Package ‘selfea’

May 31, 2015

Type Package

Title Select Features Reliably with Cohen’s Effect Sizes

Version 1.0

Date 2015-05-25

Author Lang Ho Lee, Arnold Saxton, Nathan Verberkmoes

Maintainer Lang Ho Lee <langholee@gmail.com>

Depends R (>= 3.1.3), pwr, MASS, plyr

Description Functions using Cohen's effect sizes (Cohen, J. A power primer. Psychological bulletin 112, 155 (1992)) are provided for reliable feature selection in biology data analysis. With Cohen's effect sizes, p-values are calculated and adjusted from quasi-Poisson GLM, negative binomial GLM and Normal distribution ANOVA.

License GPL-2

NeedsCompilation no

Repository CRAN

Date/Publication 2015-05-31 09:28:56

R topics documented:

selfea-package ........................................... 2
example_data ........................................... 3
example_data1 ......................................... 4
example_data2 ......................................... 4
example_group ......................................... 5
get_statistics_from_dataFrame ...................... 6
get_statistics_from_file .............................. 7
glm_anova .............................................. 9
top_table ............................................. 10

Index 12
Description

description

Details

<table>
<thead>
<tr>
<th>Package:</th>
<th>selfea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type:</td>
<td>Package</td>
</tr>
<tr>
<td>Version:</td>
<td>1.0</td>
</tr>
<tr>
<td>Date:</td>
<td>2015-03-31</td>
</tr>
<tr>
<td>License:</td>
<td>GPL-2</td>
</tr>
</tbody>
</table>

Author(s)

Lang Ho Lee, Arnold Saxton, Nathan Verberkmoes

Maintainer: Lang Ho Lee <llee27@utk.edu>

References

Lang Ho Lee, Arnold Saxton, Nathan Verberkmoes, Selfea: A R package for reliable feature selection in process

See Also

get_statistics_from_file top_table get_statistics_from_dataFrame

Examples

library(selfea)

## Load Gregori's data and test Selfea

## 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'

## Description:
## Each sample consists in 500ng of standard yeast llsate spiked with
## 100, 200, 400 and 600fm of a mix of 48 equimolar human proteins (UPS1, Sigma-Alrlich).
## The dataset contains different number of technical replimessagees of each sample
## Description

The spectral counts matrix has samples in the columns, and proteins in the rows. Each sample consists in 500ng of standard yeast lysate spiked with 100, 200, 400 and 600fm of a mix of 48 equimolar human proteins (UPS1, Sigma-Aldrich). The dataset contains a different number of technical replicates of each sample. This dataset has only 100 proteins of total 685 proteins in the original data for fast example execution. If you want to use whole dataset, go for 'example_data2'.

## Usage

```r
data(example_data1)
```

## Format

A data frame containing protein IDs and their expression profile.

## 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)

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).

## Examples

```r
data(example_data1)
```
Yeast lisate samples spiked with human proteins

Description
The spectral counts matrix has samples in the columns, and proteins in the rows. Each sample consists in 500ng of standard yeast lisate spiked with 100, 200, 400 and 600fm of a mix of 48 equimolar human proteins (UPS1, Sigma-Aldrich). The dataset contains a different number of technical replicates of each sample. This dataset has only 100 proteins of total 685 proteins in the original data for fast example execution. If you want to use whole dataset, go for ’example_data2’.

Usage

data(example_data1)

Format
Two data frames, df_contrast (protein expression profile) and df_group (experiment group information).

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).

Examples

data(example_data1)
Format

Two data frames, df_contrast (protein expression profile) and df_group (experiment group information).

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).

Examples

data(example_data1)

Description

The spectral counts matrix has samples in the columns, and proteins in the rows. Each sample consists in 500ng of standard yeast lisate spiked with 100, 200, 400 and 600fm of a mix of 48 equimolar human proteins (UPS1, Sigma-Aldrich). The dataset contains a different number of technical replicates of each sample. This dataset has only 100 proteins of total 685 proteins in the original data for fast example execution. If you want to use whole dataset, go for ‘example_data2’.

Usage

data(example_data1)

Format

A data frame containing MS Run names and their corresponding experiment groups

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).

Examples

data(example_data1)
get_statistics_from_dataFrame

Description

A function "get_statistics_from_dataFrame" computes several statistics by reading csv files obtained from input arguments.

Usage

get_statistics_from_dataFrame(df_contrast, df_group, padj = "fdr")

Arguments

df_contrast

A data frame that consists of 'ID' column and expression profile (columns after 'ID' column). 'ID' column should be unique. Column names after 'ID' column should be unique. Only positive numbers are allowed in expression data. Here is an example.

<table>
<thead>
<tr>
<th>ID</th>
<th>Y500U100_001</th>
<th>Y500U100_002</th>
<th>Y500U100_003</th>
<th>Y500U100_004</th>
<th>Y500U200_001</th>
<th>Y500U200_002</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>YKL060C</td>
<td>151</td>
<td>195</td>
<td>188</td>
<td>184</td>
<td>221</td>
</tr>
<tr>
<td>2</td>
<td>YDR155C</td>
<td>154</td>
<td>244</td>
<td>237</td>
<td>232</td>
<td>190</td>
</tr>
<tr>
<td>3</td>
<td>YOL086C</td>
<td>64</td>
<td>89</td>
<td>128</td>
<td>109</td>
<td>116</td>
</tr>
<tr>
<td>4</td>
<td>YJR104C</td>
<td>161</td>
<td>155</td>
<td>158</td>
<td>172</td>
<td>164</td>
</tr>
<tr>
<td>5</td>
<td>YGR192C</td>
<td>157</td>
<td>161</td>
<td>173</td>
<td>175</td>
<td>177</td>
</tr>
<tr>
<td>6</td>
<td>YLR150W</td>
<td>96</td>
<td>109</td>
<td>113</td>
<td>115</td>
<td>119</td>
</tr>
<tr>
<td>7</td>
<td>YPL037C</td>
<td>23</td>
<td>28</td>
<td>27</td>
<td>27</td>
<td>48</td>
</tr>
<tr>
<td>8</td>
<td>YNL007C</td>
<td>53</td>
<td>58</td>
<td>64</td>
<td>63</td>
<td>66</td>
</tr>
<tr>
<td>9</td>
<td>YBR072W</td>
<td>52</td>
<td>53</td>
<td>54</td>
<td>44</td>
<td>73</td>
</tr>
<tr>
<td>10</td>
<td>YDR418W_1</td>
<td>76</td>
<td>53</td>
<td>62</td>
<td>74</td>
<td>63</td>
</tr>
</tbody>
</table>

df_group

A data frame that consists of 'Col_Name' and 'Group' columns. This parameter is to match experiment groups to expression profiles of df_contrast. 'Col_Name' should be corresponding to column names of expression profile of df_contrast. 'Group' columns have experiment information of columns in expression profile of df_contrast. Here is an example. See the example of df_contrast together.

<table>
<thead>
<tr>
<th>Col_Name</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Y500U100_001</td>
</tr>
<tr>
<td>2</td>
<td>Y500U100_002</td>
</tr>
<tr>
<td>3</td>
<td>Y500U100_003</td>
</tr>
<tr>
<td>4</td>
<td>Y500U100_004</td>
</tr>
<tr>
<td>5</td>
<td>Y500U200_001</td>
</tr>
<tr>
<td>6</td>
<td>Y500U200_002</td>
</tr>
<tr>
<td>7</td>
<td>Y500U200_003</td>
</tr>
</tbody>
</table>
get_statistics_from_file

padj Choose one of these c("holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", "none"). "fdr" is default option. The option is same to `p.adjust`.

Value

A list that consists of the following items:

$data_table A data frame that consists of ID, GLM Negative Binomial P-value, Cohen’s W, GLM Quasi-Poisson P-value, ANOVA with Normal P-value and Cohen’s f.
$min_rep Common number of replicates in your group information. Generally, it is the minimum number of replicates.
$max_rep Maximum number of replicates in your group information.
$nt The number of total experiments in your expression profile.
$ng The number of groups in your group information.

Examples

```r
library(selfea)

## For this example we will import Gregori data
## 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

## Description:
## Each sample consists in 500ng of standard yeast listate spiked with
## 100, 200, 400 and 600fm of a mix of 48 equimolar human proteins (UPS1, Sigma-Aldrich).
## The dataset contains a different number of technical replicates of each sample

## import Gregori data
data(example_data)
df_contrast <- example_data
df_group <- example_group

## Get statistics through 'get_statistics_from_dataFrame' function
list_result <- get_statistics_from_dataFrame(df_contrast, df_group)

## Get significant features (alpha >= 0.05 and power >= 0.84)
significant_qpf <- top_table(list_result, pvalue=0.05, power_desired=0.84, method='QPF')
```

get_statistics_from_file

get_statistics_from_file

Description

A function "get_statistics_from_file" computes several statistics by reading csv files obtained from input arguments.
get_statistics_from_file

Usage

get_statistics_from_file(file_expr = "", file_group = "", padj = "fdr")

Arguments

file_expr a CSV type file, comma (,) separated file format, that has unique "ID" at the first column and expression data for the corresponding ID. Here is an short example.

ID,Y500U100_001,Y500U100_002,Y500U100_003,Y500U100_004,Y500U200_001,Y500U200_002
YKL060C,151,195,188,184,221,201
YDR155C,154,244,237,232,190,187
YOL086C,64,89,128,109,116,119

file_group a CSV type file, comma (,) separated file format, that consists of "Col_Name", column names of "file_expr" parameter, and "Group" information of the corresponding column name. The order of "Col_Name" column have to be same to order of columns in "file_expr". Here is an example. See also the example above.

Col_Name,Group
Y500U100_001,U100
Y500U100_002,U100
Y500U100_003,U100
Y500U100_004,U100
Y500U200_001,U200
Y500U200_002,U200

padj Choose one of these c("holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", "none"). "fdr" is default option.

Value

A list that consists of the following items:

$data_table A data frame that consists of ID, GLM Negative Binomial P-value, Cohen’s W, GLM Quasi-Poisson P-value, ANOVA Normal P-value and Cohen’s f.

$min_rep Common number of replicates in your group information. Generally, it is the minimum number of replicates.

$max_rep Maximum number of replicates in your group information.

$nt The number of total experiments in your expression profile.

$ng The number of groups in your group information.

Examples

library(selfea)

## For this example we will import Gregori data
## Josep Gregori, Laura Villareal, Alex Sanchez, Jose Baselga, Josep Villanueva (2013).
glm_anova

## Description

Calculate P-values from ANOVA using Normal, Quasi-Poisson and Negative Binomial distribution and Cohen's effect sizes

## Usage

```r
glm_anova(dataset.expr, dataset.ID, group, padj = "fdr")
```

## Arguments

- `dataset.expr`  
  A data frame that has column names for distinguishing experiments and numerical values for expression levels
- `dataset.ID`  
  A vector of the obtained expression profile's ID column
- `group`  
  A data frame that consists of 'Col_Name' and 'Group' obtained from the user file through `get_statistics_from_file`
- `padj`  
  Choose one of these c("holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", "none"). "fdr" is default option.

## Value

A data frame containing ID, Cohen's W, Cohen's F, Max fold change, GLM Negative Binomial P-value, GLM Quasi-Poisson P-value and ANOVA with Normal P-value.
Description

Get IDs that pass two filters, p-value and effect-size. This top_table will make a significant list that is less than p-value and greater than effect-size. Effect-size are calculated by obtained power level.
This function requires four parameters. ex) top_table(input_data,pvalue=0.05,power_desired=0.84,method='QPF')

Usage

top_table(input_listL pvalue = P.PUL power_desired = P.8TL method = BqpfBL fc_threshold = RI

Arguments

input_list The list produced by 'get_statistics_from_file' or 'get_statistics_from_dataFrame' function. See get_statistics_from_file and get_statistics_from_dataFrame for more information. It consists of the following items:

$\text{data_table}$ A data frame that consists of ID, GLM Negative Binomial P-value, Cohen’s W, GLM Quasi-Poisson P-value, ANOVA with Normal P-value, ANOVA with Normal F value, ANOVA with Normal F value.

$\text{min_rep}$ Common number of replicates in your group information. Generally, it is the minimum number of replicates.

$\text{max_rep}$ Maximum number of replicates in your group information.

$\text{nt}$ The number of total experiments in your expression profile.

$\text{ng}$ The number of groups in your group information.

pvalue p-value should be ranged between 0 to 1. default is 0.05.

power_desired Give the statistical power you desired for output significant list

method Choose statistics method you want to use for making significant list

1 "QPF" combination of Quasi-Poisson and Cohen’s f. Default.

2 "QPFC" combination of Quasi-Poisson and Fold change.

3 "NBW" combination of Negative Binomial and Cohen’s w.

4 "NBFC" combination of Negative Binomial and Fold change.

5 "NORF" combination of ANOVA with normal distribution and Cohen’s f.

6 "NORFC" combination of ANOVA with normal distribution and Fold change.

fc_threshold Fold change you want to use. Default is 2.

Value

A list containing the follow items.

top_table a data frame that consists of ID, Cohen’s W, Cohen’s F, Max fold change, GLM Negative Binomial P-value, GLM Quasi-Poisson P-value, ANOVA with Normal P-value, ANOVA with Normal F value. 

执行top_table函数的示例：

```r
top_table(input_list, pvalue = 0.05, power_desired = 0.84, method = "QPF", fc_threshold = 2)
```
### Examples

library(selfea)

```r
## For this example we will import Gregori data
## 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)

## Description:
## Each sample consists in 500ng of standard yeast listate spiked with
## 100, 200, 400 and 600fm of a mix of 48 equimolar human proteins (UP51, Sigma-Aldrich).
## The dataset contains a different number of technical replimessagees of each sample

## import Gregori data
data(example_data)
df_contrast <- example_data
df_group <- example_group

## Get statistics through 'get_statistics_from_dataFrame' function
list_result <- get_statistics_from_dataFrame(df_contrast, df_group)

## Get significant features (alpha >= 0.05 and power >= 0.84)
significant_qpf <- top_table(list_result, pvalue=0.05, power_desired=0.84, method='QPF')
```
Index

*Topic Quasi-Poisson, Negative Binomial, Cohen’s f, Cohen’s w, ANOVA
  selfea-package, 2
*Topic datasets
  example_data, 3
  example_data1, 4
  example_data2, 4
  example_group, 5
  example_data, 3
  example_data1, 4
  example_data2, 4
  example_group, 5
  get_statistics_from_dataFrame, 2, 6, 10
  get_statistics_from_file, 2, 7, 10
  glm_anova, 9
  p.adjust, 7
  selfea(selfea-package), 2
  selfea-package, 2
  top_table, 2, 10