Package 'MSstatsTMTPTM'

February 22, 2021

Type Package

Title Post Translational Modification (PTM) Significance Analysis in shotgun mass spectrometry-based proteomic experiments with tandem mass tag (TMT) labeling

Version 1.0.2

Date 2020-09-28

Description Tools for Post Translational Modification (PTM) and protein significance analysis in shotgun mass spectrometry-based proteomic experiments with tandem mass tag (TMT) labeling. The functions in this package should be used after PTM/protein summarization. They can be used to both plot the summarized results and model the summarized datasets.

License Artistic-2.0

Depends R (>= 4.0)

Imports dplyr, gridExtra, stringr, reshape2, stats, utils, ggplot2, grDevices, graphics, MSstatsTMT, Rcpp

Suggests BiocStyle, knitr, rmarkdown, testthat, MSstats, covr

LinkingTo Rcpp

VignetteBuilder knitr

biocViews ImmunoOncology, MassSpectrometry, Proteomics, Software, DifferentialExpression, OneChannel, TwoChannel, Normalization, QualityControl

BugReports https://github.com/Vitek-Lab/MSstatsTMTPTM/issues

Encoding UTF-8

LazyData TRUE

Roxygen list(markdown = TRUE)

RoxygenNote 7.1.1

git_url https://git.bioconductor.org/packages/MSstatsTMTPTM

git_branch RELEASE_3_12

git_last_commit eb86988

git_last_commit_date 2021-02-15

Date/Publication 2021-02-21

Author Devon Kohler [aut, cre],
Ting Huang [aut],
Mateusz Staniak [aut],
Meena Choi [aut],
Tsung-Heng Tsai [aut],
Olga Vitek [aut]

Maintainer Devon Kohler <kohler.d@northeastern.edu>

R topics documented:

`dataProcessPlotsTMTPTM` *Visualization for explanatory data analysis - TMT experiment*

Description

To illustrate the quantitative data and quality control of MS runs, dataProcessPlotsTMT takes the quantitative data from MSstatsTMT converter functions as input and generate two types of figures in pdf files as output : (1) profile plot (specify "ProfilePlot" in option type), to identify the potential sources of variation for each protein; (2) quality control plot (specify "QCPlot" in option type), to evaluate the systematic bias between MS runs.

Usage

```
dataProcessPlotsTMTPTM(
  data.ptm,
  data.protein,
  data.ptm.summarization,
  data.protein.summarization,
  type,
  ylimUp = FALSE,
  ylimDown = FALSE,
  x.axis.size = 10,
  y.axis.size = 10,
  text.size = 4,
  text.angle = 90,
  legend.size = 7,
  dot.size.profile = 2,
```

```
  ncol.guide = 5,
  width = 10,
  height = 12,
  which.Protein = "all",
  originalPlot = TRUE,
  summaryPlot = TRUE,
  address = ""
)
```

Arguments

data.ptm	name of the data with PTM sites in protein name, which can be the output of MSstatsTMT converter functions.
data.protein	name of the data with peptide level, which can be the output of MSstatsTMT converter functions.
data.ptm.summarization	name of the data with ptm sites in protein-level name, which can be the output of the MSstatsTMT proteinSummarization function.
data.protein.summarization	name of the data with protein-level, which can be the output of the MSstatsTMT proteinSummarization function.
type	choice of visualization. "ProfilePlot" represents profile plot of log intensities across MS runs. "QCPlot" represents box plots of log intensities across channels and MS runs.
ylimUp	upper limit for y-axis in the log scale. FALSE(Default) for Profile Plot and QC Plot uses the upper limit as rounded off maximum of log2(intensities) after normalization + 3..
ylimDown	lower limit for y-axis in the log scale. FALSE(Default) for Profile Plot and QC Plot uses 0..
x.axis.size	size of x-axis labeling for "Run" and "channel in Profile Plot and QC Plot.
y.axis.size	size of y-axis labels. Default is 10.
text.size	size of labels represented each condition at the top of Profile plot and QC plot. Default is 4.
text.angle	angle of labels represented each condition at the top of Profile plot and QC plot. Default is 0.
legend.size	size of legend above Profile plot. Default is 7.
dot.size.profile	size of dots in Profile plot. Default is 2.
ncol.guide	number of columns for legends at the top of plot. Default is 5.
width	width of the saved pdf file. Default is 10.
height	height of the saved pdf file. Default is 10.
which.Protein	Protein list to draw plots. List can be names of Proteins or order numbers of Proteins. Default is "all", which generates all plots for each protein. For QC plot, "allonly" will generate one QC plot with all proteins.
originalPlot	TRUE(default) draws original profile plots, without normalization.
summaryPlot	TRUE(default) draws profile plots with protein summarization for each channel and MS run.

address — the name of folder that will store the results. Default folder is the current working directory. The other assigned folder has to be existed under the current working directory. An output pdf file is automatically created with the default name of "ProfilePlot.pdf" or "QCplot.pdf". The command address can help to specify where to store the file as well as how to modify the beginning of the file name. If address=FALSE, plot will be not saved as pdf file but showed in window.

Value

plot or pdf

Examples

```
data(raw.ptm)
data(raw.protein)
data(quant.msstats.ptm)
data(quant.msstats.protein)

## Profile plot
dataProcessPlotsTMTPTM(data.ptm=raw.ptm,
                  data.protein=raw.protein,
                  data.ptm.summarization=quant.msstats.ptm,
                  data.protein.summarization=quant.msstats.protein,
                  which.Protein = 1,
                  type='ProfilePlot',
                  address=FALSE)
```

---

example.contrast.matrix

Example contrast matrix for input into the groupComparisonTMTPTM function

---

Description

Manually specified comparisons of interest for contrast.matrix arguement of groupComparisonTMTPTM.

Usage

```
example.contrast.matrix
```

Format

A data frame with 9 rows and 6 variables.

Details

- Condition_1, ... Condition_6 : Column names are conditions in dataset
- 1-4, ... 5-6 : Row names are comparisons of interest

Examples

```
head(example.contrast.matrix)
```

groupComparisonTMTPTM *Model PTM and/or protein data and make adjustments if needed*

Description

Takes summarized PTM data from proteinSummarization and models with groupComparisonTMT. Can also take protein level data in the same format and model with groupComparisonTMT. Including protein data allows for adjusting PTM Fold Change by the change in protein abundance without modification.

Usage

```
groupComparisonTMTPTM(
  data.ptm,
  data.protein = NULL,
  contrast.matrix = "pairwise",
  moderated = FALSE,
  adj.method = "BH"
)
```

Arguments

data.ptm	Name of the output of the MSstatsTMT proteinSummarization function with PTM data. It should have columns named Protein, TechRepMixture, Mixture, Run, Channel, Condition, BioReplicate, Abundance.
data.protein	Protein dataset returned by the MSstatsTMT proteinSummarization function
contrast.matrix	Comparison between conditions of interests. 1) default is 'pairwise', which compare all possible pairs between two conditions. 2) Otherwise, users can specify the comparisons of interest. Based on the levels of conditions, specify 1 or -1 to the conditions of interests and 0 otherwise. The levels of conditions are sorted alphabetically.
moderated	TRUE will moderate t statistic; FALSE (default) uses ordinary t statistic.
adj.method	Adjusted method for multiple comparison. "BH" is default.

Value

A list models of all modeled and adjusted datasets

Examples

```
# Load summarized datasets from MSstatsTMT proteinSummarization function
data(quant.msstats.ptm)
data(quant.msstats.protein)

# Load specific contrast matrix
data(example.contrast.matrix)

model.results.contrast <- groupComparisonTMTPTM(data.ptm=quant.msstats.ptm,
                                    data.protein=quant.msstats.protein,
                                    contrast.matrix = example.contrast.matrix)
```

`model.results.contrast` *Ouput of groupComparisonTMTPTM for specific comparisons of interest*

Description

Returns the a list with three dataframes for three statistical models. One for each Protein, PTM, and PTM adjusted for protein level.

Usage

```
model.results.contrast
```

Format

A list of three dataframes

Details

- List objects: PTM.Model, Protein.Model, Adjusted.Model (all `dataframe`). Columns as follows:
- Protein : Protein ID
- Label: Label of the pairwise comparision or contrast
- log2FC: Log2 fold change
- SE: Standard error of the comparsion of contrast results
- DF: Degree of freedom
- pvalue: Value of p statistic of the test
- adj.pvalue: adjusted p value
- issue: used for indicating the reason why a comparison is not testable. NA means the comparison is testable. 'oneConditionMissing' means the protein has no measurements in one conndition of the comparison. Furtherone, when 'issue = oneConditionMissing', 'log2FC = Inf' means the negative condition (with coefficient -1 in the Label column) is missing and 'log2FC = -Inf' means the positive condition (with coefficient 1 in the Label column) is missing. completeMissing' means the protein has no measurements in all the conndtions of the comparison. unfittableModel' means there is no enough measurements to fit the linear model. In other words, each condition has only one measurement.

Examples

```
names(model.results.contrast)
head(model.results.contrast[[1]])
```

`model.results.pairwise` *Ouput of groupComparisonTMTPTM for full pairwise test*

Description

Returns the a list with three dataframes for three statistical models. One for each Protein, PTM, and PTM adjusted for protein level.

Usage

```
model.results.pairwise
```

Format

A list of three dataframes

Details

- List objects: PTM.Model, Protein.Model, Adjusted.Model (all `dataframe`). Columns as follows:
- Protein : Protein ID
- Label: Label of the pairwise comparision or contrast
- log2FC: Log2 fold change
- SE: Standard error of the comparsion of contrast results
- DF: Degree of freedom
- pvalue: Value of p statistic of the test
- adj.pvalue: adjusted p value
- issue: used for indicating the reason why a comparison is not testable. NA means the comparison is testable. 'oneConditionMissing' means the protein has no measurements in one conndition of the comparison. Furtherone, when 'issue = oneConditionMissing', 'log2FC = Inf' means the negative condition (with coefficient -1 in the Label column) is missing and 'log2FC = -Inf' means the positive condition (with coefficient 1 in the Label column) is missing. completeMissing' means the protein has no measurements in all the connditions of the comparison. unfittableModel' means there is no enough measurements to fit the linear model. In other words, each condition has only one measurement.

Examples

```
names(model.results.pairwise)
head(model.results.pairwise[[1]])
```

`MSstatsTMTPTM` *MSstatsTMTPTM: A package for detecting differencially abundant post translational modifications (PTM) in shotgun mass spectrometry-bsed proteomic experiments with tandem mass tag (TMT) labeling.*

Description

A set of tools for detecting differentially abundant PTMs and proteins in shotgun mass spectrometry-based proteomic experiments with tandem mass tag (TMT) labeling. The functions in this package should be used after PTM/protein summarization. They can be used to both plot the summarized results and model the summarized datasets.

functions

- `dataProcessPlotsTMTPTM` : Data visualization of PTM and global protein levels. Can plot either Profile plots to identify the potential sources of variation for each protein, or quality control plots to evaluate the systematic bias between MS runs.
- `groupComparisonTMTPTM` : Tests for significant changes in PTM abundance adjusted for global protein abundance across conditions based on a family of linear mixed-effects models in TMT experiment.

`quant.msstats.protein` *Example of output from proteinSummarizaiton function for Protein data*

Description

It is made from `raw.protein`. It is the output of proteinSummarization function from MSstatsTMT. It should include the required columns as below. The variables are as follows:

Usage

```
quant.msstats.protein
```

Format

A data frame with 93258 rows and 8 variables.

Details

- Run : MS run ID
- Protein : Protein ID
- Abundance: Protein-level summarized abundance
- Channel : Labeling information (126, ... 131)
- Condition : Condition (ex. Healthy, Cancer, Time0)
- BioReplicate : Unique ID for biological subject.
- TechRepMixture : Unique ID for technical replicate of one TMT mixture.
- Mixture : Unique ID for TMT mixture.

Examples

```
head(quant.msstats.protein)
```

`quant.msstats.ptm` *Example of output from proteinSummarizaiton function for PTM data*

Description

It is made from `raw.ptm`. It is the output of proteinSummarization function from MSstatsTMT. It should include the required columns as below. The variables are as follows:

Usage

```
quant.msstats.ptm
```

Format

A data frame with 19205 rows and 8 variables.

Details

- Run : MS run ID
- Protein : Protein ID with modification site mapped in. Ex. Protein_1002_S836
- Abundance: Protein-level summarized abundance
- Channel : Labeling information (126, ... 131)
- Condition : Condition (ex. Healthy, Cancer, Time0)
- BioReplicate : Unique ID for biological subject.
- TechRepMixture : Unique ID for technical replicate of one TMT mixture.
- Mixture : Unique ID for TMT mixture.

Examples

```
head(quant.msstats.ptm)
```

raw.protein *Example of input Protein dataset for TMT experiments.*

Description

It can be the output of PDtoMSstatsTMTFormat or other MSstatsTMT converter functions. It includes peak intensities for a variety of PTMs. This is the companion file to the raw.ptm dataset, includes unmodified protein data. The variables are as follows:

Usage

```
raw.protein
```

Format

A data frame with 620476 rows and 11 variables.

Details

- ProteinName : Name of protein
- PeptideSequence
- Charge
- PSM
- Mixture : Mixture of samples labeled with different TMT reagents, which can be analyzed in a single mass spectrometry experiment. If the channal doesn't have sample, please add Empty' under Condition. \item TechRepMixture : Technical replicate of one mixture. One mixture may have multiple technical replicates. For example, if TechRepMixture' = 1, 2 are the two technical replicates of one mixture, then they should match with same Mixture' value. \item Run : MS run ID. \item Channel : Labeling information (126, ... 131). \item Condition : Condition (ex. Healthy, Cancer, Time0) \item BioReplicate : Unique ID for biological subject. If the channal doesn't have sample, please add Empty' under BioReplicate.
- Intensity

Examples

```
head(raw.protein)
```

raw.ptm *Example of input PTM dataset for TMT experiments.*

Description

It can be the output of PDtoMSstatsTMTFormat or other MSstatsTMT converter functions. It includes peak intensities for a variety of PTMs. The variables are as follows:

Usage

```
raw.ptm
```

Format

A data frame with 24704 rows and 11 variables.

Details

- ProteinName : Name of protein with modification site mapped in with an underscore. ie "Protein_4_Y474"
- PeptideSequence
- Charge
- PSM
- Mixture : Mixture of samples labeled with different TMT reagents, which can be analyzed in a single mass spectrometry experiment. If the channal doesn't have sample, please add Empty' under Condition. \item TechRepMixture : Technical replicate of one mixture. One mixture may have multiple technical replicates. For example, if TechRepMixture' = 1, 2 are the two technical replicates of one mixture, then they should match with same Mixture' value. \item Run : MS run ID. \ite nel : Labeling information (126, ... 131). \item Condition : Condition (ex. Healthy, Cancer, Time0) \item BioReplicate : Unique ID for biological subject. If the channal doesn't have sample, please add Empty' under BioReplicate.
- Intensity

Examples

```
head(raw.ptm)
```

Index