Package 'MSstatsLOBD'

October 14, 2021

Type Package

Title Assay characterization: estimation of limit of blanc(LoB) and limit of detection(LOD)

Version 1.0.0

Date 2021-04-12

Description The MSstatsLOBD package allows calculation and visualization of limit of blac (LOB) and limit of detection (LOD). We define the LOB as the highest apparent concentration of a peptide expected when replicates of a blank sample containing no peptides are measured. The LOD is defined as the measured concentration value for which the probability of falsely claiming the absence of a peptide in the sample is 0.05, given a probability 0.05 of falsely claiming its presence. These functionalities were previously a part of the MSstats package. The methodology is described in Galitzine (2018) <doi:10.1074/mcp.RA117.000322>.

License Artistic-2.0

Depends R (>= 4.0)

Imports minpack.lm, ggplot2, utils, stats, grDevices

Suggests BiocStyle, knitr, rmarkdown, covr, tinytest, dplyr

LinkingTo Rcpp

VignetteBuilder knitr

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

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

Encoding UTF-8

LazyData TRUE

Roxygen list(markdown = TRUE)

RoxygenNote 7.1.1

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

git_branch RELEASE_3_13

git_last_commit b490f04

git_last_commit_date 2021-05-19

Date/Publication 2021-10-14

Author Devon Kohler [aut, cre],
Mateusz Staniak [aut],
Cyril Galitzine [aut],
Meena Choi [aut],
Olga Vitek [aut]

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

R topics documented:

`linear_quantlim` *Calculation of the LOB and LOD with a linear fit*

Description

This function calculates the value of the LOB (limit of blank) and LOD (limit of detection) from the (Concentration, Intensity) spiked in data. The function also returns the values of the linear curve fit that allows it to be plotted. At least 2 blank samples (characterized by Intensity = 0) are required by this function which are used to calculate the background noise. T he LOB is defined as the concentration at which the value of the linear fit is equal to the 95\ The LOD is the concentration at which the latter is equal to the 90\ A weighted linear fit is used with weights for every unique concentration proportional to the inverse of variance between replicates.

Usage

```
linear_quantlim(
  datain,
  alpha = 0.05,
  Npoints = 100,
  Nbootstrap = 500,
  num_changepoint_samples = 30,
  num_prediction_samples = 200,
  max_iter = 30
)
```

Arguments

datain	Data frame that contains the input data. The input data frame has to contain the following columns : CONCENTRATION, INTENSITY (both of which are measurements from the spiked in experiment) and NAME which designates the name of the assay (e.g. the name of the peptide or protein)
alpha	Probability level to estimate the LOB/LOD
Npoints	Number of points to use to discretize the concentration line between 0 and the maximum spiked concentration
Nbootstrap	Number of bootstrap samples to use to calculate the prediction interval of the fit. This number has to be increased for very low alpha values or whenever very accurate assay characterization is required.
num_changepoint_samples	Number of bootstrap samples for the prediction inteval for the changepoint. Large values can make calculations very expensive
num_prediction_samples	Number of prediction samples to generate
max_iter	Number of trials for convergence of every curvefit algorithm

Details

The LOB and LOD can only be calculated when more than 2 blank samples are included. The data should ideally be plotted using the companion function plot_quantlim to ensure that a linear fit is suited to the data.

Value

data.frame, It contains the following columns: i) CONCENTRATION: Concentration values at which the value of the fit is calculated ii) MEAN: The value of the curve fit iii) LOW: The value of the lower bound of the 95\ iv) UP: The value of the upper bound of the 95\ v) LOB: The value of the LOB (one column with identical values) vi) LOD: The value of the LOD (one column with identical values) vii) SLOPE: Value of the slope of the linear curve fit where only the spikes above LOD are considered viii) INTERCEPT: Value of the intercept of the linear curve fit where only the spikes above LOD are considered ix) NAME: The name of the assay (identical to that provided in the input) x) METHOD which is always set to LINEAR when this function is used. Each line of the data frame corresponds to a unique concentration value at which the value of the fit and prediction interval are evaluated. More unique concentrations values than in the input data frame are used to increase the accuracy of the LOB/D calculations.

Examples

```
# Consider data from a spiked-in contained in an example dataset. This dataset contains
# a significant threshold at low concentrations that is not well captured by a linear fit

head(spikeindata)

linear_quantlim_out <- linear_quantlim(spikeindata, Nbootstrap = 10)
```

`nonlinear_quantlim` *Calculation of the LOB and LOD with a nonlinear fit*

Description

This function calculates the value of the LOB (limit of blank) and LOD (limit of detection) from the (Concentration, Intensity) spiked in data. This function should be used instead of the linear function whenever a significant threshold is present at low concentrations. Such threshold is characterized by a signal that is dominated by noise where the mean intensity is constant and independent of concentration. The function also returns the values of the nonlinear curve fit that allows it to be plotted. At least 2 blank samples (characterized by Intensity = 0) are required by this function which are used to calculate the background noise. The LOB is defined as the concentration at which the value of the nonlinear fit is equal to the 95\ of the noise. The LOD is the concentration at which the latter is equal to the 90\ bound (5\ A weighted nonlinear fit is used with weights for every unique concentration proportional to the inverse of variance between replicates. The details behind the calculation of the nonlinear fit can be found in the Reference.

Usage

```
nonlinear_quantlim(
  datain,
  alpha = 0.05,
  Npoints = 100,
  Nbootstrap = 2000,
  num_changepoint_samples = 30,
  num_prediction_samples = 200,
  max_iter = 30
)
```

Arguments

`datain`	Data frame that contains the input data. The input data frame has to contain the following columns : CONCENTRATION, INTENSITY (both of which are measurements from the spiked in experiment) and NAME which designates the name of the assay (e.g. the name of the peptide or protein)
`alpha`	Probability level to estimate the LOB/LOD
`Npoints`	Number of points to use to discretize the concentration line between 0 and the maximum spiked concentration
`Nbootstrap`	Number of bootstrap samples to use to calculate the prediction interval of the fit. This number has to be increased for very low alpha values or whenever very accurate assay characterization is required.
`num_changepoint_samples`	Number of bootstrap samples for the prediction inteval for the changepoint. Large values can make calculations very expensive
`num_prediction_samples`	Number of prediction samples to generate
`max_iter`	Number of trials for convergence of every curvefit algorithm

Details

The LOB and LOD can only be calculated when more than 2 blank samples are included. The data should ideally be plotted using the companion function plot_quantlim to ensure that the fit is suited to the data.

Value

data.frame It contains the following columns: i) CONCENTRATION: Concentration values at which the value of the fit is calculated ii) MEAN: The value of the curve fit iii) LOW: The value of the lower bound of the 95\ iv) UP: The value of the upper bound of the 95\ v) LOB: The value of the LOB (one column with identical values) vi) LOD: The value of the LOD (one column with identical values) vii) SLOPE: Value of the slope of the linear curve fit where only the spikes above LOD are considered viii) INTERCEPT: Value of the intercept of the linear curve fit where only the spikes above LOD are considered ix) NAME: The name of the assay (identical to that provided in the input) x) METHOD which is always set to NONLINEAR when this function is used. Each line of the data frame corresponds to a unique concentration value at which the value of the fit and prediction interval are evaluated. More unique concentrations values than in the input data frame are used to increase the accuracy of the LOB/D calculations.

Examples

```
# Consider data from a spiked-in contained in an example dataset. This dataset contains
# a significant threshold at low concentrations that is not well captured by a linear fit

head(spikeindata)

nonlinear_quantlim_out <- nonlinear_quantlim(spikeindata, Nbootstrap = 10)
```

`plot_quantlim` *Plot results of nonlinear_quantlim() and linear_quantlim()*

Description

This function allows to plot the curve fit that is used to calculate the LOB and LOD with functions nonlinear_quantlim() and linear_quantlim(). The function outputs for each calibration curve, two pdf files each containg one plot. On the first, designated by _overall.pdf, the entire concentration range is plotted. On the second plot, designated by _zoom.pdf, the concentration range between 0 and xlim_plot (if specified in the argument of the function) is plotted. When no xlim_plot value is specified, the region close to LOB and LOD is automatically plotted.

Usage

```
plot_quantlim(
  spikeindata,
  quantlim_out,
  alpha,
```

```
  xlim_plot,
  width = 12,
  height = 4,
  address = ""
)
```

Arguments

spikeindata	Data frame that contains the experimental spiked in data. This data frame should be identical to that used as input by function functions nonlinear_quantlim() or linear_quantlim(). The data frame has to contain the following columns : CONCENTRATION, INTENSITY (both of which are measurements from the spiked in experiment) and NAME which designates the name of the assay (e.g. the name of the peptide or protein)
quantlim_out	Data frame that was output by functions nonlinear_quantlim() or linear_quantlim(). It has to contain at least the following columns: i) CONCENTRATION: Concentration values at which the value of the fit is calculated ii) MEAN: The value of the curve fit iii) LOW: The value of the lower bound of the 95\ the upper bound of the 95\ (one column with identical values) vi) LOD: The value of the LOD (one column with identical values) vii) NAME: The name of the assay (identical to that provided in the input) viii) METHOD which is LINEAR or NONLINEAR
alpha	Probability level to estimate the LOB/LOD
xlim_plot	Optional argument containing the maximum xaxis value of the zoom plot. When no value is specified, a suitable value close to LOD is automatically chosen.
width	width of the saved file. Default is 10.
height	height of the saved file. Default is 10.
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 "QuantLim.pdf" and "QuantLim_Zoom.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

list of two ggplot2 object

Examples

```
## Run LOBD analysis and plot
quant_out = nonlinear_quantlim(spikeindata, Nbootstrap = 10)
plot_quantlim(spikeindata = spikeindata, quantlim_out = quant_out,
              address = FALSE)
```

raw_data *Example of dataset that contains spike in data for 43 distinct peptides.*

Description

This example dataset is from CPTAC(Clinical Proteomic Tumor Analysis Consortium) assay portan (Thomas and others 2015). The dataset contains spike in data for 43 distinct peptides. For each peptide, 8 distinct concentration spikes for 3 different replicates are measured. The Skyline files for the assay along with details about the experiment can be obtained from this webpage: https://assays.cancer.gov. The particular dataset examined here (called JHU_DChan_HZhang_ZZhang) can be found at https://panoramaweb.org/labkey/project/CPTAC%20Assay%20Portal/JHU_DChan_HZhang_ZZhang/Serum_QExactive_GlycopeptideEnrichedPRM/begin.view?. It should be downloaded from the MSStats website http://msstats.org/?smd_process_download=1&download_id=548. The data is then exported in a csv file (calibration_data_raw.csv) from Skyline. The csv file contains the measured peak area for each fragment of each light and heavy version of each peptide. Depending on the format of the Skyline file and depending on whether standards were used, the particular outputs obtained in the csv file may vary. In this particular case the following variables are obtained in the output file calibration_data_raw.csv: File Name, Sample Name, Replicate Name, Protein Name, Peptide Sequence, Peptide Modified Sequence, Precursor Charge, Product Charge, Fragment Ion, Average Measured Retention Time, SampleGroup, IS Spike, Concentration, Replicate, light Area, heavy Area. A number of variables are byproducts of the acquisition process and will not be considered for the following, i.e. File Name, Sample Name, Replicate Name, SampleGroup, IS Spike. Variables that are important for the assay characterization are detailed below (others are assumed to be self explanatory):

Usage

```
raw_data
```

Format

A data frame with 3870 rows and 16 variables.

Details

- Pepdidesequence Name of the peptide sequence
- Concentration Value of the known spiked concentration in pmol.
- Replicate Number of the technical replicate
- light Area Peak area of the light (measured)
- heavy Area Peak area of the heavy (reference) peptide

Examples

```
head(raw_data)
```

spikeindata *Example of normalized datasets from raw_data,*

Description

We normalize the intensity of the light peptides using that of the heavy peptides. This corrects any systematic errors that can occur during a run or across replicates. The calculation is greatly simplified by the use of the tidyr and dplyr packages. The area from all the different peptide fragments is first summed then log transformed. The median intensity of the reference heavy peptides medianlog2heavy is calculated. Their intensities should ideally remain constant across runs since the spiked concentration of the heavy peptide is constant. The difference between the median for all the heavy peptide spikes is calculated. It is then used to correct (i.e. to normalize) the intensity of the light peptides log2light to obtain the adjusted intensity log2light_norm. The intensity is finally converted back to original space. Details are available in vignette. The variables are as follows:

Usage

```
spikeindata
```

Format

A data frame with 30 rows and 4 variables.

Details

- CONCENTRATION: Concentration values at which the value of the fit is calculated
- MEAN: The value of the curve fit
- LOW: The value of the lower bound of the 95\
- UP: The value of the upper bound of the 95\
- LOB: The value of the LOB (one column with identical values)
- LOD: The value of the LOD (one column with identical values)
- SLOPE: Value of the slope of the linear curve fit where only the spikes above LOD are considered
- INTERCEPT: Value of the intercept of the linear curve fit where only the spikes above LOD are considered
- NAME: The name of the assay (identical to that provided in the input)
- METHOD which is always set to NONLINEAR when this function is used.
- Each line of the data frame corresponds to a unique concentration value at which the value of the fit and prediction interval are evaluated.
- More unique concentrations values than in the input data frame are used to increase the accuracy of the LOB/D calculations.

Examples

```
head(spikeindata)
```

Index