Package ‘msmsTests’

March 23, 2017

Type Package

Title LC-MS/MS Differential Expression Tests

Version 1.12.0

Date 2013-10-02

Author Josep Gregori, Alex Sanchez, and Josep Villanueva

Maintainer Josep Gregori <josep.gregori@gmail.com>

Depends R (>= 3.0.1), MSnbase, msmsEDA

Imports edgeR, qvalue

Description Statistical tests for label-free LC-MS/MS data by spectral counts, to discover differentially expressed proteins between two biological conditions. Three tests are available: Poisson GLM regression, quasi-likelihood GLM regression, and the negative binomial of the edgeR package. The three models admit blocking factors to control for nuisance variables. To assure a good level of reproducibility a post-test filter is available, where we may set the minimum effect size considered biologically relevant, and the minimum expression of the most abundant condition.

License GPL-2

biocViews Software, MassSpectrometry, Proteomics

NeedsCompilation no

R topics documented:

msmsTests-package .................................................. 2
msms.edgeR ................................................................. 2
msms.glm.pois .............................................................. 4
msms.glm.qlll ............................................................... 6
msms.spk ................................................................. 7
pval.by.fc ................................................................. 8
res.volcanoplot .......................................................... 9
test.results ............................................................ 11

Index 14
msmsTests-package

**LC-MS/MS Differential Expression Tests**

**Description**

Statistical tests for label-free LC-MS/MS data by spectral counts, to discover differentially expressed proteins between two biological conditions. Three tests are available: Poisson GLM regression, quasi-likelihood GLM regression, and the negative binomial of the edgeR package. The three models admit blocking factors to control for nuisance variables. To assure a good level of reproducibility a post-test filter is available, where we may set the minimum effect size considered biologically relevant, and the minimum expression of the most abundant condition.

**Details**

- **Package**: msmsTests
- **Type**: Package
- **Version**: 0.99.1
- **Date**: 2013-07-26
- **License**: GPL-2

- **msms.glm.pois**: Poisson based GLM regression
- **msms.glm.qlll**: Quasi-likelihood GLM regression
- **msms.edgeR**: The binomial negative of edgeR
- **pval.by.fc**: Table of cumulative frequencies of features by p-values in bins of log fold change
- **test.results**: Multitest p-value adjustment and post-test filter
- **res.volcanoplot**: Volcanoplot of the results

**Author(s)**

Josep Gregori, Alex Sanchez, and Josep Villanueva
Maintainer: Josep Gregori <josep.gregori@gmail.com>

**References**

msms.edgeR

Description
Given a null and an alternative model, with a two level treatment factor as the two conditions to compare, executes the negative binomial test by edgeR functions to discover differentially expressed proteins between the two conditions. The null and alternative models may include blocking factors. The reference level of the main factor is considered to be the control condition.

Usage
msms.edgeR(msnset, form1, form0, facs = NULL, div = NULL, fnm = NULL)

Arguments
msnset A MSnSet object with spectral counts in the expression matrix.
form1 The alternative hypothesis model as an standard R formula, with the treatment factor of interest, and eventual blocking factors.
form0 The null hypothesis model as an standard R formula. It may be the standard null model (y~.) or contain one or multiple blocking factors.
facs NULL or a data frame with the factors in its columns.
div NULL or a vector with the divisors used to compute the offsets.
fnm NULL or a character string with the treatment factor name, as used in the column names of the factors data frame, and in the formula.

Details
The right hand site of the formulas is expected to be "y~", with the combination of factors after the tilde. If facs is NULL the factors are taken as default from pData(msnset). If div is NULL all divisors are taken equal to one. If fnm is NULL it is taken to be the first factor in facs.

Value
A data frame with column names 'LogFC', 'LR', 'p.value', with the estimated log fold changes, likelihood ratio statistic and corresponding p-value as obtained from a call to glmLRT() from the edgeR package.

Author(s)
Josep Gregori i Font

References
See Also

MSnSet, edgeR, glmLRT, msmsEDA

Examples

```r
## Example
library(msmsTests)
data(msms.dataset)
e <- pp.msms.data(msms.dataset)
e

null.f <- "y~batch"
alt.f <- "y~treat+batch"
div <- apply(exprs(e),2,sum)
res <- msms.edgeR(e,alt.f,null.f,div=div,fnm="treat")
str(res)
head(res)
```

---

**msms.glm.pois**

*Spectral counts differential expression by Poisson GLM*

### Description

Given a null and an alternative model, with a two level treatment factor as the two conditions to compare, executes a Poisson based GLM regression to discover differentially expressed proteins between the two conditions. The null and alternative models may include blocking factors. The reference level of the main factor is considered to be the control condition.

### Usage

```r
msms.glm.pois(msnset,form1,form0,facs=NULL,div=NULL)
```

### Arguments

- **msnset**: A MSnSet object with spectral counts in the expression matrix.
- **form1**: The alternative hypothesis model as an standard R formula, with the treatment factor of interest, and eventual blocking factors.
- **form0**: The null hypothesis model as an standard R formula. It may be the standard null model (y~.) or contain one or multiple blocking factors.
- **facs**: NULL or a data frame with the factors in its columns.
- **div**: NULL or a vector with the divisors used to compute the offsets.

### Details

The right hand site of the formulas is expected to be "y~", with the combination of factors after the tilde. If facs is NULL the factors are taken as default from pData(msnset). If div is NULL all divisors are taken equal to one.
Value
A data frame with the following columns:

- **LogFC**: Log fold change estimated from the model parameters.
- **D**: Residual deviance as statistic of the test.
- **p.value**: The p-values obtained from the test.

Author(s)
Josep Gregori i Font

References
- Josep Gregori, Laura Villareal, Alex Sanchez, Jose Baselga, Josep Villanueva (2013). An Effect Size Filter Improves the Reproducibility in Spectral Counting-based Comparative Proteomics. Journal of Proteomics, DOI [http://dx.doi.org/10.1016/j.jprot.2013.05.030](http://dx.doi.org/10.1016/j.jprot.2013.05.030)

See Also
MSnSet, glm

Examples
```r
library(msmsTests)
data(msms.dataset)
# Pre-process expression matrix
e <- pp.msms.data(msms.dataset)
# Factors
pData(e)
# Control condition
levels(pData(e)$treat)[1]
# Treatment condition
levels(pData(e)$treat)[2]

# Models and normalizing condition
null.f <- "y~batch"
alt.f <- "y+treat+batch"
div <- apply(exprs(e),2,sum)

#Test
res <- msms.glm.pois(e,alt.f,null.f,div=div)
str(res)
head(res)
```
### Spectral counts differential expression by quasi-likelihood GLM

#### Description
Given a null and an alternative model, with a two level treatment factor as the two conditions to compare, executes a quasi-likelihood based GLM regression to discover differentially expressed proteins between the two conditions. The null and alternative models may include blocking factors. The reference level of the main factor is considered to be the control condition.

#### Usage
```
msms.glm.qlll(msnset, form1, form0, facs=NULL, div=NULL)
```

#### Arguments
- **msnset**: A MSnSet object with spectral counts in the expression matrix.
- **form1**: The alternative hypothesis model as a standard R formula, with the treatment factor of interest, and eventual blocking factors.
- **form0**: The null hypothesis model as a standard R formula. It may be the standard null model (y~.) or contain one or multiple blocking factors.
- **facs**: NULL or a data frame with the factors in its columns.
- **div**: NULL or a vector with the divisors used to compute the offsets.

#### Details
The right hand site of the formulas is expected to be "y~", with the combination of factors after the tilde. If `facs` is NULL the factors are taken as default from `pData(msnset)`. If `div` is NULL all divisors are taken equal to one.

#### Value
A data frame with the following columns:
- **LogFC**: Log fold change estimated from the model parameters.
- **D**: Residual deviance as statistic of the test.
- **p.value**: The p-values obtained from the test.

#### Author(s)
Josep Gregori i Font

#### References


See Also

`MSnSet, glm`

Examples

```r
library(msmsTests)
data(msms.dataset)
# Pre-process expression matrix
e <- pp.msms.data(msms.dataset)
# Factors
pData(e)
# Control condition
levels(pData(e)$treat)[1]
# Treatment condition
levels(pData(e)$treat)[2]

# Models and normalizing condition
null.f <- "y~batch"
alt.f <- "y~treat+batch"
div <- apply(exprs(e),2,sum)

# Test
res <- msms.glm.qlll(e,alt.f,null.f,div=div)
str(res)
head(res)
```

Descrizione

Un MSnSet con una matrice di conti spettrali nella matrice di espressione, e un fattore di trattamento nella slot `phenoData`. La matrice di conti spettrali ha campioni nella colonna, e proteine nelle righe. Ogni campione è costituito da 500ng di lisata standard di lievi spazzolati con 100, 200, 400 e 600fm di un mix di 48 proteine umane equimolari (UPS1, Sigma-Aldrich). Il dataset contiene un numero diverso di ripetizioni tecniche per ogni campione.

Uso

`data(msms.spk)`
pval.by.fc

Format
A MSnSet

References

Laurent Gatto and Kathryn S. Lilley, MSnbase - an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation, Bioinformatics 28(2), 288-289 (2012).

See Also
See MSnSet for detail on the class, and the exprs and pData accessors.

Examples
data(msms.spk)
msms.spk
dim(msms.spk)
table(pData(msms.spk))
head(exprs(msms.spk))

pval.by.fc

Table of cumulative frequencies of p-values by log fold change bins

Description
Given the set of p-values and log fold changes that result from a test, computes a table of cumulative frequencies of features by p-values in bins of log fold changes.

Usage
pval.by.fc(pvals,lfc)

Arguments
lfc
The log fold changes estimated from the tests.
pvals
The p-values, adjusted or not, obtained from the tests.

Value
A matrix of cumulated frequencies with descriptive row and column names.

Author(s)
Josep Gregori i Font
References

See Also
test.results

Examples
library(msmsTests)
data(msms.spk)
# Subset
treat <- pData(msms.spk)
jdx <- which(treat=="U200" | treat=="U600")
e <- msms.spk[,jdx]
pData(e)$treat <- treat[jdx,1,drop=TRUE]
# Pre-process expression matrix
e <- pp.msms.data(e)
# Models and normalizing condition
null.f <- "y-1"
alt.f <- "y-treat"
div <- apply(exprs(e),2,sum)
#Test
res <- msms.glm.pois(e,alt.f,null.f,div=div)
# Post-test filter
lst <- test.results(res,e,pData(e)$treat,"U600","U200",div,
                    alpha=0.05,minSpC=2,minLFC=1,
                    method="BH")

## On all features, with multitest adjusted p-values
pval.by.fc(lst$tres$adjp, lst$tres$LogFC)

### On all features deemed significant and biologically relevant
flags <- lst$tres$DEP
pval.by.fc(lst$tres$adjp[flags], lst$tres$LogFC[flags])

---

res.volcanoplot Volcanoplot

Description
Given the data frame obtained from test.results() a volcano plot is drawn. The features are colored according to significance and relevance.

Usage
res.volcanoplot(tres, max.pval=0.05, min.LFC=1, maxx=3, maxy=10, ylblr=20)
Arguments

 tres The dataframe with test results as obtained from test.results(). Or a dataframe with, at least, the following columns: LogFC with log fold changes, adjp with multitest adjusted p-values, and DEP with TRUE or FALSE as post test filter results, being the TRUE features both statistically significant and relevant for reproducibility.

 max.pval The maximum adjusted p-value considered as statistically significant.

 min.LFC The minimum absolute log fold change considered as biologically relevant.

 maxx The maximum value in abscissas (i.e. log2(fold change)).

 maxy The maximum value in ordinates (i.e. -log10(p.val)).

 yblbs All features with -log10(p.val) above this value will be plotted with feature labels.

Details

Abscissas and ordinates may be limited giving a value other than NULL to the parameters maxx and maxy. All features deemed significant and relevant are plotted by a blue dot, all features deemed significant but not passing the post test filter are plotted by a red dot. The non-significant features are plotted as smaller black dots. All features deemed significant and relevant and with a -log10 p-value above yblbs are plotted with a label showing their row index in the test results dataframe. The borders limiting the values given by max.pval and min.LFC are plotted as dash-and-dot red lines.

Value

No return value.

Author(s)

Josep Gregori i Font

References


See Also

test.results, volcanoplot

Examples

library(msmsTests)
data(msms.dataset)
# Pre-process expression matrix
e <- pp.msms.data(msms.dataset)
# Models and normalizing condition
null.f <- "y~batch"
alt.f <- "y~treat+batch"
div <- apply(exprs(e),2,sum)
# Test
test.results

res <- msms.glm.qlll(e, alt.f, null.f, div = div)
lst <- test.results(res, e, pData(e)$treat, "U600", "U200", div,
alpha = 0.05, minSpC = 2, minLFC = log2(1.8),
method = "BH")
# Plot
res.volcanoplot(lst$tres, max.pval = 0.05, min.LFC = 1, maxx = 3, maxy = NULL,
ylbd = 4)

Description

Operates on the statistic tests results obtained from msms.glm.pois(), msms.glm.qlll() or msms.edgeR().
The following variables are computed: Raw expression mean values for each condition (control and
treatment), log fold change based on these expression levels and taking into account the normal-
izing divisors (div), multtest adjusted p-values with FDR control, and a post test filter based on
minimum spectral counts and minimum absolute log fold change as estimated by the statistic test.
According to the results of this post-test filter, features are flagged as T or F depending on whether
they result relevant or not, beyond their statistic significance.

Usage

test.results(test, msnset, gpf, gp1, gp2, div, alpha = 0.05,
minSpC = 2, minLFC = 1, method = "BH")

Arguments

test The dataframe obtained from either msms.glm.pois(), msms.glm.qlll() or msms.edgeR()
msnset A MSnSet object with spectral counts in the expression matrix.
gpf The factor used in the tests.
gp1 The treatment level name.
gp2 The control level name. Should be the factor’s reference level. See R function
relevel.
div The weights used as divisors (offsets) in the GLM model. Usually the sum of
spectral counts of each sample.
alpha The multi test adjusted p-value significance threshold.
minSpC The minimum spectral counts considered as relevant in the most abundant con-
tdition. This filter aims at reaching good reproducibility.
minLFC The minimum absolute log fold change considered both, relevant and biologi-
cally significant. This filter aims at assuring enough biological effect size and at
reaching good reproducibility.
method One among BH or qval. The p-values are FDR adjusted by the Benjamini-
Hochberg method (BH) or by qvalue (qval).

Details

No feature is removed in the filter, but instead they are flagged as TRUE or FALSE depending
on whether they are considered as differentially expressed or not, in the DEP column, taking into
account statistic significance and reproducibility metrics.
Value

A data frame with the following columns:

- **first column**: Column named as the treatment level with the mean raw spectral counts observed for this condition.
- **second column**: Column named as the control level with the mean raw spectral counts observed for this condition.
- **lFC.Av**: Log fold change computed from the mean expression levels taking into account the given normalization factors.
- **logFC**: Log fold change estimated by fitting the given GLM model. The reference level of the main factor is taken as control.
- **D or LR**: The statistic obtained from the tests. The residual deviance D for Poisson and quasi-likelihood, or the likelihood ratio LR for edgeR.
- **p.val**: The unadjusted p-values obtained from the tests.
- **adjp**: The multitest adjusted p-values with FDR control.
- **DEP**: A logical flagging the features considered both as statistically significant and relevant for reproducibility.

Author(s)

Josep Gregori i Font

References


Alan Dabney, John D. Storey and with assistance from Gregory R. Warnes. qvalue: Q-value estimation for false discovery rate control. R package version 1.30.0.

See Also

pval.by.fc, p.adjust, qvalue.

Examples

```r
library(msmsTests)
data(msms.dataset)
# Pre-process expression matrix
e <- pp.msms.data(msms.dataset)
# Factors
pData(e)
# Control condition
levels(pData(e)$treat)[1]
# Treatment condition
levels(pData(e)$treat)[2]
# Models and normalizing condition
null.f <- "y=batch"
alt.f <- "y=treat+batch"
```
test.results

```r
div <- apply(exprs(e),2,sum)
# Test
res <- msms.glm.qlll(e,alt.f,null.f,div=div)
# Post-test filter
lst <- test.results(res,e,pData(e)$treat,"U600","U200",div,
alpha=0.05, minSpC=2, minLFC=1,
method="BH")
str(lst)
lst$cond
head(lst$tres)
rownames(lst$tres)[which(lst$tres$DEP)]
```
Index

*Topic datasets
  msms.spk, 7

*Topic design
  msms.edgeR, 2
  msms.glm.pois, 4
  msms.glm.qlll, 6

*Topic hplot
  msmsTests-package, 2
  res.volcanoplot, 9

*Topic htest
  msmsTests-package, 2
  pval.by.fc, 8
  res.volcanoplot, 9

*Topic models
  msms.edgeR, 2
  msms.glm.pois, 4
  msms.glm.qlll, 6

*Topic univar
  msms.edgeR, 2
  msms.glm.pois, 4
  msms.glm.qlll, 6
  pval.by.fc, 8
  res.volcanoplot, 9

design, 7
edgeR, 4
glm, 5, 7
glmLRT, 4

msms.edgeR, 2
msms.glm.pois, 4
msms.glm.qlll, 6
msms.spk, 7
msmsEDA, 4
msmsTests (msmsTests-package), 2
msmsTests-package, 2
MSnSet, 4, 5, 7, 8

p.adjust, 12
pval.by.fc, 8, 12
qvalue, 12
res.volcanoplot, 9