Package 'BOG'

March 3, 2015

Type Package

Title Bacterium and Virus Analysis of Orthologous Groups (BOG) is a Package for Identifying Differentially Regulated Genes in the Light of Gene Functions

Version 2.0

Date 2015-02-23

Author Jincheol Park (Keimyung University, South Korea), Cenny Taslim, Shili Lin (The Ohio State University, USA)

Maintainer Jincheol Park <park.jincheol@gmail.com>

Description An implementation of three statistical tests for identification of COG (Cluster of Orthologous Groups) that are over represented among genes that show differential expression under conditions. It also provides tabular and graphical summaries of the results for easy visualisation and presentation.

LazyLoad yes

Depends R(>= 3.1.2), hash, DIME

License GPL-3

NeedsCompilation no

Repository CRAN

Date/Publication 2015-03-03 14:32:46

R topics documented:

`BOG-package` *BOG is a package for identifying differentially regulated genes in the light of gene functions.*

Description

The BOG package is designed to identify COG (Cluster of Orthologous Groups) that are over represented among genes that show differential expression under two different conditions using statistical analysis and graphical representations. Particular statistical analysis includes the hypergeometric test (HG), Mann-Whitney rank sum test (RANK), and Gene set enrichment analysis test (GSEA). The results are then organized and presented tabularly and graphically for easy visualization and presentation.

Details

Package:	BOG
Type:	Package
Version:	1.0
Date:	2015-2-23
License:	GPL-3

Author(s)

Jincheol Park (Keimyung University, South Korea), Cenny Taslim, Shili Lin (The Ohio State University, USA)

Maintainer: Jincheol Park <park.jincheol@gmail.com>

References

Carlson, P. et al. (2009) Transcriptional Profiling of Bacillus anthracis Sterne (34F2) during Iron Starvation. PLOS ONE 4(9): e6988. Doi:10.1371/journal.pone.0006988.

Khalili, A. et al. (2009) A robust unified approach to analyzing methylation and gene expression data. Comput. Stat. Data Anal., 53, 1701-1710.

Subramian, A. et al.(2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS vol.102 no.43, 15545-15550.

Taslim, C. et al. (2010) DIME: R-package for Identifying Differential ChIP-seq Based on an Ensemble of Mixture Models. Bioinformatics, 27, 1569-1570.

`anthracis` *Bacillus anthracis str. 'Ames Ancestor'*

Description

This is protein details for Bacillus anthracis str. 'Ames Ancestor' with COG annotated. If a user does not specify cog_file argument in BOG() command, this built-in COG file will be loaded by default.

Usage

```
data(anthracis)
```

Format

List of two elements with names: geneID, COG

Source

http://www.ncbi.nlm.nih.gov/genome/proteins/181?project_id=58083

Examples

```
data(anthracis)
```

`anthracis_iron` *anthracis_iron*

Description

This is the anthracis data of (Carson et.al,2009) with adjusted p-value produced by DIME. It consists of anthracis geneID and adjusted p-value for gene expression between two conditions. If a user does not specify the data argument in BOG() command, this built-in data will be loaded by default as an example.

Usage

```
data(anthracis_iron)
```

Format

A data frame with 5537 observations on the following 2 variables: geneID, adj.pval

References

Carlson, P. et al. (2009) Transcriptional Profiling of Bacillus anthracis Sterne (34F2) during Iron Starvation. PLOS ONE 4(9): e6988. Doi:10.1371/journal.pone.0006988.

Examples

```
data(anthracis_iron)
```

BOG *BOG*

Description

This function is the flagship function of BOG. It reads data and COG annotation files with user specified setting for analysis. It performs hypergeometric test, rank test (Mann-Whiteny test), and GSEA test (optional).

Usage

```
BOG(data = NULL, data.type = c("data", "pval"), cog.file = NULL,
hg.thresh = 0.05, gsea = FALSE,
DIME.K = 5, DIME.iter = 50, DIME.rep = 5)
```

Arguments

`data`	This input file can either be a dataframe or a text file consisting of two columns. The first column is the geneIDs (characters). The second column provides numerical measures for the corresponding genes, which has three possible options controlled by the data_type argument. If data is not specified, BOG load a built-in data, anthracis_adjpval, by default, as an example data set.
`data.type`	1. data.type="data" : normalized "differences" of gene expressions between two comparison groups. 2. data.type="pval" : raw p-values or multiple testing adjusted p-values for each gene if differential analysis is carried out beforehand If the data is specified for option(1), then DIME will be called to perform the differental analysis. Under option(2), no preprocessing is needed before carrying out the tests. Default data.type is "data".
`cog.file`	This can either be a user specified input file (R dataframe), a raw text file, or simply the specification of the name of one of the six built-in COGs: anthracis, brucella, coxiella, difficile, ecoli, or francisella. If the virus/bateria being analyzed is not one of the six built-in varieties, then a data frame or a text file with two columns is required: the first column provides geneIDs as in the input data file; the second column specifies the cluster of orthologous groups to which each gene belongs. BOG will perform statistical tests by first merging the data and cog_file using geneID as the key, hence it is important that geneIDs in both dataframes match. If cog_file is not specified, BOG loads "anthracis" by default.

hg.thresh	In the statistical analysis, BOG uses local-fdr(or p-value) as a score for strength of evidence for differences between groups. The smaller value of the score, the stronger is the evidence for differences in gene expression. hg.thresh is a threshhold used in hypergeometric test. By default, it is set to be 0.05.
gsea	By default, gsea is set FALSE so that unless user specify it to be TRUE, BOG does not perform GSEA test.
DIME.K	The number of mixture components in fitting an ensemble of mixture models, if DIME processing is activated. If user select data.type="pval", this is irrelevant. The default is 5.
DIME.iter	The number of iterations in fitting an ensemble of mixture models. This is only relevant if data.type="data". The default is 50.
DIME.rep	The number of repetitions in fitting an ensemble of mixture models. This is only relevant if data.type="data". The default is 5.

Value

List with three elements : stat, dime, dime_data.

stat	stat is a list consisting of statistical analysis outputs.
dime	dime is a list consisting of DIME outputs.
dime_data	dime_data is a list of the data specified in the data argument.

Examples

```
bog=BOG(data="anthracis_iron",data.type="pval",cog.file="anthracis",gsea=FALSE)
```

BOGest *BOGest*

Description

This is an internal function so that it is not expected for a user to use it.

Usage

```
BOGest(data, data.type, cog.file, hg.thresh, gsea, DIME.K, DIME.iter, DIME.rep)
```

Arguments

The definition of all the arguments are same as the ones described in the BOG() command so that a user may refer to BOG() command for details.

This input file can be either a dataframe or a text file consisting of two columns. The first column is the geneIDs (charaters). The second column provides numerical measures for the corresponding genes, which has three possible options controlled by the data_type argument. If data is not specified, BOG load a built-in data, anthracis_adjpval, by default.

data.type	1. data.type="data" : normalized "differences" of gene expressions between two comparison groups. 2. data.type="pval" : raw p-values or multiple testing adjusted p-values for each gene if differential analysis is carried out beforehand If the data is specified for option(1), then DIME will be called to perform the differental analysis. Under option(2), no preprocessing is needed before carrying out the tests. Default data.type is "data".
cog.file	This can either be a user specified input file (R dataframe), a raw text file, or simply the specification of the name of one of the six built-in COGs: anthracis, brucella, coxiella, difficile, ecoli, or francisella. If the virus/bateria being analyzed is not one of the six built-in varieties, then a data frame or a text file with two columns is required: the first column provides geneIDs as in the input data file; the second column specifies the cluster of orthologous groups to which each gene belongs. BOG will perform statistical tests by first merging the data and cog_file using geneID as the key, hence it is important that geneIDs in both dataframes match. If cog_file is not specified, BOG loads "anthracis" by default.
hg.thresh	In statistical analysis, BOG uses local-fdr(or p-value) as a score for strength of evidence for differences between groups. The smaller absolute value of the score, the stronger is the evidence for differences in gene expression. hg.thresh is a threshhold used in hypergeometric test. By default, it is set 0.05.
gsea	By default, gsea is set FALSE so that unless user specify it to be TRUE, BOG does not perform GSEA test.
DIME.K	The number of mixture components in fitting an ensemble of mixture models. The default is 5.
DIME.iter	The number of iterations in fitting an ensemble of mixture models. The default is 50.
DIME.rep	The number of repitions in fitting an ensemble of mixture models. The default is 5.

Value

List with three elements : stat, dime, dime_data.

stat	stat is a list consisting of statistical analysis outputs.
dime	dime is a list consisting of DIME outputs.
dime_data	dime_data is a list of the data specified in the data argument.

BOGstat *BOGstat*

Description

This is an internal function so that it is not expected for a user to use it.

Usage

```
BOGstat(db_Gene, hg.thresh, gsea)
```

Arguments

db_Gene	Internally defined list object.
hg.thresh	In statistical analysis, BOG uses local-fdr as a score for strength of evidence for differences between groups. The smaller absolute value of local fdr, the stronger is the evidence for differences in gene expression. fdr.cutoff is a threshhold used in hypergeometric test. By default, it is set 0.05.
gsea	By default, gsea is set FALSE so that unless user specify it to be TRUE, BOG does not perform GSEA test.

---

brucella *Brucella suis 1330*

Description

This is protein details for Brucella suis 1330 Sequence with COG annotated.

Usage

```
data(brucella)
```

Format

List of two elements with names: geneID, COG

Source

http://www.ncbi.nlm.nih.gov/genome/proteins/806?project_id=57927

Examples

```
data(brucella)
```

coxiella *Coxiella burnetii RSA 493*

Description

This is protein details for Coxiella burnetii RSA 493 with COG annotated.

Usage

```
data(coxiella)
```

Format

List of two elements with names: geneID, COG

Source

http://www.ncbi.nlm.nih.gov/genome/proteins/543?project_id=57631

Examples

```
data(coxiella)
```

difficile *Clostridium difficile 630*

Description

This is protein details for Clostridium difficile 630 with COG annotated.

Usage

```
data(difficile)
```

Format

List of two elements with names: geneID, COG

Source

http://www.ncbi.nlm.nih.gov/genome/proteins/535?project_id=57679

Examples

```
data(difficile)
```

DIMEplot *DIMEplot*

Description

This function provides visualization for the best fitted mixture model and enables a user to evaluate the fitted mixure model graphically. Inset in the plot shows a zoomed-in plot of individual components of the model. This is relevant only if data.type="data".

Usage

```
DIMEplot(x)
```

Arguments

x	This is a BOG object.

ecoli *Escherichia coli O157:H7 str. Sakai*

Description

This is protein details for Escherichia coli O157:H7 str. Sakai

Usage

```
data(ecoli)
```

Format

List of two elements with names: geneID, COG

Source

http://www.ncbi.nlm.nih.gov/genome/proteins/167?project_id=57781

Examples

```
data(ecoli)
```

francisella *Francisella tularensis subsp. holarctica LVS*

Description

This is protein details for Francisella tularensis subsp. holarctica LVS

Usage

```
data(francisella)
```

Format

List of two elements with names: geneID, COG

Source

http://www.ncbi.nlm.nih.gov/genome/proteins/511?project_id=58595

Examples

```
data(francisella)
```

gseaplot *gseaplot*

Description

This is the command to visualize the path of GSEA scores. The GSEA takes all genes into account by constructing a test statistic based on their local-fdr without pre-selection of threshold.

Usage

```
gseaplot(x, cat = NULL)
```

Arguments

x	This is a bog object.
cat	To plot an enrichment score behavior, a user needs to specify a cat argument from one of COG groups. The default of cat is set NULL to raise BOG specific warning instead of R-system warning when a user fogets to specify COG, which should match one of the COGs specified in the input file or in the built-in files. The warning message is ""BOG : Category needs to be specified.""

References

Subramian, A. et al.(2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS vol.102 no.43, 15545-15550.

hgplot *hgplot*

Description

A user can visualize the hypergeometric test results using this command. It display the most significant COG groups (adjusted p-value < 0.1) with observed and expected counts displayed.

Usage

```
hgplot(x)
```

Arguments

x	This is a BOG object.

Examples

```
bog=BOG(gsea=FALSE)
hgplot(bog)
```

internalhgplot *internalhgplot*

Description

This is an internal function so that it is not expected for a user to use it.

Usage

```
internalhgplot(stat)
```

Arguments

stat	This is an internally defined list object.

printGSEA *printGSEA*

Description

If gsea argument is set TRUE in BOG() command, this command print the outcome of GSEA test.

Usage

```
printGSEA(x)
```

Arguments

x	It is a BOG object.

printHG *printHG*

Description

This command print the outcome of HG(Hypergeometric) test.

Usage

```
printHG(x)
```

Arguments

x	It is a BOG object.

Examples

```
bog=BOG(data="anthracis_iron",cog.file="anthracis")
printHG(bog)
```

printRANK *printRANK*

Description

This command print the outcome of RANK(Mann-Whiteny) test.

Usage

```
printRANK(x)
```

Arguments

x	It is a BOG object.

Examples

```
bog=BOG(data="anthracis_iron",cog.file="anthracis")
printRANK(bog)
```

Index