

Package ‘statTarget’

November 9, 2024

Type Package

Title Statistical Analysis of Molecular Profiles

Version 1.36.0

Author Hemi Luan

Maintainer Hemi Luan <hemi.luan@gmail.com>

Depends R (>= 3.6.0)

Imports randomForest,plyr,pdist,ROC,utils,grDevices,graphics,rrcov,stats,
pls,impute

VignetteBuilder knitr

Suggests testthat, BiocStyle, knitr, rmarkdown

Description A streamlined tool provides a graphical user interface for quality control based signal drift correction (QC-RFSC), integration of data from multi-batch MS-based experiments, and the comprehensive statistical analysis in metabolomics and proteomics.

License LGPL (>= 3)

URL <https://stattarget.github.io>

biocViews ImmunoOncology, Metabolomics, Proteomics, Machine Learning, Lipidomics, MassSpectrometry, QualityControl, Normalization, QC-RFSC, ComBat, DifferentialExpression, BatchEffect, Visualization, MultipleComparison,Preprocessing, Software

RoxygenNote 7.2.1

LazyData true

NeedsCompilation no

BuildVignettes true

git_url <https://git.bioconductor.org/packages/statTarget>

git_branch RELEASE_3_20

git_last_commit 26706b6

git_last_commit_date 2024-10-29

Repository Bioconductor 3.20

Date/Publication 2024-11-08

Contents

| | |
|-------------------------------------|---|
| statTarget-package | 2 |
| mdsPlot | 3 |
| predict_RF | 3 |
| pvimPlot | 4 |
| rForest | 5 |
| shiftCor | 6 |
| shiftCor_dQC (deprecated) | 7 |
| statAnalysis | 8 |

| | |
|--------------|-----------|
| Index | 11 |
|--------------|-----------|

statTarget-package *Statistical Analysis of Molecular Profiles*

Description

An streamlined tool provides graphical user interface for quality control based signal correction, integration of MS-based data from multiple batches, and the comprehensive statistical analysis for omics studies.

Usage

```
statTarget()
```

Details

Package: statTarget

Type: package

License: LGPL (>= 3)

Value

A description of statTarget. See the details at <https://stattarget.github.io>

Author(s)

Hemi Luan

Maintainer: Hemi Luan hemi.luan@gmail.com

mdsPlot *MDSplot in statTarget*

Description

Multi-dimensional scaling plot of proximity matrix from randomForest.

Usage

```
mdsPlot(rForest,pimpModel,Labels = TRUE,slink = FALSE,
slinkDat, ...)
```

Arguments

| | |
|-----------|---|
| rForest | An object of class randomForest that contains the proximity component from statTarget_rForest function. |
| pimpModel | An object of permutation-based variable Gini importance measures (PIMP-algorithm) from statTarget_rForest function. |
| Labels | Labels is TRUE for visible the sample name in the figure else with the index for class. |
| slink | Logical indicating if slinkDat is active for extenal classID. |
| slinkDat | A data frame for the extenal classID. |
| ... | A generic MDSplot function in randomForest package |

Value

The output of cmdscale on $1 - rf\$proximity$ is returned invisibly.

Author(s)

Hemi Luan, hemi.luan@gmail.com

See Also

MDSplot

predict_RF *preidict function for random forest objects in statTarget*

Description

Prediction of test data using random forest in statTarget.

Usage

```
predict_RF(object, newdata, type='response',...)
```

Arguments

| | |
|---------|---|
| object | An object created by the function <code>statTarget_rForest</code> . |
| newdata | A data frame or matrix containing new data. (Note: If not given, the out-of-bag prediction in object is returned. see <code>randomForest</code> package.) |
| type | One of <code>response</code> , <code>prob.</code> or <code>votes</code> , indicating the type of output: predicted values, matrix of class probabilities, or matrix of vote counts. <code>class</code> is allowed, but automatically converted to <code>'response'</code> , for backward compatibility. |
| ... | A generic <code>predict</code> function from <code>randomForest</code> package. |

Value

A class of predicted values is returned. Object type is `classification`, for detail see `randomForest` package.

Author(s)

Hemi Luan, hemi.luan@gmail.com

See Also

`randomForest`

| | |
|----------|---|
| pvimPlot | <i>Gini importance and permutation-based variable importance measures plots</i> |
|----------|---|

Description

Create plots for Gini importance and permutation-based variable Gini importance measures.

Usage

```
pvimPlot(rForest, pimpModel, nvarRF = 6, border = NA,
         space = 0.3, ...)
```

Arguments

| | |
|-----------|---|
| rForest | an object of class <code>randomForest</code> that contains the proximity component from <code>statTarget_rForest</code> function. |
| pimpModel | an object of permutation-based variable Gini importance measures (PIMP-algorithm) from <code>statTarget_rForest</code> function. |
| nvarRF | The number of variables in importance plot of <code>randomForest</code> . |
| border | The color to be used for the border of the bars. Use <code>border = NA</code> to omit borders. see also <code>barplot</code> . |
| space | The amount of space (as a fraction of the average bar width) left before each bar. May be given as a single number or one number per bar. see also <code>barplot</code> |
| ... | A generic <code>barplot</code> function from <code>graphics</code> package. |

Value

The output of the name of selected variable importance.

Author(s)

Hemi Luan, hemi.luan@gmail.com

Examples

```
datpath <- system.file('extdata',package = 'statTarget')
statFile <- paste(datpath,'data_example_long.csv', sep='/')
getFile <- read.csv(statFile,header=TRUE)
rFtest <- rForest(getFile,ntree = 10,times = 5)
pvimPlot(rFtest$randomForest,rFtest$pimpTest)
```

rForest

Random Forest classification in statTarget

Description

rForest provides the Breiman's random forest algorithm for classification and permutation-based variable importance measures (PIMP-algorithm).

Usage

```
rForest(file,ntree = 100,times = 100, gDist = TRUE,
seed = 123,...)
```

Arguments

| | |
|-------|---|
| file | An data frame or 'Stat File' from statTarget software. |
| ntree | Number of trees to grow. This should not be set to too small a number, to ensure that every input row gets predicted at least a few times. |
| times | The number of permutations for permutation-based variable importance measures. |
| gDist | If gDist is TRUE the null importance distributions are approximated with Gaussian distributions else with empirical cumulative distributions. |
| seed | For the same set of random variables and reproducible results. |
| ... | A generic function in randomForest package |

Value

Objects Two objects from statTarget_rForest (1. randomForest,rfModel; 2. PIMPresult, pimp-Model)

VarImp The original Gini importance

PerVarImp A matrix, where the permuted VarImp measures for the predictor variable.

p-value The probability of observing the original VarImp or a larger value, given the fitted null importance distribution.

p.ks.test The p-values of the Kolmogorov-Smirnov Tests for each row PerVarImp.

Author(s)

Hemi Luan, hemi.luan@gmail.com

References

Altmann A., Tolosi L., Sander O. and Lengauer T. (2010) Permutation importance: a corrected feature importance measure, *Bioinformatics* 26 (10), 1340-1347.

Ender Celik. (2015) vita: Variable Importance Testing Approaches. R package version 1.0.0 <https://CRAN.R-project.org/package=vita>

Examples

```
datpath <- system.file('extdata', package = 'statTarget')
statFile <- paste(datpath, 'data_example_long.csv', sep='/')
getFile <- read.csv(statFile, header=TRUE)
rFtest <- rForest(getFile, ntree = 10, times = 5)
```

 shiftCor

shiftCor

Description

shiftCor provides the QC based signal correction for large scale metabolomics and targeted proteomics.

Usage

```
shiftCor(
  samPeno,
  samFile,
  Frule = 0.8,
  MLmethod = "QCRFSC",
  ntree = 500,
  imputeM = "KNN",
  coCV = 30,
  plot = FALSE
)
```

Arguments

| | |
|----------|---|
| samPeno | File path. The file with the meta information including the sample name, batches, class and order. |
| samFile | File path. The file with the expression information. |
| Frule | Modified n percent rule function. A variable will be kept if it has a non-zero value for at least n percent of samples in any one group. (Default: 0.8) |
| MLmethod | The machine learning method for QC based signal correction, such as QC based random forest signal correction (QC-RFSC). QC-RLSC was deprecated . |
| ntree | Number of trees to grow in random forest model. |

| | |
|---------|---|
| imputeM | The parameter for imputation method i.e., nearest neighbor averaging, 'KNN'; minimum values, 'min'; Half of minimum values, 'minHalf'; median values, 'median'. |
| coCV | Define the cutoff value (0-100) of CV for controlling the number of features. |
| plot | Defines if images of feature quality should be generated (TRUE) or not (FALSE). Defaults to FALSE. |

Value

the shiftCor files. See the details at <https://stattarget.github.io>

Examples

```
datpath <- system.file('extdata',package = 'statTarget')
samPeno <- paste(datpath,'MTBLS79_sampleList.csv', sep='/')
samFile <- paste(datpath,'MTBLS79.csv', sep='/')
samPeno
samFile
shiftCor(samPeno,samFile, MLmethod = 'QCRFSC', imputeM = 'KNN',coCV = 30)
```

shiftCor_dQC (deprecated)

QC-free based signal correction

Description

shiftCor_dQC provides the QC-free based signal correction for large scale mass spectrometry-based omics data.

Usage

```
shiftCor_dQC(
  samPeno,
  samFile,
  Frule = 0.8,
  imputeM = "KNN",
  MLmethod = "Combat",
  par.prior = TRUE,
  prior.plots = FALSE,
  mod.covariates = FALSE,
  batch.Num = NULL
)
```

Arguments

| | |
|---------|---|
| samPeno | The file with the meta information including the sample name, batches, class and order (denoting other covariates besides batch). |
| samFile | The file with the expression information. |
| Frule | Modified n percent rule function. A variable will be kept if it has a non-zero value for at least n percent of samples in any one group. (Default: 0.8) |

| | |
|----------------|--|
| imputeM | The parameter for imputation method i.e., nearest neighbor averaging, 'KNN'; minimum values, 'min'; Half of minimum values, 'minHalf'; median values, 'median'. |
| MLmethod | 'ComBat' allows users to adjust for batch effects in datasets where the batch covariate is known, using methodology described in Johnson et al. 2007. It uses either parametric or non-parametric empirical Bayes frameworks for adjusting data for batch effects. Users are returned an expression matrix that has been corrected for batch effects. The function was revised according to 'sva' package (version = "3.8"). |
| par.prior | TRUE indicates parametric adjustments will be used, FALSE indicates non-parametric adjustments will be used |
| prior.plots | (Optional) TRUE give prior plots. |
| mod.covariates | TRUE indicates model matrix for outcome of interest and other covariates besides batch (Column 'order' denotes covariates the in samPeno file). |
| batch.Num | (Optional) NULL If given, will use the selected batch as a reference for batch adjustment. |

Value

the shiftCor files. See the details at <https://stattarget.github.io>

Examples

```
datpath <- system.file('extdata', package = 'statTarget')
samPeno <- paste(datpath, 'MTBLS79_dQC_sampleList.csv', sep = '/')
samFile <- paste(datpath, 'MTBLS79.csv', sep = '/')
shiftCor_dQC(samPeno, samFile, Frule = 0.8, MLmethod = "Combat", mod.covariates = FALSE)
shiftCor_dQC(samPeno, samFile, Frule = 0.8, MLmethod = "Combat", mod.covariates = TRUE, batch.Num = 1)
```

statAnalysis

statAnalysis for statistical analysis for omics data or others.

Description

statAnalysis provides the statistical analysis for metabolomics data or others.

Usage

```
statAnalysis(
  file,
  Frule = 0.8,
  normM = "NONE",
  imputeM = "KNN",
  glog = TRUE,
  FDR = TRUE,
  ntree = 500,
  nvarRF = 5,
  scaling = "Pareto",
  plot.volcano = TRUE,
  save.boxplot = FALSE,
```



```

    silt = 20,
    pcax = 1,
    pcay = 2,
    Labels = TRUE,
    upper.lim = 2,
    lower.lim = 0.5,
    sig.lim = 0.05
)

```

Arguments

| | |
|--------------|---|
| file | The file with the expression information with long format |
| Frule | Modified n percent rule function. A variable will be kept if it has a non-zero value for at least n percent of samples in any one group. (Default: 0.8) |
| normM | The parameter for normalization method (i.e median quotient normalization, 'PQN'; integral normalization, 'SUM', and 'NONE'). |
| imputeM | The parameter for imputation method i.e., nearest neighbor averaging, 'KNN'; minimum values, 'min'; Half of minimum values, 'minHalf'; median values, 'median'. |
| glog | Generalised logarithm (glog) transformation, with the default value TRUE. The glog is a better behaved log transformation when some data values are zero or just near zero. |
| FDR | The false discovery rate for conceptualizing the rate of type I errors in null hypothesis testing when conducting multiple comparisons. |
| nree | Number of trees to grow for randomForest model. This should not be set to too small a number, to ensure that every input row gets predicted at least a few times. |
| nvarRF | The number of the variables with top importance in randomforest model |
| scaling | Scaling method before statistic analysis (PCA or PLS-DA). 'pareto', 'Pareto', 'p' or 'P' can be used for specifying the Pareto scaling. 'auto', 'Auto', 'auto', 'a' or 'A' can be used for specifying the Auto scaling (or unit variance scaling). 'vast', 'Vast', 'v' or 'V' can be used for specifying the vast scaling. 'range', 'Range', 'r' or 'R' can be used for specifying the Range scaling. |
| plot.volcano | if TRUE, the volcano plot is performed |
| save.boxplot | if TRUE, the box plot is performed |
| silt | The number of permutation for PLS-DA model and variable importance of randomForest. |
| pcax | Principal components in PCA model for the x-axis. |
| pcay | Principal components in PCA model for the y-axis. |
| Labels | Name labels for score plot of multiple statistical analysis |
| upper.lim | The up-regulated metabolites using Fold Changes cut off values in the Volcano plot. Fold change values will be calculated before normalization step. |
| lower.lim | The down-regulated metabolites using Fold Changes cut off values in the Volcano plot. Fold change values will be calculated before normalization step. |
| sig.lim | The significance level for metabolites in the Pvalues file in the Volcano plot. |

Value

The statAnalysis output files. See the details at <https://stattarget.github.io>

Author(s)

Hemi Luan, hemi.luan@gmail.com

Examples

```
datpath <- system.file('extdata',package = 'statTarget')
file <- paste(datpath,'data_example_long.csv', sep='/')
statAnalysis(file,Frule = 0.8, normM = 'NONE', imputeM = 'KNN', glog = TRUE,scaling = 'Pareto')
```

Index

* **Controls,Correction**

shiftCor, 6

shiftCor_dQC (deprecated), 7

* **P-value**

statAnalysis, 8

* **PCA**

statAnalysis, 8

* **PLSDA**

statAnalysis, 8

* **Quality**

shiftCor, 6

shiftCor_dQC (deprecated), 7

mdsPlot, 3

predict_RF, 3

pvimPlot, 4

rForest, 5

shiftCor, 6

shiftCor_dQC (shiftCor_dQC
(deprecated)), 7

shiftCor_dQC (deprecated), 7

statAnalysis, 8

statTarget (statTarget-package), 2

statTarget-package, 2