Package 'empiricalFDR.DESeq2'

May 27, 2015

Type Package

Title Simulation-Based False Discovery Rate in RNA-Seq

Version 1.0.3

Date 2015-05-26

Author Mikhail V. Matz

Maintainer Mikhail V. Matz <matz@utexas.edu>

Description

Auxiliary functions for the DESeq2 package to simulate read counts according to the null hypothesis (i.e., with empirical sample size factors, per-gene total counts and dispersions, but without effects of predictor variables) and to compute the empirical false discovery rate.

License GPL-3

Depends DESeq2,GenomicRanges

NeedsCompilation no

Repository CRAN

Date/Publication 2015-05-27 08:55:36

R topics documented:

empiricalFDR.DESeq2-package

Simulation-Based False Discovery Rate in RNA-Seq

Description

Auxiliary functions for the DESeq2 package to simulate read counts according to the null hypothesis (i.e., with empirical sample size factors, per-gene total counts and dispersions, but without effects of predictor variables) and to compute the empirical false discovery rate.

Details

Package:	empiricalFDR.DESeq2
Type:	Package
Version:	1.0.2
Date:	2015-04-01
License:	GPL-3

The key function is simulateCounts, which takes a fitted DESeq2 data object as an input and returns a simulated data object with the same sample size factors, total counts and dispersions for each gene as in real data, but without the effect of predictor variables. Functions fdrTable, fdrBiCurve and empiricalFDR compare the DESeq2 results obtained for the real and simulated data, compute the empirical false discovery rate (the ratio of the number of differentially expressed genes detected in the simulated data and their number in the real data) and plot the results.

Author(s)

Mikhail V. Matz Maintainer: Mikhail V. Matz <matz@utexas.edu>

References

R. M. Wright, G. V. Aglyamova, E. Meyer and M. V. Matz (2015) Local and systemic gene expression responses to a white-syndrome-like disease in a reef-bulding coral, Acropora hyacinthus.

Examples

```
dds = makeExampleDESeqDataSet(betaSD=1, n=100)
dds = DESeq(dds)
sims = simulateCounts(dds)
sims = DESeq(sims)
res = results(dds)
sim.res=results(sims)

# how similar is the simulation to real data?
plot(sizeFactors(sims)~sizeFactors(dds))
```

```
plot(log(dispersions(sims),10)~log(dispersions(dds),10))

# computing and plotting empirical FDR
fdrt = fdrTable(res$pvalue,sim.res$pvalue)
fdrBiCurve(fdrt,maxLogP=4,main="DEG discovery rates")
efdr = empiricalFDR(fdrt,plot=TRUE,main="False discovery rate")

# how many genes pass empirical 0.1 FDR cutoff?
table(res$pvalue<efdr)

# how many genes pass multiplicity-corrected 0.1 FDR cutoff?
table(res$padj<0.1)
```

`empiricalFDR` *Computing the p-value cutoff to achieve a given FDR.*

Description

This function calculates the cutoff at which a particular false discovery rate is observed using loess smoothing and interpolation.

Usage

```
empiricalFDR(fdr.table, FDR = 0.1, maxLogP = 5, plot = FALSE, span = 0.1, ...)
```

Arguments

`fdr.table`	The output of fdrTable(): a dataframe listing p-value cutoffs and the number of null hypothesis rejections at each cutoff in the real and simulated datasets.
`FDR`	The target false discovery rate.
`maxLogP`	Maximal negative decimal logarithm of the p-value for plotting (the default, 5, implies the data for p-values better than 10e-5 will not be plotted)
`plot`	Whether to produce a plot.
`span`	span parameter for loess smoothing.
`...`	Additional plotting parameters.

Value

The function returns a single value, which is the p-value cutoff at which the target FDR is observed. With plot = TRUE, also plots the observed experimental FDRs with the loess smoother.

Author(s)

Mikhail V. Matz

References

R. M. Wright, G. V. Aglyamova, E. Meyer and M. V. Matz (2015) Local and systemic gene expression responses to a white-syndrome-like disease in a reef-bulding coral, Acropora hyacinthus.

Examples

```
dds = makeExampleDESeqDataSet(betaSD=1, n=100)
dds = DESeq(dds)
sims = simulateCounts(dds)
sims = DESeq(sims)
res = results(dds)
sim.res=results(sims)

# how similar is the simulation to real data?
plot(sizeFactors(sims)~sizeFactors(dds))
plot(log(dispersions(sims),10)~log(dispersions(dds),10))

# computing and plotting empirical FDR
fdrt = fdrTable(res$pvalue,sim.res$pvalue)
fdrBiCurve(fdrt,maxLogP=4,main="DEG discovery rates")
efdr = empiricalFDR(fdrt,plot=TRUE,main="False discovery rate")

# how many genes pass empirical 0.1 FDR cutoff?
table(res$pvalue<efdr)

# how many genes pass multiplicity-corrected 0.1 FDR cutoff?
table(res$padj<0.1)
```

`fdrBiCurve` *Plots the numbers of null hypothesis rejections*

Description

Plots the numbers of null hypothesis rejections in real data and data simulated under null hypothesis (false positives)

Usage

```
fdrBiCurve(fdr.table, maxLogP = 5, ...)
```

Arguments

`fdr.table`	The output of fdrTable(): a dataframe listing p-value cutoffs and the number of null hypothesis rejections at each cutoff in the real and simulated datasets.
`maxLogP`	Maximal negative decimal logarithm of the p-value for plotting (the default, 5, implies the data for p-values better than 10e-5 will not be plotted)
`...`	Additional plotting parameters.

Value

The plot is designed to ascertain that the number of discoveries in real data (black line) indeed exceeds the number of false positives (red line) across the range of p-value cutoffs. The grey dotted line is the number of discoveries expected under uniform distribution of p-values.

Author(s)

Mikhail V. Matz

References

R. M. Wright, G. V. Aglyamova, E. Meyer and M. V. Matz (2015) Local and systemic gene expression responses to a white-syndrome-like disease in a reef-bulding coral, Acropora hyacinthus.

Examples

```
dds = makeExampleDESeqDataSet(betaSD=1, n=100)
dds = DESeq(dds)
sims = simulateCounts(dds)
sims = DESeq(sims)
res = results(dds)
sim.res=results(sims)

# how similar is the simulation to real data?
plot(sizeFactors(sims)~sizeFactors(dds))
plot(log(dispersions(sims),10)~log(dispersions(dds),10))

# computing and plotting empirical FDR
fdrt = fdrTable(res$pvalue,sim.res$pvalue)
fdrBiCurve(fdrt,maxLogP=4,main="DEG discovery rates")
efdr = empiricalFDR(fdrt,plot=TRUE,main="False discovery rate")

# how many genes pass empirical 0.1 FDR cutoff?
table(res$pvalue<efdr)

# how many genes pass multiplicity-corrected 0.1 FDR cutoff?
table(res$padj<0.1)
```

---

fdrTable *Computes false discovery rates for a series of p-value cutoffs.*

Description

Given vectors of p-values from real data and data simulated under a null hypothesis, produces a table listing the number of null hypothesis rejections under a range of p-value cutoffs in real and simulated data.

Usage

```
fdrTable(real.p, sim.p)
```

Arguments

real.p	Vector of p-values in the real data.
sim.p	Vector of p-values in the simulated data.

Value

The function returns a dataframe listing p-value cutoffs, in steps of 0.1 on the decimal log scale, the number of null hypothesis rejections at each cutoff in real and simulated datasets, and their ratio (the false discovery rate).

Author(s)

Mikhail V. Matz

References

R. M. Wright, G. V. Aglyamova, E. Meyer and M. V. Matz (2015) Local and systemic gene expression responses to a white-syndrome-like disease in a reef-bulding coral, Acropora hyacinthus.

Examples

```
dds = makeExampleDESeqDataSet(betaSD=1, n=100)
dds = DESeq(dds)
sims = simulateCounts(dds)
sims = DESeq(sims)
res = results(dds)
sim.res=results(sims)

# how similar is the simulation to real data?
plot(sizeFactors(sims)~sizeFactors(dds))
plot(log(dispersions(sims),10)~log(dispersions(dds),10))

# computing and plotting empirical FDR
fdrt = fdrTable(res$pvalue,sim.res$pvalue)
fdrBiCurve(fdrt,maxLogP=4,main="DEG discovery rates")
efdr = empiricalFDR(fdrt,plot=TRUE,main="False discovery rate")

# how many genes pass empirical 0.1 FDR cutoff?
table(res$pvalue<efdr)

# how many genes pass multiplicity-corrected 0.1 FDR cutoff?
table(res$padj<0.1)
```

`simulateCounts` *Simulating RNA-seq read counts*

Description

This function takes a fitted DESeq2 data object as an input and returns a simulated data object with the same sample size factors, total counts and dispersions for each gene as in real data, but without the effect of predictor variables.

Usage

```
simulateCounts(deseq.object)
```

Arguments

`deseq.object`	DESeq2 data object, with estimated size factors and dispersions (output of DESeq() function).

Details

For each gene, the total counts are randomly resampled into different samples. The estimated per-gene dispersions are understood as the suare of the coefficient of variation and used to simulate random deviations in per-sample assignment probability. The probabilities of per-sample assignment are also weighted by the empirical sample size factors.

Value

DESeq2 data object

Author(s)

Mikhail V. Matz

References

R. M. Wright, G. V. Aglyamova, E. Meyer and M. V. Matz (2015) Local and systemic gene expression responses to a white-syndrome-like disease in a reef-bulding coral, Acropora hyacinthus.

Examples

```
dds = makeExampleDESeqDataSet(betaSD=1, n=100)
dds = DESeq(dds)
sims = simulateCounts(dds)
sims = DESeq(sims)
res = results(dds)
sim.res=results(sims)

# how similar is the simulation to real data?
```

```
plot(sizeFactors(sims)~sizeFactors(dds))
plot(log(dispersions(sims),10)~log(dispersions(dds),10))

# computing and plotting empirical FDR
fdrt = fdrTable(res$pvalue,sim.res$pvalue)
fdrBiCurve(fdrt,maxLogP=4,main="DEG discovery rates")
efdr = empiricalFDR(fdrt,plot=TRUE,main="False discovery rate")

# how many genes pass empirical 0.1 FDR cutoff?
table(res$pvalue<efdr)

# how many genes pass multiplicity-corrected 0.1 FDR cutoff?
table(res$padj<0.1)
```

Index