Package ‘sRAP’

April 10, 2015

Type Package

Title Simplified RNA-Seq Analysis Pipeline

Version 1.6.0

Date 2013-08-21

Author Charles Warden

Maintainer Charles Warden <`cwarden45@gmail.com`>

Depends WriteXLS

Imports gplots, pls, ROCR, qvalue

Description This package provides a pipeline for gene expression analysis (primarily for RNA-Seq data). The normalization function is specific for RNA-Seq analysis, but all other functions (Quality Control Figures, Differential Expression and Visualization, and Functional Enrichment via BD-Func) will work with any type of gene expression data.

License GPL-3

LazyLoad yes

biocViews GeneExpression, RNAseq, Microarray, Preprocessing, QualityControl, Statistics, DifferentialExpression, Visualization, GeneSetEnrichment, GO

R topics documented:

`bdfunc.enrichment.human`

BD-Func Enrichment File for Human Gene Symbols

Description

Contains genes lists of paired up-regulated and down-regulated genes. Gene Lists come from Gene Ontology (GO [1]) and MSigDB [2].

BD-Func [3] will compare the expression patterns for the up-regulated genes to the down-regulated genes.

Usage

```
data(bdfunc.enrichment.human)
```

Source

GO Gene Lists: http://www.geneontology.org/GO.downloads.annotations.shtml MSigDB Gene Lists: http://www.broadinstitute.org/gsea/downloads.jsp

References

[1] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, and Sherlock G.(2000). Gene Ontology: tool for the unification of biology *Nat Genet*, 25:25-29

[2] Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, and Mesirov JP.(2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27:1739-1740.

[3] Warden CD, Kanaya N, Chen S, and Yuan Y-C. (2013). BD-Func: A Streamlined Algorithm for Predicting Activation and Inhibition of Pathways. *peerJ*, 1:e159

`bdfunc.enrichment.mouse`

BD-Func Enrichment File for Mouse Gene Symbols

Description

Contains genes lists of paired up-regulated and down-regulated genes. Gene Lists come from Gene Ontology (GO [1]).

BD-Func [2] will compare the expression patterns for the up-regulated genes to the down-regulated genes.

Usage

```
data(bdfunc.enrichment.mouse)
```

Source

GO Gene Lists: http://www.geneontology.org/GO.downloads.annotations.shtml

References

[1] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, and Sherlock G.(2000). Gene Ontology: tool for the unification of biology *Nat Genet*, 25:25-29

[2] Warden CD, Kanaya N, Chen S, and Yuan Y-C. (2013). BD-Func: A Streamlined Algorithm for Predicting Activation and Inhibition of Pathways. *peerJ*, 1:e159

`RNA.bdfunc.fc` *Functional Enrichment for a Table of Fold-Change Values*

Description

Bi-Directional FUNCtional enrichment [1] compares expression values for up- and down-regulated genes are compared for at least one gene set, using fold-change valuesvalues. Gene sets are already defined for human and mouse gene symbols. All other gene sets must be specified by the user. The user can optionally output density plots to visualze enrichment scores across samples in different groups.

Usage

```
RNA.bdfunc.fc(stat.table, project.name, project.folder, species = NULL, enrichment.file = NULL, p.meth
```

Arguments

`stat.table`	Table of fold-change values. This assumes that the variable hypothesized to control gene expression can be defined as two groups, with a specified reference group. Gene symbols are in the first column. Fold-change values are in the second column. This file is automatically created by the RNA.deg function.
`project.name`	Name for sRAP project. This determines the names for output files.
`project.folder`	Folder for sRAP output files
`species`	Name for species used for analysis. If species is set to "human" or "mouse," then pre-defined gene lists provided by the sRAP package are used. The default human gene list is created from gene ontology [2] and MSigDB [3] databases. The default mouse gene list is created from the gene ontology [2] database.
`enrichment.file`	Table of gene lists including up- and down-regulated genes. This is only necessary when defining a custom species. This parameter is ignored when the species is set to "human" or "mouse"

`p.method`	Method for calculating p-values "t-test" (Default) = t-test between up-regulated and down-regualted genes "mann-whitney" = Non-parametric Mann-Whitney U test between up-and down-regulated genes "ks" = Kolmogorov-Smirnov test between up- and down-regulated genes
`p.adjust.method`	Method for calculating false discovery rate (FDR): "fdr" (Default)= B-H "Step-Up" FDR [4] "q-value" = Storey q-value [5] "none" = use unadjusted p-value without multiple hypothesis correction
`plot.flag`	Logical value: Should density plots be created for all gene sets?

Author(s)

Charles Warden <cwarden@coh.org>

References

[1] Warden CD, Kanaya N, Chen S, and Yuan Y-C. (2013). BD-Func: A Streamlined Algorithm for Predicting Activation and Inhibition of Pathways. *peerJ*, 1:e159

[2] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, and Sherlock G.(2000). Gene Ontology: tool for the unification of biology *Nat Genet*, 25:25-29

[3] Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, and Mesirov JP.(2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27:1739-1740.

[4] Benjamini Y, and Hochberg Y.(1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B*, 57:289-300.

[5] Storey JD, and Tibshirani R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100:9440-9445.

See Also

sRAP goes through an entire analysis for an example dataset provided with the sRAP package.

Examples

```
library("sRAP")

dir <- system.file("extdata", package="sRAP")
expression.table <- file.path(dir,"MiSeq_cufflinks_genes_truncate.txt")
sample.table <- file.path(dir,"MiSeq_Sample_Description.txt")
project.folder <- getwd()
project.name <- "MiSeq"

expression.mat <- RNA.norm(expression.table, project.name, project.folder)

stat.table <- RNA.deg(sample.table, expression.mat, project.name, project.folder, box.plot=FALSE, ref.group=TRUE
```

```
RNA.bdfunc.fc(stat.table, project.name, plot.flag=FALSE, project.folder, species="human")
```

`RNA.bdfunc.signal` *Functional Enrichment for a Table of Normalized Gene Expression Values*

Description

Bi-Directional FUNCtional enrichment [1] compares expression values for up- and down-regulated genes are compared for at least one gene set, using normalized expression values. Gene sets are already defined for human and mouse gene symbols. All other gene sets must be specified by the user. The user can optionally output box-plots to visualze enrichment scores across samples in different groups.

Usage

```
RNA.bdfunc.signal(expression.table, sample.file, project.name, project.folder, species = NULL, enrich
```

Arguments

`expression.table`	Data frame with genes in columns and samples in rows. Data should be log2 transformed. The RNA.norm function automatically creates this file.
`sample.file`	Tab-delimited text file providing group attributions for all samples considered for analysis.
`project.name`	Name for sRAP project. This determines the names for output files.
`project.folder`	Folder for sRAP output files
`species`	Name for species used for analysis. If species is set to "human" or "mouse," then pre-defined gene lists provided by the sRAP package are used. The default human gene list is created from gene ontology [2] and MSigDB [3] databases. The default mouse gene list is created from the gene ontology [2] database.
`enrichment.file`	Table of gene lists including up- and down-regulated genes. This is only necessary when defining a custom species. This parameter is ignored when the species is set to "human" or "mouse".
`p.method`	Method for calculating p-values "t-test" (Default) = t-test between up-regulated and down-regualted genes "mann-whitney" = Non-parametric Mann-Whitney U test between up-and down-regulated genes "ks" = Kolmogorov-Smirnov test between up- and down-regulated genes
`p.adjust.method`	Method for calculating false discovery rate (FDR): "fdr" (Default)= B-H "Step-Up" FDR [4] "q-value" = Storey q-value [5] "none" = use unadjusted p-value without multiple hypothesis correction

`plot.flag`	Logical value: Should box-plots be created for all gene sets? If primary variable is two groups called "positive" and "negative", this value also determines if ROC plot will be created.
`color.palette`	Colors for primary variable (specified in the second column of the sample file). If method is set to "t-test," this variable is ignored. In this special case, groups with an average t-test statistic above 2 are colored red, groups with an average t-test statistic below -2 are colored green, and all other groups are colored grey.

Author(s)

Charles Warden <cwarden@coh.org>

References

[1] Warden CD, Kanaya N, Chen S, and Yuan Y-C. (2013). BD-Func: A Streamlined Algorithm for Predicting Activation and Inhibition of Pathways. *peerJ*, 1:e159

[2] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, and Sherlock G.(2000). Gene Ontology: tool for the unification of biology *Nat Genet*, 25:25-29

[3] Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, and Mesirov JP.(2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27:1739-1740.

[4] Benjamini Y, and Hochberg Y.(1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B*, 57:289-300.

[5] Storey JD, and Tibshirani R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100:9440-9445.

See Also

sRAP goes through an entire analysis for an example dataset provided with the sRAP package.

Examples

```
library(”sRAP”)

dir <- system.file(”extdata”, package=”sRAP”)
expression.table <- file.path(dir,”MiSeq_cufflinks_genes_truncate.txt”)
sample.table <- file.path(dir,”MiSeq_Sample_Description.txt”)
project.folder <- getwd()
project.name <- ”MiSeq”

expression.mat <- RNA.norm(expression.table, project.name, project.folder)

stat.table <- RNA.deg(sample.table, expression.mat, project.name, project.folder, box.plot=FALSE, ref.group=TRUE,

RNA.bdfunc.signal(expression.mat, sample.table, plot.flag=FALSE, project.name, project.folder, species=”human”)
```

`RNA.deg` *Differential Expression Statistics*

Description

Provides a table of diffrerenitally expressed genes (in .xlsx format) as well as differential expression statistics for all genes (in .xlsx format as well as returned data frame). Function automatically creates a heatmap for differentially expressed genes and user can optionally also create box-plots for each individual differentially expressed gene. The efficacy of this protocol is described in [1].

Output files will be created in the "DEG" and "Raw_Data" subfolders.

Usage

```
RNA.deg(sample.file, expression.table, project.name, project.folder, log2.fc.cutoff = 0.58, pvalue.cut
```

Arguments

`sample.file`	Tab-delimited text file providing group attributions for all samples considered for analysis.
`expression.table`	Data frame with genes in columns and samples in rows. Data should be log2 transformed. The RNA.norm function automatically creates this file.
`project.name`	Name for sRAP project. This determines the names for output files.
`project.folder`	Folder for sRAP output files
`log2.fc.cutoff`	If the primary variable contains two groups with a specified reference, this is the cut-off to define differentially expressed genes (default = 1.5, on a linear scale). Otherwise, this variable is ignored
`pvalue.cutoff`	Minimum p-value to define differentially expressed genes
`fdr.cutoff`	Minimum false discovery rate (FDR) to define differentially expressed genes.
`box.plot`	A logical value: Should box-plots be created for all differenitally expressed genes? If TRUE, then box-plots will be created in a separate subfolder.
`ref.group`	A logical value: Is the primary variable 2 groups, with a reference group?
`ref`	If the primary variable contains two groups (indicated by ref.group = FALSE), this is the reference used to calculate fold-change values (so, the mean expression for the reference group is substracted from the treatment group). Otherwise, this variable is ignored
`method`	Method for calculating p-values: "lm" (Default) = linear regression "aov" = ANOVA
`color.palette`	Colors for primary variable (specified in the second column of the sample file). If the primary variable is a continuous variable, this parameter is ignored.

Value

Data frame containing differential expression statistics.

First column contains gene name.

If the primary variable contains two groups (with a specified reference), then fold-change values are provided in the second column.

P-values and FDR values are provided for each variable in subsequent columns, starting with the primary variable.

Author(s)

Charles Warden <cwarden@coh.org>

References

[1] Warden CD, Yuan Y-C, and Wu X. (2013). Optimal Calculation of RNA-Seq Fold-Change Values. *in preparation*,

See Also

sRAP goes through an entire analysis for an example dataset provided with the sRAP package.

Examples

```
library("sRAP")

dir <- system.file("extdata", package="sRAP")
expression.table <- file.path(dir,"MiSeq_cufflinks_genes_truncate.txt")
sample.table <- file.path(dir,"MiSeq_Sample_Description.txt")
project.folder <- getwd()
project.name <- "MiSeq"

expression.mat <- RNA.norm(expression.table, project.name, project.folder)

stat.table <- RNA.deg(sample.table, expression.mat, project.name, project.folder, box.plot=FALSE, ref.group=TRUE,
```

`RNA.norm` *Normalization for RNA-Seq Data*

Description

Takes a table of RPKM (Read Per Kilobase per Million reads [1]) gene expression values. Rounds RPKM values based upon RPKM.cutoff(to avoid bias from low-coverage genes), and then performs a log2 transformation of the data (so that the data more closely follows a normal distribution). The efficacy of this protocol is described in [2].

Output files will be created in the "Raw_Data" subfolder.

Usage

```
RNA.norm(input.file, project.name, project.folder, RPKM.cutoff = 0.1)
```

Arguments

`input.file`	Table of RPKM expression values. Genes are represented in columns. Samples are represented in rows.
`project.name`	Name for sRAP project. This determines the names for output files.
`project.folder`	Folder for sRAP output files
`RPKM.cutoff`	Cut-off for rounding RKPM expression values. If the default of 0.1 is used, genes with expression values consistently below 0.1 will essentially be ignored.

Value

Data frame of normalized expression values on a log2 scale.

Just like the input table, genes are represented on columns, samples are represented in rows.

This data frame is used for quality control and differential expression analysis.

Author(s)

Charles Warden <cwarden@coh.org>

References

[1] Mortazavi A, Williams BA, McCue K, Schaeffer L, and Wold B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth*, 5:621-628.

[2] Warden CD, Yuan Y-C, and Wu X. (2013). Optimal Calculation of RNA-Seq Fold-Change Values. *in preparation*,

See Also

sRAP goes through an entire analysis for an example dataset provided with the sRAP package.

Examples

```
library("sRAP")

dir <- system.file("extdata", package="sRAP")
expression.table <- file.path(dir,"MiSeq_cufflinks_genes_truncate.txt")
sample.table <- file.path(dir,"MiSeq_Sample_Description.txt")
project.folder <- getwd()
project.name <- "MiSeq"

expression.mat <- RNA.norm(expression.table, project.name, project.folder)
```

`RNA.qc` *Gene Expression Quality Control Statistics*

Description

Provides descriptive statistics (median, top/bottom quartiles, mininum,maximum), sample histograms and box-plot, sample dendrogram, principal component analysis plot.

Output files will be created in the "QC" subfolder.

Usage

```
RNA.qc(sample.file, expression.table, project.name, project.folder, plot.legend = TRUE, color.palette
```

Arguments

`sample.file`	Tab-delimited text file providing group attributions for all samples considered for analysis.
`expression.table`	Data frame with genes in columns and samples in rows. Data should be log2 transformed. The RNA.norm function automatically creates this file.
`project.name`	Name for sRAP project. This determines the names for output files.
`project.folder`	Folder for sRAP output files
`plot.legend`	A logical value: Should legend be plotted within QC figures?
`color.palette`	Colors for primary variable (specified in the second column of the sample file). If the primary variable is a continuous variable, this parameter is ignored.

Author(s)

Charles Warden <cwarden@coh.org>

See Also

sRAP goes through an entire analysis for an example dataset provided with the sRAP package.

Examples

```
library("sRAP")
library("WriteXLS")

dir <- system.file("extdata", package="sRAP")
expression.table <- file.path(dir,"MiSeq_cufflinks_genes_truncate.txt")
sample.table <- file.path(dir,"MiSeq_Sample_Description.txt")
project.folder <- getwd()
project.name <- "MiSeq"

expression.mat <- RNA.norm(expression.table, project.name, project.folder)
```

```
RNA.qc(sample.table, expression.mat, project.name, project.folder, plot.legend=FALSE, color.palette=c("green","o
```

Index