LazyLoad:	yes

Author(s)

Karan Uppal Maintainer: <kuppal3gt@gmail.com>

---

data_preprocess   *data_preprocess*

Description

This function performs data transformation, normalization

Usage

```
function(Xmat = NA, Ymat = NA, feature_table_file = NA,
                parentoutput_dir, class_labels_file = NA,
                num_replicates = 3, feat.filt.thresh = NA,
                summarize.replicates = TRUE, summary.method = "mean",
                all.missing.thresh = 0.5, group.missing.thresh = 0.7,
                log2transform = TRUE, medcenter = FALSE,
                znormtransform = FALSE, quantile_norm = TRUE,
                lowess_norm = FALSE, rangescaling = FALSE,
                paretoscaling = FALSE, mstus = FALSE, sva_norm =
                FALSE, TIC_norm = FALSE, eigenms_norm = FALSE,
                madscaling = FALSE, vsn_norm = FALSE, cubicspline_norm
                = FALSE, missing.val = 0, samplermindex = NA,
                rep.max.missing.thresh = 0.5, summary.na.replacement =
                "zeros", featselmethod = NA, pairedanalysis = FALSE,
                normalization.method = "none", input.intensity.scale =
                "raw", create.new.folder = TRUE)
```

Arguments

Xmat	R object for feature table. If this is given, then feature table can be set to NA.
Ymat	R object for response/class labels matrix. If this is given, then class can be set to NA.
feature_table_file	Feature table that includes the mz, retention time, and measured intensity in each sample for each analyte. The first 2 columns should be the mz and time. The remaining columns should correspond to the samples in the class labels file with each column including the intensity profile of a sample. Full path required. Eg: C:/My Documents/test.txt The feature table should be in a tab-delimited format. An example of the input file is provided under the "example" folder.
parentoutput_dir	Provide full path of the folder where you want the results to be written. Eg: C:/My Documents/ProjectA/results/
class_labels_file	File with class labels information for each sample. Samples should be in the same order as in the feature table. Please use the same format as in the example folder.
num_replicates	Number of technical replicates
feat.filt.thresh	Percent Intensity Difference or Coefficient of variation threshold; feature filtering Use NA to skip this step.
summarize.replicates	Do the technical replicates per sample need to be averaged or median summarized?
summary.method	Method for summarizing the replicates. Options: "mean" or "median"
summary.na.replacement	How should the missing values be represented? Options: "zeros", "halffeaturemin", "halfsamplemin","halfdatamin", "none" "zeros": replaces missing values by 0 "halfsamplemin": replaces missing value by one-half of the lowest signal intensity in the corresponding sample "halfdatamin": replaces missing value by one-half of the lowest signal intensity in the complete dataset "halffeaturemin": replaces missing value by one-half of the lowest signal intensity for the current feature "none": keeps missing values as NAs Users are recommended to perform imputation prior to performing biomarker discovery.
missing.val	How are the missing values represented in the input data? Options: "0" or "NA"
samplermindex	Column index of any additional or irrelevant columns to be deleted. Options: "NA" or list of column numbers. eg: c(1,3,4) Default=NA
rep.max.missing.thresh	What propotion of replicates are allowed to have missing values during the averaging or median summarization step of each biological sample? If the number of replicates with missing values is greater than the defined threshold, then the summarized value is represented by the "missing.val" parameter. If the number of replicates with missing values is less than or equal to the defined threshold, then the summarized value is equal to the mean or the median of the non-missing values. Default: 0.5
all.missing.thresh	What propotion of total number of samples should have an intensity? Default: 0.5

group.missing.thresh	
	What propotion of samples in either of the two groups should have an intensity? If at least x for further analysis. Default: 0.7
log2transform	Data transformation: Please refer to http://www.biomedcentral.com/1471-2164/7/142 Try different combinations; such as log2transform=TRUE, znormtransfrom=FALSE or log2transform=FALSE, znormtransfrom=TRUE
medcenter	Median centering of metabolites
znormtransform	Auto scaling; each metabolite will have a mean of 0 and unit variance
quantile_norm	Performs quantile normalization. Normalization options: Please set only one of the options to be TRUE
lowess_norm	Performs lowess normalization. Normalization options: Please set only one of the options to be TRUE
madscaling	Performs median adjusted scale normalization. Normalization options: Please set only one of the options to be TRUE

Value

Pre-processed data matrix.

Author(s)

Karan Uppal <kuppal2@emory.edu>

---

diffexp	*diffexp*

Description

This is a wrapper function for the data analysis workflow. The function performs data preprocessing (filtering, normalization, transformation), PCA, biomarker discovery, correlation-based network analysis, clustering analysis, and functional class scoring. The "featselmethod" allows users to select the method for selecting features for classification, regression, ANOVA, and time-series designs. The function evaluates the k-fold cross-validation accuracy using Support Vector Machine, performs hierarchical clustering analysis, PCA analysis (R2/Q2 diagnostics), and generates boxplots, bar plots, and line plots for discriminatory features.

Usage

```
function(Xmat=NA,Ymat=NA,feature_table_file,parentoutput_dir,
class_labels_file,num_replicates=3,summarize.replicates=TRUE,
summary.method="mean",summary.na.replacement="zeros",missing.val=0,
rep.max.missing.thresh=0.3,
 all.missing.thresh=0.1,group.missing.thresh=0.7,
input.intensity.scale="raw",
log2transform=FALSE,medcenter=FALSE,znormtransform=FALSE,
quantile_norm=FALSE,lowess_norm=FALSE,madscaling=FALSE,TIC_norm=FALSE,
rangescaling=FALSE,mstus=FALSE,paretoscaling=FALSE,sva_norm=FALSE,
eigenms_norm=FALSE,vsn_norm=FALSE,
normalization.method=c("log2transform"),rsd.filt.list=1,
pairedanalysis=FALSE,featselmethod=c("limma","pls"),fdrthresh=0.05,
```

```
fdrmethod="BH",cor.method="spearman",networktype="complete",
abs.cor.thresh=0.4,cor.fdrthresh=0.05,kfold=10,
pred.eval.method="BER",globalcor=TRUE,
target.metab.file=NA,target.mzmatch.diff=10,target.rtmatch.diff=NA,
max.cor.num=100,
numtrees=20000,analysismode="classification",net_node_colors=c("green","red"), net_legend=TRUE,
svm_kernel="radial",heatmap.col.opt="brewer.RdBu",
manhattanplot.col.opt=c("darkblue","red3"),boxplot.col.opt=c("journal"),
barplot.col.opt=c("journal"),sample.col.opt="journal",
lineplot.col.opt="journal",hca_type="two-way",pls_vip_thresh=2,
num_nodes=2,max_varsel=100,
pls_ncomp=5,pca.stage2.eval=FALSE,scoreplot_legend=TRUE,pca.global.eval=TRUE,
rocfeatlist=seq(2,6,1),rocfeatincrement=TRUE,rocclassifier="svm",
foldchangethresh=1,wgcnarsdthresh=20,WGCNAmodules=FALSE,
optselect=TRUE,max_comp_sel=1,saveRda=FALSE,legendlocation="topleft",
pcacenter=TRUE,pcascale=TRUE,pca.cex.val=6,
pca.ellipse=FALSE,ellipse.conf.level=0.95,pls.permut.count=NA,
svm.acc.tolerance=5,limmadecideTests=TRUE,pls.vip.selection="max",
globalclustering=FALSE,plots.res=600,plots.width=8,plots.height=8,
plots.type="cairo",
output.device.type="pdf",pvalue.thresh=0.05,individualsampleplot.col.opt="journal",
pamr.threshold.select.max=FALSE,aggregation.method="RankAggreg",
aggregation.max.iter=1000,mars.gcv.thresh=1,error.bar=TRUE,cex.plots=1,
lme.modeltype="RI",
barplot.xaxis="Factor1",lineplot.lty.option=c("solid", "dashed", "dotted", "dotdash", "longdash",
alphabetical.order=FALSE,
kegg_species_code="hsa",database="pathway",reference_set=NA,metab_annot=NA,
add.pvalues=FALSE,add.jitter=FALSE,fcs.permutation.type=1,fcs.method="zscore",
fcs.min.hits=2,names_with_mz_time=NA,ylab_text="Abundance",xlab_text=NA,
boxplot.type="ggplot",samplermindex=NA,
differential.network.analysis = FALSE, degree.centrality.method = "hybrid.DEC", log2.transform.con
balance.classes = FALSE, balance.classes.sizefactor = 10, balance.classes.seed = 1, cv.perm.count =
multiple.figures.perpanel = FALSE, hca.labRow.value = FALSE, hca.labCol.value = FALSE, alpha.col = 
similarity.matrix = "correlation", outlier.method = c("pcout", "sumtukey", "pcatukey", "MDchisq"),
removeRda = TRUE, color.palette = c("journal", "npg", "nejm", "jco", "lancet", "custom1", "brewer.Rc
"brewer.RdBu", "brewer.PuOr", "brewer.PRGn", "brewer.PiYG", "brewer.BrBG", "brewer.Set2", "brewer.
"brewer.Dark2", "brewer.YlGnBu", "brewer.YlGn", "brewer.YlOrRd", "brewer.YlOrBr", "brewer.PuBuGn",
"brewer.PuRd", "brewer.PuBu", "brewer.OrRd", "brewer.GnBu", "brewer.BuPu", "brewer.BuGn", "brewer.
"black", "grey65", "terrain", "rainbow", "heat", "topo"), plot_DiNa_graph = FALSE,
limma.contrasts.type = c("contr.sum", "contr.treatment"), hca.cex.legend = 0.7, plot.boxplots.raw =
vcovHC.type = "HC3", ggplot.type1 = TRUE, facet.nrow = 1, pairwise.correlation.analysis = FALSE,
generate.boxplots = TRUE, pvalue.dist.plot = TRUE, ...)
```

Arguments

Xmat	R object for feature table. If this is given, then feature table can be set to NA.
Ymat	R object for response/class labels matrix. If this is given, then class can be set to NA.
feature_table_file	
	Feature table that includes the mz, retention time, and measured intensity in each sample for each analyte. The first 2 columns should be the mz and time. The remaining columns should correspond to the samples in the class labels file with

each column including the intensity profile of a sample. Full path required. Eg: C:/My Documents/test.txt The feature table should be in a tab-delimited format. An example of the input file is provided under the "example" folder.

parentoutput_dir

Provide full path of the folder where you want the results to be written. Eg: C:/My Documents/ProjectA/results/

class_labels_file

File with class labels information for each sample. Samples should be in the same order as in the feature table. Please use the same format as in the example folder. If you want to adjust for covariates in "lmreg" option, then you can add additional columns, one per covariate. Categorical variables should be strings (eg: "male", "female"). Please see "classlabels_gender.txt" file as an example.

num_replicates   Number of technical replicates

feat.filt.thresh

Percent Intensity Difference or Coefficient of variation threshold; feature filtering Use NA to skip this step.

summarize.replicates

Do the technical replicates per sample need to be averaged or median summarized?

summary.method   Method for summarizing the replicates. Options: "mean" or "median"

summary.na.replacement

How should the missing values be represented? Options: "zeros", "halffeaturemin", "halfsamplemin","halfdatamin", "none" "zeros": replaces missing values by 0 "halfsamplemin": replaces missing value by one-half of the lowest signal intensity in the corresponding sample "halfdatamin": replaces missing value by one-half of the lowest signal intensity in the complete dataset "halffeaturemin": replaces missing value by one-half of the lowest signal intensity for the current feature "none": keeps missing values as NAs

Users are recommended to perform imputation prior to performing biomarker discovery.

missing.val   How are the missing values represented in the input data? Options: "0" or "NA"

rep.max.missing.thresh

What propotion of replicates are allowed to have missing values during the averaging or median summarization step of each biological sample? If the number of replicates with missing values is greater than the defined threshold, then the summarized value is represented by the "missing.val" parameter. If the number of replicates with missing values is less than or equal to the defined threshold, then the summarized value is equal to the mean or the median of the non-missing values. Default: 0.5

all.missing.thresh

What propotion of total number of samples should have an intensity? Default: 0.5

input.intensity.scale

Are the intensities in the input feature table at raw scale or log2 scale? eg: "raw" or "log2" Default: "raw"

group.missing.thresh

What propotion of samples in either of the two groups should have an intensity? If at least x for further analysis. Default: 0.7

normalization.method

Data transformation and normalization method. Options:

"log2transform": log2 transformation "log2quantilenorm": log2 transformation and quantile normalization "znormtransform": auto scaling; each variable will have a mean of 0 and unit variance "quantile_norm": Performs quantile normalization "lowess_norm": Performs lowess normalization "rangescaling": Performs range scaling; scale by the min and max range "paretoscaling": Performs Pareto scaling; scale by the square root of the standard deviation "mstus": MS Total Useful Signal (MSTUS) normalization "sva": Surrogate Variable Analysis (SVA) normalization "eigenms_norm": EigenMS normalization "vsn_norm": variance stabilizing normalization "tic_norm": totial intensity normalization "cubicspline_norm": Cubic spline normalization "mad_norm": Median absolute deviation normalization

rsd.filt.list	This parameter allows to perform feature filtering based on overall variance (across all samples) prior to performing hypothesis testing. Eg: seq(0,30,5).
pairedanalysis	Is this a paired-study design? TRUE or FALSE If samples are paired, then the feature table and the class labels file should be organized so that the paired samples are arranged in the same order in each group. For example, the first sample in group A and the first sample in group B should be paired.
featselmethod	Options: "limma": for one-way ANOVA using LIMMA (mode=classification) "limma2way": for two-way ANOVA using LIMMA (mode=classification) "limma1wayrepeat": for one-way ANOVA repeated measures using LIMMA (mode=classification) "limma2wayrepeat": for two-way ANOVA repeated measures using LIMMA (mode=classification) "lm1wayanova": for one-way ANOVA using linear model (mode=classification) "lm2wayanova": for two-way ANOVA using linear model (mode=classification) "lm1wayanovarepeat": for one-way ANOVA repeated measures using linear model (mode=classification) "lm2wayanovarepeat": for two-way ANOVA repeated measures using linear model (mode=classification) "lmreg": variable selection based on p-values calculated using a linear regression model; allows adjustment for covariates (mode= regression or classification) "logitreg": variable selection based on p-values calculated using a logistic regression model; allows adjustment for covariates (mode= classification) "rfesvm": uses recursive feature elimination SVM algorithm for variable selection; (mode=classification) "RF": for random forest based feature selection (mode= regression or classification) "RFconditional": for conditional random forest based feature selection (mode= regression or classification) "pamr": for prediction analysis for microarrays algorithm based on the nearest shrunken centroid method (mode= classification) "MARS": for multiple adaptive regression splines (MARS) based feature selection (mode= regression or classification) "pls": for partial least squares (PLS) based feature selection (mode= regression or classification) "spls": for sparse partial least squares (PLS) based feature selection (mode= regression or classification) "spls1wayrepeat": for sparse partial least squares (PLS) based feature selection for one-way repeated measures (mode= regression or classification) "spls2wayrepeat": for sparse partial least squares (PLS) based feature selection for two-way repeated measures (mode= regression or classification) "o1pls": for orthogonal partial least squares (OPLS) based feature selection (mode= regression or classification)
pvalue.thresh	p-value threshold. Eg: 0.05^M
fdrthresh	False discovery rate threshold. Eg: 0.05
fdrmethod	Options: "BH", "ST", "Strimmer", "none" "BH": Benjamini-Hochberg (1995) (Default: more conservative than "ST" and "Strimmer") "ST": Storey & Tibshirani (Storey 2001, PNAS) algorithm implemented in the qvalue package "Strimmer": (Strimmer 2008, Bioinformatics) algorithm implemented in the fdrtool

	package "none": No FDR correction will be performed. fdrthresh will be treated as raw p-value cutoff
cor.method	Correlation method. Options: "pearson" or "spearman". Default: "spearman"
networktype	Options: "complete" or "GGM" "complete": performs network analysis using ordinary Pearson or Spearman correlation statistic "GGM": generates network based on partial correlation analysis using the GeneNet package
abs.cor.thresh	Absolute Pearson or Spearman correlation coefficient for network analysis. Default: 0.4
cor.fdrthresh	False discovery rate threshold for correlation analysis. Default: 0.05
kfold	Number of subsets in which the data should be divided for cross-validation. If kfold=10, then the data set will be divided into 10 subsets of size (N/10), where N is the total number of samples. 9 subsets are used for training and the remaining 1 is used for testing. This process is repeated 10 times and the CV-accuracy would be the mean of the classification accuracy from the 10 iterations. The same will be true for any other value of k. WARNING: The kfold value should be less than or equal to the total number of samples.
pred.eval.method	
	Criteria for evaluating the performance of the model. CV: Overall Cross-validation classification accuracy, balanced error rate (BER): (sum of accuracy in each class)/(number of classes) area under the curve (AUC) Eg: "CV", "BER", or "AUC". Default: "BER"
globalcor	Perform correlation analysis between selected features and all other features? Options: TRUE or FALSE
target.metab.file	
	File that includes the mz and/or retention time of the targeted metabolites. See example.
target.mzmatch.diff	
	+/- ppm mass tolerance for searching the target m/z in the current feature table
target.rtmatch.diff	
	+/- retention time tolerance for searching the target m/z in the current feature table
max.cor.num	Maximum number of correlated metabolites to be included in the network figure. Default: 100
pcacenter	Data centering for PCA. Options: "TRUE" or "FALSE". Default=TRUE
pcascale	Data scaling for PCA. Options: "TRUE" or "FALSE". Default=TRUE
samplermindex	Column index of any additional or irrelevant columns to be deleted. Options: "NA" or list of column numbers. eg: c(1,3,4) Default=NA
numtrees	Number of trees to be used for random forest method. Default=500
analysismode	"classification" for group-wise comparison (case vs control) or "regression" for continuous response variables. Default: "classification"
net_node_colors	
	Colors of nodes in the correlation networks. Eg: c("pink", "skyblue"), or ("red","green")
net_legend	Should the network be displayed for the correlation network? eg: TRUE or FALSE
max_var	Max number of variables to be used for sPLS, rfesvm, and Random Forest? eg:150
svm_kernel	SVM kernel eg: "radial" or "linear"

rf_selmethod	Random forest VIP based selection method. If rankbased option is selected, variables are ranked based on the Variable Importance Measure. Only the top "max_varsel" variables are selected. If absVIMthresh is selected, then all features with VIM greater than the absolute value of the lowest VIM are selected. eg: "absVIMthresh" or "rankbased"
pls_vip_thresh	Threshold for VIP score from PLS/O1PLS. eg: 1
max_varsel	Maximum number of variables to keep if "rankbased" RF or spls is used. eg: 100
pls_ncomp	Maximum number of components to be considered during the PLS optimal number of components selection step. eg: 2
pca.stage2.eval	
	Should PCA diagnostics be performed in stage 2? eg: TRUE or FALSE
scoreplot_legend	
	Should legends be included in score plots? eg: TRUE or FALSE
pca.global.eval	
	Should global PCA evaluation be performed? Default:TRUE eg: TRUE or FALSE
rocfeatlist	Vector indicating number of features to be used for ROC evaluation: eg: c(2,4,6) will generate ROC for top 2, top 4, and top 6 feautres. Default: seq(2,10,1)
rocclassifier	Classifier to be used for ROC evaluation. Options: "svm" or "logitreg". Default: "svm"
foldchangethresh	
	Secondary feature selection criteria based on fold change threshold. This is performed after statistical significance or importance evaluation.
wgcnarsdthresh	Relative standard deviation or coefficient of variation (across all samples) based filtering threshold before performing WGCNA module preservation analysis. Default: 20
WGCNAmodules	Perform WGCNA module preservation analysis. TRUE or FALSE Default: TRUE
optselect	Determine optimal number of PLS components. Default: TRUE
max_comp_sel	Number of PLS components to use for VIP or sparse loading selection (sPLS). Default=1
saveRda	Should the results be saved in a binary R object. Default: TRUE
legendlocation	Legend location for PLS or PCA plots
pca.cex.val	Size of points on PCA plots. eg: 4
pca.ellipse	Should ellipse be plotted on PCA plots? eg: TRUE or FALSE
ellipse.conf.level	
	Confidence interval for PCA ellipses eg: 0.95
pls.permut.count	
	Number of permutations for calculating p-values for PLS or sPLS models. eg: 1000
svm.acc.tolerance	
	Stopping criteria for forward feature selection using "rfeSVM" method. If the difference between best accuracy and current accuracy based on the newly added feature drops below the tolerance level, the forward selection process is terminated. eg: 5

pamr.threshold.select.max

        If two or more thresholds for shrinking the d statistic in the PAM algorithm (Tibshirani et al. Statistical Science 2003) have equal accuracy, should the maximum value (lowest number of features) be used? Default: FALSE

aggregation.method

        Method for combining the results from mutliple feature selection methods Options: Consensus: will only keep features that are selected in all methods RankAggreg: will use the cross entropy algorithm with Spearman footrule distance as the distance measure (RankAggreg; Pihur et al. BMC Bioinformatics 2009)

aggregation.max.iter

        Maximum number of iterations used in the rank aggregation step. Default: 1000

mars.gcv.thresh

        Minimum generalized cross-validation value (range: 0 to 100) for a feature to be selected. Default: 10

limmadecideTests

        Perform decide tests for LIMMA to perform multiple testing and assign up, down, or not significant. TRUE or FALSE.

pls.vip.selection

        How to summarize VIP across multiple PLS components? Options: "max" to take the maximum VIP across all selected components or "mean" to take the average VIP across all selected components. Default: "max"

globalclustering

        Perform global clustering using all features based on EM and hierachical clustering analysis. TRUE or FALSE. Default:FALSE

plots.res   Resolution of PNG files. Default: 600

plots.width   Width dimension for PNG files. Default: 8

plots.height   Height dimension for PNG files. Default: 8

output.device.type

        Output device: "png" or "pdf" Default:"pdf"

heatmap.col.opt

        Color scheme for heatmaps:

        Default: "brewer.RdBu"

error.bar   Plot error bars. TRUE or FALSE. Default: TRUE

cex.plots   Relative factor to change font size of text in plots e.g.: 0.8 or 2 Default: 1

lme.modeltype   Options for mixed-effects models: RI:Random intercept RIRS: random intercept and random slope models Default: "RI"

barplot.xaxis   Label for x-axis in barplots Default: "Factor1"

lineplot.lty.option

        Default: c("solid", "dashed", "dotted", "dotdash", "longdash", "twodash")

timeseries.lineplots

        Generate lineplots showing longitudinal pattern: TRUE or FALSE Default: FALSE

alphabetical.order

        Arrange class labels in alphabetical order versus arranging them based on which class appears first in the class labels file. TRUE or FALSE Default: TRUE

boxplot.type   Type of boxplots: "simple" using the boxplot() function in R or "ggplot" for ggboxplot and geom_boxplot functions

add.pvalues   Add p-values in boxplots: TRUE or FALSE Default: FALSE

add.jitter	Add jitter in boxplots: TRUE or FALSE Default: FALSE
ylab_text	Y-axis label in barplots, boxplots, and lineplots Default: "Abundance"
differential.network.analysis	
	Compare network centrality of each variable between the groups (classes). e.g. TRUE or FALSE
degree.centrality.method	
	Method used for calculating degree centraliy. e.g. "eigenvector", "betweenness", "hybrid.DEC","hybrid.DBC"
	hybrid.DEC uses local connectivity and eigenvector centrality
	hybrid.DBC uses local connectivity and betweenness centrality
log2.transform.constant	
	Small constant to add to values for log2 transformation to avoid getting undefined values when calculating log2 of 0 e.g. 1
balance.classes	
	Balance classes using the ROSE package for cases when the classes are imbalanced. Only useful for binary or two-class comparisons. e.g. FALSE or TRUE
balance.classes.sizefactor	
	Multiplicative factor (balance.classes.sizefactor x original N) for defining the number of samples in the resulting dataset. e.g. 10
balance.classes.seed	
	Random number for the ROSE method e.g. 999
cv.perm.count	Number of iterations for calculating the permuted k-fold CV accuracy e.g. 100 or 1000
	Set to NA to turn off this step
multiple.figures.perpanel	
	Generate multiple figures per panel in box plots e.g. FALSE or TRUE
hca.labRow.value	
	Show variable (row) names in hierarchical clustering analysis heatmaps e.g. TRUE or FALSE
hca.labCol.value	
	Show sample (column) names in hierarchical clustering analysis heatmaps e.g. TRUE or FALSE
alpha.col	Numeric valute to adjust color transparency e.g. 1 or 0.8
similarity.matrix	
	Similarity matrix to use in hierarchical clustering analysis e.g. "correlation" for Pearson or Spearman or "TOM" for topological overlap matrix as used in WGCNA
outlier.method	Method for outlier detection for samples e.g. "pcout" for the pcout function in the mvoutlier package "MDchisq" for using Mahalanobis distance with a Chi-square test "pcatukey" for using PCA scores with Tukey's +/- 1.5 interquartile range (IQR) rule "sumtukey" for applying Tukey's +/- 1.5 IQR rule on summed abundance or intensities across all variables for each sample
removeRda	Remove intermediate .Rda files e.g. TRUE or FALSE
color.palette	Color theme for plots. default: "journal" Options: 1. "journal": color-blind friendly palette 2. built-in R color palettes: "rainbow", "terrain", "heat","topo" 3. RColorBrewer pallettes: "brewer.YlOrRd", "brewer.Purples","brewer.YlGn", "brewer.BuPu","brewer.BuGn","brewer.GnBu", "brewer.YlGnBu", "brewer.RdBu",

"brewer.RdYlBu","brewer.PuOr","brewer.PRGn" (color codes: Yl-yellow; Rd-red, Bu-blue, Or-orange, Gn-green, PR-purple) 4. Generate a custom palette by providing colors (e.g. c("orange","blue","green"))

Please the color_palettes_xmsPANDA.pdf file on the GitHub page under xmsPANDA/inst

plot_DiNa_graph

Show network plots generated using differential network analysis e.g. TRUE or FALSE

limma.contrasts.type

Contrasts method for LIMMA e.g. "contr.sum" for ANOVA like sum contrasts method "contr.treatment" to treat the first group as the reference group and all other groups are compared to the reference

hca.cex.legend Numeric value indicating the amount by which plotting text and symbols should be scaled relative to the default. e.g 0.7

Set to NA to hide the HCA legend

plot.boxplots.raw

Should the box plots using untransformed (raw) data be generated? e.g. TRUE or FALSE

vcovHC.type Heteroscedasticity-consistent covariance matrix estimation type Please see the type argument in the vcovHC function in R package sandwich for more details e.g. "HC3", "HC0"

ggplot.type1 Should the boxplots be sub-grouped (faceted) by Factor2? Applicable for two-factor study designs. e.g. TRUE or FALSE

facet.nrow Number of rows in box plots if using ggplots for two-factor study designs. e.g. 1

pairwise.correlation.analysis

Should the pairwise correlation analysis between selected variables be performed? e.g. TRUE or FALSE

Note that this is only for the variables meeting the feature selection criteria.

generate.boxplots

Should the boxplots be generated? e.g. TRUE or FALSE

pvalue.dist.plot

Should the p-value distribution plots be generated? e.g. TRUE or FALSE

Details

This function performs data transformation, normalization, FDR analysis using LIMMA, variable selection using random forests, evaluates the predictive accuracy of the FDR significant features using k-fold cross-validation with a Support Vector Machine classifier, performs two-way hierarchical clustering analysis, and principal component analysis. An optimizaiton scheme is used to select the best set of features from different log2 fold change filtering thresholds. Finally, metabolome-wide and targeted correlation based network analysis of the FDR significant features is performed.

Value

diffexp_metabs   Best set of discriminatory features.

all_metabs   Results for all features.

mw.an.fdr   Metabolome-wide significant correlation network of differentially expressed metabolites.

targeted.an.fdr

        Correlation network of differentially expressed metabolites with targeted metabolites.

Following files are generated in the parent output location: Manhattan plots: showing metabolome wide p-values; Heatmap from Two-way hierarchical clustering analysis; Pairwise score plots from Principal Component Analysis; PCA score distribution plots; ROC plots; List of differentially expressed metabolites; Boxplots of differentially expressed metabolites; Correlation network figure and matrix; Pairwise correlation matrix CIRCOS format ready to be uploaded to: http://mkweb.bcgsc.ca/tableviewer/visu Or uploaded to Cytoscape gml format

Author(s)

Karan Uppal <kuppal2@emory.edu>

---

diffexp.lite   *diffexp.lite*

Description

A "lite" version of the diffexp() function with key features for data preprocessing, feature selection, PCA/HCA, and correlation analysis

Usage

```
diffexp.lite(Xmat = NA, Ymat = NA, outloc = NA, summary.na.replacement = "halffeaturemin", missing.v
```

Arguments

Xmat	R object for feature table. If this is given, then feature table can be set to NA.
Ymat	R object for response/class labels matrix. If this is given, then class can be set to NA.
outloc	Provide full path of the folder where you want the results to be written. Eg: C:/My Documents/ProjectA/results/

summary.na.replacement

        How should the missing values be represented? Options: "zeros", "halffeaturemin", "halfsamplemin","halfdatamin", "none" "zeros": replaces missing values by 0 "halfsamplemin": replaces missing value by one-half of the lowest signal intensity in the corresponding sample "halfdatamin": replaces missing value by one-half of the lowest signal intensity in the complete dataset "halffeaturemin": replaces missing value by one-half of the lowest signal intensity for the current feature "none": keeps missing values as NAs

        Users are recommended to perform imputation prior to performing biomarker discovery.

missing.val	How are the missing values represented in the input data? Options: "0" or "NA"

all.missing.thresh

        What propotion of total number of samples should have an intensity? Default: 0.5

group.missing.thresh

        What propotion of samples in either of the two groups should have an intensity? If at least x for further analysis. Default: 0.7

input.intensity.scale

> Are the intensities in the input feature table at raw scale or log2 scale? eg: "raw" or "log2" Default: "raw"

normalization.method

> Data transformation and normalization method. Options:
>
> "log2transform": log2 transformation "log2quantilenorm": log2 transformation and quantile normalization "znormtransform": auto scaling; each variable will have a mean of 0 and unit variance "quantile_norm": Performs quantile normalization "lowess_norm": Performs lowess normalization "rangescaling": Performs range scaling; scale by the min and max range "paretoscaling": Performs Pareto scaling; scale by the square root of the standard deviation "mstus": MS Total Useful Signal (MSTUS) normalization "sva": Surrogate Variable Analysis (SVA) normalization "eigenms_norm": EigenMS normalization "vsn_norm": variance stabilizing normalization "tic_norm": totial intensity normalization "cubicspline_norm": Cubic spline normalization "mad_norm": Median absolute deviation normalization

pairedanalysis   Is this a paired-study design? TRUE or FALSE If samples are paired, then the feature table and the class labels file should be organized so that the paired samples are arranged in the same order in each group. For example, the first sample in group A and the first sample in group B should be paired.

pca.global.eval

> Perform PCA using all variables? TRUE or FALSE

featselmethod   Options: "limma": for one-way ANOVA using LIMMA (mode=classification) "limma2way": for two-way ANOVA using LIMMA (mode=classification) "limma1wayrepeat": for one-way ANOVA repeated measures using LIMMA (mode=classification) "limma2wayrepeat": for two-way ANOVA repeated measures using LIMMA (mode=classification) "lm1wayanova": for one-way ANOVA using linear model (mode=classification) "lm2wayanova": for two-way ANOVA using linear model (mode=classification) "lm1wayanovarepeat": for one-way ANOVA repeated measures using linear model (mode=classification) "lm2wayanovarepeat": for two-way ANOVA repeated measures using linear model (mode=classification) "lmreg": variable selection based on p-values calculated using a linear regression model; allows adjustment for covariates (mode= regression or classification) "logitreg": variable selection based on p-values calculated using a logistic regression model; allows adjustment for covariates (mode= classification) "rfesvm": uses recursive feature elimination SVM algorithm for variable selection; (mode=classification) "RF": for random forest based feature selection (mode= regression or classification) "RFconditional": for conditional random forest based feature selection (mode= regression or classification) "pamr": for prediction analysis for microarrays algorithm based on the nearest shrunken centroid method (mode= classification) "MARS": for multiple adaptive regression splines (MARS) based feature selection (mode= regression or classification) "pls": for partial least squares (PLS) based feature selection (mode= regression or classification) "spls": for sparse partial least squares (PLS) based feature selection (mode= regression or classification) "spls1wayrepeat": for sparse partial least squares (PLS) based feature selection for one-way repeated measures (mode= regression or classification) "spls2wayrepeat": for sparse partial least squares (PLS) based feature selection for two-way repeated measures (mode= regression or classification) "o1pls": for orthogonal partial least squares (OPLS) based feature selection (mode= regression or classification)

pvalue.thresh   p-value threshold. Eg: 0.05^M

fdrthresh	False discovery rate threshold. Eg: 0.05
fdrmethod	Options: "BH", "ST", "Strimmer", "BY","none" "BH": Benjamini-Hochberg (1995) (Default: more conservative than "ST" and "Strimmer") "ST": Storey & Tibshirani (Storey 2001, PNAS) algorithm implemented in the qvalue package "Strimmer": (Strimmer 2008, Bioinformatics) algorithm implemented in the fdrtool package "none": No FDR correction will be performed. fdrthresh will be treated as raw p-value cutoff
foldchangethresh	
	Secondary feature selection criteria based on fold change threshold. This is performed after statistical significance or importance evaluation.
analysismode	"classification" for group-wise comparison (case vs control) or "regression" for continuous response variables. Default: "classification"
pls_vip_thresh	Threshold for VIP score from PLS/O1PLS. eg: 1
num_nodes	Number of CPU cores to use e.g. 2
optselect	Determine optimal number of PLS components. Default: TRUE
max_comp_sel	Number of PLS components to use for VIP or sparse loading selection (sPLS). Default=1
hca_type	"one-way" or "two-way" HCA
output.device.type	
	"pdf" or "png"
timeseries.lineplots	
	Generate lineplots showing longitudinal pattern: TRUE or FALSE Default: FALSE
alphabetical.order	
	Arrange class labels in alphabetical order versus arranging them based on which class appears first in the class labels file. TRUE or FALSE
ylab_text	Y-axis label in barplots, boxplots, and lineplots Default: "Abundance"
boxplot.type	Type of boxplots: "simple" using the boxplot() function in R or "ggplot" for ggboxplot and geom_boxplot functions
color.palette	Color theme for plots. default: "journal" Options: 1. "journal": color-blind friendly palette 2. built-in R color palettes: "rainbow", "terrain", "heat","topo" 3. RColorBrewer pallettes: "brewer.YlOrRd", "brewer.Purples","brewer.YlGn", "brewer.BuPu","brewer.BuGn","brewer.GnBu", "brewer.YlGnBu", "brewer.RdBu", "brewer.RdYlBu","brewer.PuOr","brewer.PRGn" (color codes: Yl-yellow; Rd-red, Bu-blue, Or-orange, Gn-green, PR-purple) 4. Generate a custom palette by providing colors (e.g. c("orange","blue","green"))
	Please the color_palettes_xmsPANDA.pdf file on the GitHub page under xmsPANDA/inst
generate.boxplots	
	Should the boxplots be generated? e.g. TRUE or FALSE
hca.cex.legend	Numeric value indicating the amount by which plotting text and symbols should be scaled relative to the default. e.g 0.7
	Set to NA to hide the HCA legend
lme.modeltype	Options for mixed-effects models: RI:Random intercept RIRS: random intercept and random slope models Default: "RI"
globalcor	Perform correlation analysis between selected features and all other features?
	Options: TRUE or FALSE
abs.cor.thresh	Absolute Pearson correlation coefficient for network analysis. Default: 0.4

`cor.fdrthresh`	False discovery rate threshold for correlation analysis. Default: 0.05
`net_legend`	Should the network be displayed for the correlation network? eg: TRUE or FALSE
`rocclassifier`	Set to NA to turn off k-fold CV classification accuracy and ROC analysis Default: "svm"

`limma.contrasts.type`

Contrasts method for LIMMA e.g. "contr.sum" for ANOVA like sum contrasts method "contr.treatment" to treat the first group as the reference group and all other groups are compared to the reference

`limmadecideTests`

Perform decide tests for LIMMA to perform multiple testing and assign up, down, or not significant. TRUE or FALSE.

`cex.plots`	Relative factor to change font size of text in plots e.g.: 0.8 or 2 Default: 1

`hca.labRow.value`

Show variable (row) names in hierarchical clustering analysis heatmaps e.g. TRUE or FALSE

`hca.labCol.value`

Show sample (column) names in hierarchical clustering analysis heatmaps e.g. TRUE or FALSE

Details

The "lite" version requires fewer computational resources. The function performs data transformation, normalization, feature selection, evaluates the predictive accuracy of the FDR significant features using k-fold cross-validation with a Support Vector Machine classifier, performs hierarchical clustering analysis, correlation analysis, and principal component analysis.

Value

`diffexp_metabs`	Best set of discriminatory features.
`all_metabs`	Results for all features.
`mw.an.fdr`	Metabolome-wide significant correlation network of differentially expressed metabolites.

`targeted.an.fdr`

Correlation network of differentially expressed metabolites with targeted metabolites.

Following files are generated in the parent output location: Manhattan plots: showing metabolome wide p-values; Heatmap from Two-way hierarchical clustering analysis; Pairwise score plots from Principal Component Analysis; PCA score distribution plots; ROC plots; List of differentially expressed metabolites; Boxplots of differentially expressed metabolites; Correlation network figure and matrix; Pairwise correlation matrix CIRCOS format ready to be uploaded to: http://mkweb.bcgsc.ca/tableviewer/visu Or uploaded to Cytoscape gml format

Author(s)

Karan Uppal <kuppal2@emory.edu>

---

do_wgcna   *do_wgcna*

Description

This function performs module preservation analysis using WGCNA.

Usage

```
do_wgcna(feature_table_file = NA, class_labels_file = NA, X = NA,
Y = NA, sigfeats = NA)
```

Arguments

feature_table_file

> Path and name of feature table that includes the mz, retention time, and mea-
> sured intensity in each sample for each analyte. The first 2 columns should be
> the mz and time. The remaining columns should correspond to the samples in
> the class labels file with each column including the intensity profile of a sample.
> Full path required. Eg: C:/My Documents/test.txt The feature table should be
> in a tab-delimited format. An example of the input file is provided under the
> "example" folder.

class_labels_file

> File with class labels information for each sample. Samples should be in the
> same order as in the feature table. Please use the same format as in the example
> folder. If you want to adjust for covariates in "lmreg" option, then you can add
> additional columns, one per covariate. Categorical variables should be strings
> (eg: "male", "female"). Please see "classlabels_gender.txt" file as an example.

X   R object for feature table. If this is given, then feature table can be set to NA.

Y   R object for response/class labels matrix. If this is given, then class labels file
             can be set to NA.

sigfeats   List of differentially expressed features. Default: NA

Details

This function calls WGCNA to perform module preservation analysis between different classes or
groups.

Value

PDF plots for module preservation from WGCNA and preservation matrix

Author(s)

Karan Uppal

References

WGCNA (Horvath 2007)

get.classification.accuracy

get.classification.accuracy

Description

This function performs classification accuracy analysis using the training and test sets. Users can choose support vector machine, logistic regression, random forest, and naive bayes as classifiers. The performance evaluation is determined based on the total classification rate, balanced accuracy rate, and AUC.

Usage

```
get.classification.accuracy(kfold, featuretable, classlabels,
kernelname = "radial", errortype = "AUC", conflevel = 95,
 classifier = "svm", seednum = 555,
 testfeaturetable = NA, testclasslabels = NA)
```

Arguments

kfold   Number of folds for cross-validation. e.g. 5 or 10

featuretable   R object for feature table with only differentially expressed features. This is the training set. The first two columns should be m/z and time.

classlabels   Class labels vector. e.g. c("case","control","case")

kernelname   Kernel for SVM: e.g. "radial" or "linear"

errortype   total: total classification accuracy rate; (number of correct classifications/total N) BAR: balanced accuracy rate; accounts of number of correct classification per class; BAR=$(1/m)*((C1/N1)+(C2/N2)+...+(Cm/Nm))$ where m is the number of classes, Cm is the number of correct classifications in class m, and Nm is the total number of subjects in class m. AUC: area under the curve

conflevel   Confidence level for k-fold classification accuracy e.g: 95

classifier   Classification algorithm to be used for ROC analysis. svm: Support Vector Machine logitreg: Logistic Regression rf: random forest naivebayes: naive bayes eg: "svm", "logitreg", "rf", "naivebayes"

seednum   Starting point used in the generation of a sequence of random numbers. e.g. 555

testfeaturetable

                   R object for test feature table with only differentially expressed features. This is the test set. The first two columns should be m/z and time. The order of features should be same as the training set.

testclasslabels

                   Class labels vector for samples in the test set.

Details

Function to evaluate classification. This function performs classification accuracy analysis using the training and test sets. Users can choose support vector machine, logistic regression, random forest, and naive bayes as classifiers. The performance evaluation is determined based on the total classification rate, balanced accuracy rate, and AUC.

Value

Classification accuracy in training and test sets

Author(s)

Karan Uppal; kuppal2@emory.edu

---

get_boxplots *get_boxplots*

Description

This function generates boxplots for m/z features. The input intensity matrix could be transformed or non-transformed intensities. Sample labels in the class labels file should be in the same order as the intensity matrix or feature table.

Usage

```
get_boxplots(feature_table_file, parentoutput_dir, class_labels_file, sample.col.opt = "rainbow",
```

Arguments

feature_table_file

Feature table that includes the mz, retention time, and measured intensity in each sample for each analyte. The first 2 columns should be the mz and time. The remaining columns should correspond to the samples in the class labels file with each column including the intensity profile of a sample. Full path required. Eg: C:/My Documents/test.txt The feature table should be in a tab-delimited format. An example of the input file is provided under the "example" folder.

parentoutput_dir

Provide full path of the folder where you want the results to be written. Eg: C:/My Documents/ProjectA/results/

class_labels_file

File with class labels information for each sample. Samples should be in the same order as in the feature table. Please use the same format as in the example folder.

sample.col.opt Color scheme for PCA and heatmap sample axis eg: "heat" or "topo"

alphacol=0.3 Color scaling parameter eg:0.3

Value

Creates a PDF with boxplots for each m/z feature.

Author(s)

Karan Uppal <kuppal2@emory.edu>

get_fcs   *get_fcs*

Description

Functional class scoring

Usage

```
get_fcs(target.data, target.data.annot = NA, kegg_species_code = "hsa",
database = c("pathway","module","brite","lipidmaps_mainclass","lipidmaps_subclass",
"refmet_superclass","refmet_mainclass","refmet_subclass","reactome_compound",
"reactome_atlas","kegg_atlas","custom"),
reference_set = NA, type.statistic = "pvalue", fcs.min.hits = 2, itrs = 100, numnodes = 2)
```

Arguments

target.data	Data frame with variable ID in column A and the statistic (eg. pvalue, t-statistic, fold change, VIP) in column B
target.data.annot	
	Optional argument with annotation data for the variables in target.data
kegg_species_code	
	KEGG species code if using KEGG as the reference database (e.g. hsa)
database	Options include: c("pathway","module","brite","lipidmaps_mainclass", "lipidmaps_subclass", "refmet_superclass","refmet_mainclass", "refmet_subclass","reactome_compound", "reactome_atlas","kegg_atlas","custom")
	The reactome_atlas and kegg_atlas options include all genes, compounds, and proteins associated with each pathway in the databases.
reference_set	If cutom database option is selected, then a data frame with the reference database should be provided with the functional class/pathway ID in column A, variable ID (e.g. compound name/ID) in column B, and set/pathway name in column C
type.statistic	Type of test statistic (e.g. pvalue, t-statistic,fold change)
fcs.min.hits	Minimum number of hits in a functional class (e.g. 2)
itrs	Number of permutations to generate the null distribution (e.g. 100)
numnodes	Number of CPUs to use (e.g. 2)

Details

The algorithm uses the z-statistic method proposed by Irizarry et al.(2009) and the max-mean method proposed by Efron and Tibshirani (2006) for determining significance of each functional class. The p-values from the two methods are aggregated using the Chi-square method.

Value

A data frame with aggregated statistic, p-value

Author(s)

Karan Uppal

---

get_hca *get_hca*

Description

This function performs two-way hierarchical clustering analysis and generates a heatmap showing the clustering results. The input intensity matrix could be transformed or non-transformed intensities. Sample labels in the class labels file should be in the same order as the intensity matrix or feature table.

Usage

```
get_hca(feature_table_file, parentoutput_dir, class_labels_file, heatmap.col.opt = "RdBu", cor.met
is.data.znorm = FALSE, analysismode = "classification", sample.col.opt = "rainbow", plots.width = 20
plots.height = 2000, plots.res = 300, alphacol = 0.3,hca_type="two-way",similarity.matrix="correlat
```

Arguments

feature_table_file

Feature table that includes the mz, retention time, and measured intensity in each sample for each analyte. The first 2 columns should be the mz and time. The remaining columns should correspond to the samples in the class labels file with each column including the intensity profile of a sample. Full path required. Eg: C:/My Documents/test.txt The feature table should be in a tab-delimited format. An example of the input file is provided under the "example" folder.

parentoutput_dir

Provide full path of the folder where you want the results to be written. Eg: C:/My Documents/ProjectA/results/

class_labels_file

File with class labels information for each sample. Samples should be in the same order as in the feature table. Please use the same format as in the example folder.

heatmap.col.opt

Color scheme for HCA hetmap eg: "RdBu", "topo", "heat", or "terrain"

cor.method Correlation method. Options: "person" or "spearman". Default: "spearman"

is.data.znorm Is the data already auto-scaled or z-scaled? eg: TRUE or FALSE

analysismode "classification" for group-wise comparison (case vs control) or "regression" for continuous response variables. Default: "classification"

sample.col.opt Color scheme for PCA and heatmap sample axis eg: "heat" or "topo" or "rainbow"

plots.width Width of the tiff file. eg: 2000

plots.height Height of the tiff file. eg: 2000

plots.res Resolution of the tiff file. eg: 300

alphacol Color scaling parameter eg:0.3

hcatype Color scaling parameter eg:"two-way" or "one-way"

similarity.matrix

Correlation or topological overlap matrix eg:"correlation" or "TOM"

Value

Heatmap from Two-way hierarchical clustering analysis; Intensity matrix in the same order as the dendrograms in heatmap; Sample cluster labels

Author(s)

Karan Uppal <kuppal2@emory.edu>

---

get_manhattanplots   *get_manhattanplots*

Description

Function to generate Manhattan plots.

Usage

```
get_manhattanplots(xvec, yvec, up_or_down, maintext = "", ythresh = 0.05,
y2thresh=NA, ylab, xlab, colorvec = c("darkgreen", "firebrick1"),
col_seq = c("brown", "chocolate3", "orange3", "coral", "pink", "skyblue",
"blue", "darkblue", "purple", "violet"), xincrement = 150, yincrement = 1)
```

Arguments

xvec	Vector with values for the x-axis. eg: m/z or retention time values
yvec	Vector with values for the y-axis. eg: (-)Log10 of p-values, VIP, loadings, regression coefficients, etc.
up_or_down	Vector indicating directionality of change. eg: Fold change values
maintext	Text for the plot title
ythresh	Y-axis threshold for significance or differential expression. eg: 3 for p=0.001; y=(-1)*log10(0.001) or 2 for VIP from PLS
y2thresh	Secondary Y-axis threshold for significance or differential expression. eg: 1.3 for p=0.05; y=(-1)*log10(0.05) or 1 for VIP from PLS
ylab	Y-axis label
xlab	X-axis label
colorvec	Vector of colors for representing up-regulation and down-regulation. eg: c("darkgreen", "firebrick1") In this case, features that are up-regulated in class A will have "darkgreen" color, and features that are up-regulated in class B will have "firebrick1" color.
col_seq	Vector of colors for plotting different segments of the x-axis
xincrement	Window size for breaking the x-axis into different segments for visualization purposes. eg: 150
yincrement	Window size for breaking the y-axis into different segments for visualization purposes. eg: 1

Details

This function can be used to generate Type 1 Manhattan plots: signficance vs m/z Type 2 Manhattan plots: significance vs retention time Type 3 Manhattan plots: significance vs intensity

Value

Manhattan plots

Note

#Example pdf("Manhattanplot.pdf") get_manhattanplots(...) #pass arguments dev.off()

Author(s)

Karan Uppal

---

get_pca   *Perfors PCA analysis*

Description

This function uses the pca function implemented in the mixOmics package for PCA analysis

Usage

```
get_pca(X, samplelabels, legendlocation = "topright", filename = NA,
ncomp = 5, center = TRUE, scale = TRUE, legendcex = 0.5,
outloc = getwd(), col_vec = NA, sample.col.opt = "default",
alphacol = 0.3, class_levels = NA, pca.cex.val = 3,
pca.ellipse = TRUE, ellipse.conf.level = 0.5, samplenames = FALSE)
```

Arguments

X	Data matrix without m/z and time.
samplelabels	Vector with class label for each sample.
legendlocation	Location of the legend on PCA score plots
filename	eg: "all", "signficantfeats"
ncomp	Number of components; please use ?pca for more information
center	Should the data be centered?; please use ?pca for more information
scale	Should the data be scaled?; please use ?pca for more information
legendcex	Size of the legend text in the PCA score plots. e.g.: 0.5 or 0.7
outloc	Output folder location
col_vec	Provide vector of colors for each group. eg: NA or c("red","green") for cases and controls, respectively. This argument is ignored if sample.col.opt is provided
sample.col.opt	Select R color palette. eg: "rainbow", "terrain", "topo". "heat", "default"
alphacol	Semi-transparent colors eg: 0.2
class_levels	Vector with names of different sample groups. eg: c("case", "control") or NA
pca.cex.val	Size of dots in PCA score plots. eg: 0.4
pca.ellipse	Should the score confidence interval for each group be drawn? eg: TRUE or FALSE
ellipse.conf.level	
	Confidence interval level eg: 0.95
samplenames	Should the sample names be included in PCA plots? eg: TRUE or FALSE

Details

This function performs PCA analysis. The results are saved in a RDA file.

Value

The function returns PCA results as an object and generates pairwise score plots for the first three components

Author(s)

Karan Uppal

References

mixOmics

---

get_pcascoredistplots   *get_pcascoredistplots*

Description

PCA score distribution (25th percentile, median, 75th percentile) plots

Usage

```
get_pcascoredistplots(X = NA, Y = NA, feature_table_file, parentoutput_dir,
class_labels_file, sample.col.opt = "rainbow", plots.width = 2000,
plots.height = 2000, plots.res = 300, alphacol = 0.3, col_vec,
pairedanalysis = FALSE, pca.cex.val = 3, legendlocation = "topright",
pca.ellipse = TRUE, ellipse.conf.level = 0.5, filename = "all")
```

Arguments

X   R object for feature table. If this is given, then feature table can be set to NA.

Y   R object for response/class labels matrix. If this is given, then class labels file can be set to NA.

feature_table_file

                        Path and name of feature table that includes the mz, retention time, and measured intensity in each sample for each analyte. The first 2 columns should be the mz and time. The remaining columns should correspond to the samples in the class labels file with each column including the intensity profile of a sample. Full path required. Eg: C:/My Documents/test.txt The feature table should be in a tab-delimited format. An example of the input file is provided under the "example" folder.

parentoutput_dir

                        Provide full path of the folder where you want the results to be written. Eg: C:/My Documents/ProjectA/results/

class_labels_file

> File with class labels information for each sample. Samples should be in the same order as in the feature table. Please use the same format as in the example folder. If you want to adjust for covariates in "lmreg" option, then you can add additional columns, one per covariate. Categorical variables should be strings (eg: "male", "female"). Please see "classlabels_gender.txt" file as an example.

sample.col.opt   Color scheme for PCA and heatmap sample axis eg: "rainbow", "heat" or "topo"

alphacol   Color scaling parameter eg:0.3

col_vec   Vector of colors for each sample.

pairedanalysis   Is this a paired-study design? TRUE or FALSE If samples are paired, then the feature table and the class labels file should be organized so that the paired samples are arranged in the same order in each group. For example, the first sample in group A and the first sample in group B should be paired.

pca.cex.val   Size of points on PCA plots. eg: 4

legendlocation   Legend location on PCA plots

pca.ellipse   Should ellipse be plotted on PCA plots? eg: TRUE or FALSE

ellipse.conf.level

> Confidence interval for PCA ellipses eg: 0.95

filename   Name of output PDF file

Details

This function performs PCA and generates pariwise score plots as well as score distribution plots (per group). It uses the Y vector and classlabels for color coding the samples in the pairwise score plots.

Value

The output includes: Pairwise PCA score plots, PCA score distribution plots, PCA scores and loadings text files.

Note

The plots can be sent to an external device by running the following commands: pdf("get_pcascoredistplots.pdf") get_pcascoredistplots(...) dev.off()

Author(s)

Karan Uppal

---

get_roc   *get_roc*

Description

This function generates Receiver Operating Characteristic curves using SVM and Logistic Regression as classifiers.

Usage

```
get_roc(dataA, classlabels, classifier = "svm", kname = "radial",
rocfeatlist = NA, rocfeatincrement = TRUE, testset = NA,
testclasslabels = NA, mainlabel = NA, col_lab = NA,
legend = TRUE, newdevice = FALSE, mz_names = NA,
svm.type = "nu-classification")
```

Arguments

dataA	R object for feature table with only differentially expressed features. This is the training set. The first two columns should be m/z and time.
classlabels	Class labels vector. e.g. c("case","control","case")
classifier	Classification algorithm to be used for ROC analysis. svm: Support Vector Machine logitreg: Logistic Regression eg: "svm" or "logitreg"
kname	Kernel for SVM. eg: "radial"
rocfeatlist	Vector indicating number of features to be used for ROC evaluation: eg: c(2,4,6) will generate ROC for top 2, top 4, and top 6 feautres. Default: seq(2,10,1)
rocfeatincrement	
	Turns on or off forward selection. eg: TRUE or FALSE
testset	R object for test feature table with only differentially expressed features. This is the test set. The first two columns should be m/z and time. The order of features should be same as the training set.
testclasslabels	
	Class labels vector for samples in the test set.
mainlabel	Main text label for the ROC plot. eg: "Group A vs B ROC curve"

Details

Function to perform ROC curve analysis using only traning set or using both training and test set.

Value

PDF file with ROC plot

Author(s)

Karan Uppal; kuppal2@emory.edu

---

get_volcanoplots   *get_volcanoplots*

Description

Function to generate volcano plots

Usage

```
get_volcanoplots(xvec, yvec, up_or_down, maintext = "", ythresh = 0.05,
 y2thresh = NA, ylab, xlab, colorvec = c("darkblue", "red3"),
col_seq = c("brown", "chocolate3", "orange3", "coral", "pink",
 "skyblue", "blue", "darkblue", "purple", "violet"),
xincrement = 1, yincrement = 1, xthresh = 1, pchvec = c(21, 21),
 background.points.col = "gray50", bad.feature.index = NA)
```

Arguments

xvec	Vector with log2 fold change values for the x-axis.
yvec	Vector with values for the y-axis. eg: (-)log10 of p-values, VIP, loadings, regression coefficients.
up_or_down	Same as xvec.
maintext	Text for the plot title
ythresh	Y-axis threshold for significance or differential expression. eg: 3 for p=0.001; y=(-1)*log10(0.001) or 2 for VIP from PLS
y2thresh	Optional secondary Y-axis threshold for significance or differential expression. eg: 1.3 for p=0.05; y=(-1)*log10(0.05) or 1 for VIP from PLS
ylab	Y-axis label
xlab	X-axis label
colorvec	Vector of colors for representing up-regulation and down-regulation. eg: c("darkgreen", "firebrick1") In this case, features that are up-regulated in class A will have "darkgreen" color, and features that are up-regulated in class B will have "firebrick1" color.
xincrement	Window size for breaking the x-axis into different segments for visualization purposes. eg: 150
yincrement	Window size for breaking the y-axis into different segments for visualization purposes. eg: 1
xthresh	Absolute value of the threshold for log2 fold change. e.g. 0, 1, 2

Details

This function generates volcano plots.

Value

Volcano plot

Author(s)

Karan Uppal; kuppal2@emory.edu

---

metabnet   *metabnet*

Description

Function for correlation (complete or partial) based metabolome-wide network analysis. Additionally, users have the option to provide a matrix of m/z features corresponding to chemicals of interest such as (phenylalanine, choline, etc) and/or a matrix of m/z features corresponding to discriminatory metabolites.

Usage

```
metabnet(feature_table_file, target.metab.file, sig.metab.file,
 class_labels_file = NA, parentoutput_dir, num_replicates = 1,
 cor.method = "spearman", abs.cor.thresh = 0.4, cor.fdrthresh = 0.05,
 target.mzmatch.diff = 10, target.rtmatch.diff = NA, max.cor.num = 100,
feat.filt.thresh = NA, summarize.replicates = TRUE, summary.method = "mean", all.missing.thresh = 0.
 medcenter = TRUE, znormtransform = FALSE, quantile_norm = TRUE,
 lowess_norm = FALSE, madscaling = FALSE, missing.val = 0,
networktype = "complete", samplermindex = NA, rep.max.missing.thresh = 0.3, summary.na.replacement
net_legend = FALSE, netrandseed = 555, TIC_norm = FALSE,
normalization.method = c("none"), input.intensity.scale = "raw")
```

Arguments

feature_table_file

> Feature table that includes the mz, retention time, and measured intensity in each sample for each analyte. The first 2 columns should be the mz and time. The remaining columns should correspond to the samples in the class labels file with each column including the intensity profile of a sample. Full path required. Eg: C:/My Documents/test.txt The feature table should be in a tab-delimited format. An example of the input file is provided under the "example" folder.

target.metab.file

> File that includes the mz and/or retention time of the targeted metabolites corresponding to pathways or chemicals of interest. See example.

sig.metab.file   File that includes the mz and/or retention time of the discriminatory metabolites. See example.

class_labels_file

> File with class labels information for each sample. Samples should be in the same order as in the feature table. Please use the same format as in the example folder.

parentoutput_dir

> Provide full path of the folder where you want the results to be written. Eg: C:/My Documents/ProjectA/results/

num_replicates   Number of technical replicates

cor.method   Correlation method. Options: "pearson" or "spearman". Default: "spearman"

abs.cor.thresh   Absolute Pearson correlation coefficient for network analysis. Eg: 0.5

cor.fdrthresh   False discovery rate threshold for correlation analysis. Eg: 0.05

target.mzmatch.diff

    +/- ppm mass tolerance for searching the target m/z in the current feature table

target.rtmatch.diff

    +/- retention time tolerance for searching the target m/z in the current feature table

max.cor.num   Maximum number of correlated metabolites to be included in the network figure. Default: 100

feat.filt.thresh

    Percent Intensity Difference or Coefficient of variation threshold; feature filtering Use NA to skip this step.

summarize.replicates

    Do the technical replicates per sample need to be averaged or median summarized?

summary.method   Method for summarizing the replicates. Options: "mean" or "median"

summary.na.replacement

    How should the missing values be represented? Options: "zeros", "halfsamplemin","halfdatamin", "none" "zeros": replaces missing values by 0 "halfsamplemin": replaces missing value by one-half of the lowest signal intensity in the corresponding sample "halfdatamin": replaces missing value by one-half of the lowest signal intensity in the complete dataset "none": keeps missing values as NAs

    Users are recommended to perform imputation prior to performing biomarker discovery.

all.missing.thresh

    What propotion of total number of samples should have an intensity? Default: 0.5

group.missing.thresh

    What propotion of samples in either of the two groups should have an intensity? If at least x for further analysis. Default: 0.7

log2transform   Data transformation: Please refer to http://www.biomedcentral.com/1471-2164/7/142 Try different combinations; such as log2transform=TRUE, znormtransfrom=FALSE or log2transform=FALSE, znormtransfrom=TRUE

medcenter   Median centering of metabolites

znormtransform   Auto scaling; each metabolite will have a mean of 0 and unit variance

quantile_norm   Performs quantile normalization. Normalization options: Please set only one of the options to be TRUE

lowess_norm   Performs lowess normalization. Normalization options: Please set only one of the options to be TRUE

madscaling   Performs median adjusted scale normalization. Normalization options: Please set only one of the options to be TRUE

missing.val   How are the missing values represented in the input data? Options: "0" or "NA"

networktype   Options: "complete" or "GGM" "complete": performs network analysis using ordinary Pearson or Spearman correlation statistic "GGM": generates network based on partial correlation analysis using the GeneNet package

samplermindex   Column index of any additional or irrelevant columns to be deleted. Options: "NA" or list of column numbers. eg: c(1,3,4) Default=NA

rep.max.missing.thresh

What propotion of replicates are allowed to have missing values during the averaging or median summarization step of each biological sample? If the number of replicates with missing values is greater than the defined threshold, then the summarized value is represented by the "missing.val" parameter. If the number of replicates with missing values is less than or equal to the defined threshold, then the summarized value is equal to the mean or the median of the non-missing values. Default: 0.5

net_node_colors

Colors of nodes in the correlation networks. Eg: c("pink", "skyblue"), or ("red","green")

net_legend   Should the network be displayed for the correlation network? eg: TRUE or FALSE

Details

Function for metabolomic network analysis

Value

Correlation matrix and network of metabolites.

Author(s)

Karan Uppal <kuppal2@emory.edu>

---

runApp.xmsPANDA   *runApp.xmsPANDA*

Description

Funtion to run the full xmsPANDA Shiny App

Usage

```
runApp.xmsPANDA()
```

Details

This function runs the Shiny App which includes capabilities for replicate summarization, data pre-processing, feature selection, PCA, HCA, classification accuracy evaluation, differential network analysis, module preservation analysis, global clustering analysis, functional class scoring, comparing normalization methods, and metabolite quantification analysis.

Author(s)

Karan Uppal (kuppal3gt@gmail.com)

runApp.xmsPANDALite	*runApp.xmsPANDALite*

Description

Funtion to run the Shiny App for the Lite version of xmsPANDA

Usage

```
runApp.xmsPANDALite()
```

Details

This function runs the Lite Shiny App which includes capabilities for data preprocessing, feature selection, network analysis, PCA, HCA, and classification accuracy evaluation.

Author(s)

Karan Uppal (kuppal3gt@gmail.com)

Index