Package 'SCBN'

November 15, 2024

Type Package

Title A statistical normalization method and differential expression analysis for RNA-seq data between different species

Version 1.24.0

Author Yan Zhou

Maintainer Yan Zhou <2160090406@email.szu.edu.cn>

Description This package provides a scale based normalization (SCBN) method to identify genes with differential expression between different species. It takes into account the available knowledge of conserved orthologous genes and the hypothesis testing framework to detect differentially expressed orthologous genes. The method on this package are described in the article 'A statistical normalization method and differential expression analysis for RNA-seq data between different species' by Yan Zhou, Jiadi Zhu, Tiejun Tong, Junhui Wang, Bingqing Lin, Jun Zhang (2018, pending publication).

License GPL-2

Encoding UTF-8

LazyData true

Depends R (>= 3.5.0)

Suggests knitr,rmarkdown,BiocStyle,BiocManager

VignetteBuilder knitr

RoxygenNote 6.0.1

Imports stats

biocViews DifferentialExpression, GeneExpression, Normalization

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

git_branch RELEASE_3_20

git_last_commit b745577

git_last_commit_date 2024-10-29

Repository Bioconductor 3.20

Date/Publication 2024-11-14

Contents

`generateDataset` *Generate simulation data for different species*

Description

To generate RNA-seq genes between different species.

Usage

```
generateDataset(commonTags=15000, uniqueTags=c(1000, 3000),
                        unmapped=c(4000, 2000),group=c(1, 2),
                        libLimits=c(.9, 1.1)*1e6, empiricalDist=NULL,
                        genelength, randomRate=1/100,
                        pDifferential=.05, pUp=.5, foldDifference=2)
```

Arguments

`commonTags`	The number of genes have the same expression level.
`uniqueTags`	The number of genes only expressed in one species.
`unmapped`	The number of genes only in one species.
`group`	The number of species.
`libLimits`	The limits for two species.
`empiricalDist`	Define where to take random sample from (empirical distribution OR random exponential), if NULL, the reads take from random exponential.
`genelength`	A vector of gene length for each gene of two species.
`randomRate`	The parameter for exponential distribution.
`pDifferential`	The propotion of differential expression genes.
`pUp`	The probably for the reads in first species fold than the second species.
`foldDifference`	The fold for fold expression genes.

Value

list(.) A list of output, "DATAN" represents the read counts for the first species, "DATAM" represents the read counts for the second species, "trueFactors" represents the true scaling factor for data, "group" represents the number of species, "libSizes" represents the library size for data, "differentialInd" represents the ID for differential expression genes, "commonInd" represents the ID for common expression genes.

Examples

```
data(orthgenes)
orthgenes[, 6:9] <- round(orthgenes[, 6:9])
orthgenes1 <- orthgenes[!(is.na(orthgenes[,6])|is.na(orthgenes[,7])|
                      is.na(orthgenes[,8])|is.na(orthgenes[,9])), ]
sim_data <- generateDataset(commonTags=5000, uniqueTags=c(100, 300),
                            unmapped=c(400, 200),group=c(1, 2),
                            libLimits=c(.9, 1.1)*1e6,
                            empiricalDist=orthgenes1[, 6],
                            genelength=orthgenes1[, 2], randomRate=1/100,
                            pDifferential=.05, pUp=.5, foldDifference=2)
```

`Iter_optimal` *Iteration to find the optimal value A iteration process to compute the normalization factor to identify difference expression(DE) of genes between different species*

Description

Iteration to find the optimal value A iteration process to compute the normalization factor to identify difference expression(DE) of genes between different species

Set the initial value Using median method to compute the normalization factor to identify difference expression (DE) of genes between different species

Compute the false discovery rate Compute the p-value for each orthologous genes between different species

Usage

```
Iter_optimal(scale, orth_gene, hkind, a)

MediancalcNorm(orth_gene, hkind)

sageTestNew(x, y, lengthx, lengthy, n1, n2, scale)
```

Arguments

`scale`	A value for normalization factor.
`orth_gene`	Matrix or data.frame containing read counts and gene length for each orthologous gene between different species. The first and third column containing gene length, the second and the fourth column containing read counts.
`hkind`	A vector shows conserved genes position in orthologous genes.
`a`	P-value cutoff in iteration process to find the optimal normalization factor.
`x`	The read counts for the first species.
`y`	The read counts for the second species.
`lengthx`	The gene length for the first species.
`lengthy`	The gene length for the second species.
`n1`	The total read counts for the first species.
`n2`	The total read counts for the second species.

Value

factor Computed normalization factor.

scale Computed Normalization factor.

p_value P-values for each orthologous genes between different species.

Functions

- `Iter_optimal`: obtain the optimal normalization value.
- `MediancalcNorm`: get scaling factor for different species.
- `sageTestNew`: obtain the p-value for each orthologous genes between different species.

References

Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, et al. The evolution of gene expression levels in mammalian organs. Nature. 2011;478:343-348.

Examples

```
data(sim_data)
scale <- MediancalcNorm(orth_gene=sim_data, hkind=1:1000)
Iter_optimal(scale=scale, orth_gene=sim_data, hkind=1:1000, a=0.05)
data(sim_data)
MediancalcNorm(orth_gene=sim_data, hkind=1:1000)
data(sim_data)
orth_gene <- sim_data
hkind <- 1:1000
scale <- MediancalcNorm(orth_gene=orth_gene, hkind=hkind)
x <- orth_gene[, 2]
y <- orth_gene[, 4]
lengthx <- orth_gene[, 1]
lengthy <- orth_gene[, 3]
n1 <- sum(x)
n2 <- sum(y)
p_value <- sageTestNew(x, y, lengthx, lengthy, n1, n2, scale)
```

`orthgenes` *A real dataset of orthologous genes between the different species.*

Description

This data set gives 27821 orthologous genes which include read counts and genes length between the two different species.

Usage

```
orthgenes
```

Format

A data.frame containing 27821 orthologous genes.

Source

Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, et al. The evolution of gene expression levels in mammalian organs. Nature. 2011;478:343-348.

`SCBN` *Compute the normalization factor to identify difference expression (DE) of genes between different species*

Description

To normalize read counts and identify difference expression(DE) of orthologous genes between different species.

Usage

```
SCBN(orth_gene, hkind, a=0.05)
```

Arguments

`orth_gene`	Matrix or data.frame containing read counts and gene length for each orthologous gene between different species. The first and third column containing gene length, the second and the fourth column containing read counts.
`hkind`	A vector shows conserved genes position in orthologous genes.
`a`	P-value cutoff in iteration process to find the optimal normalization factor.

Value

list(.) A list of computed normalization factors, "median_val" represents foctors computed by median methods, "scbn_val" represents factors computed by SCBN methods.

Examples

```
data(sim_data)
SCBN(orth_gene=sim_data, hkind=1:1000, a=0.05)
```

`sim_data` *A simulation dataset of orthologous genes between the different species.*

Description

This data set gives 4149 orthologous genes which include read counts and genes length between the two different species.

Usage

```
sim_data
```

Format

A data.frame containing 4149 orthologous genes.

Source

Zhou Y, Zhu JD, Tong TJ, Wang JH, Lin BQ, Zhang J(2018, pending publication). A Novel Normalization Method and Differential Expression Analysis of RNA-seq Data between Different Species.

Index