Package 'sSeq'

March 18, 2018

Type Package

Title Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size

Version 1.16.0

Date 2013-04-17

Author Danni Yu <dyu@purdue.edu>, Wolfgang Huber <whuber@embl.de> and Olga Vitek <ovitek@purdue.edu>

Maintainer Danni Yu <dyu@purdue.edu>

Depends R (>= 3.0), caTools, RColorBrewer

Description The purpose of this package is to discover the genes that are differentially expressed between two conditions in RNA-seq experiments. Gene expression is measured in counts of transcripts and modeled with the Negative Binomial (NB) distribution using a shrinkage approach for dispersion estimation. The method of moment (MM) estimates for dispersion are shrunk towards an estimated target, which minimizes the average squared difference between the shrinkage estimates and the initial estimates. The exact per-gene probability under the NB model is calculated, and used to test the hypothesis that the expected expression of a gene in two conditions identically follow a NB distribution.

License GPL (>= 3)

LazyLoad yes

biocViews RNASeq

NeedsCompilation no

R topics documented:

`sSeq-package` *Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size*

Description

This package is to discover the genes that differentially expressed between two conditions based on RNA-seq experiments. Gene expression is measured in counts of transcripts and modeled with the Negative Binomial (NB) distribution using a shrinkage approach for dispersion estimation. The method of moment (MM) estimates for dispersion are simply shrunk toward a target, which minimizes the average squared difference between the shrinkage estimates and the initial estimates. The exact per-gene probability under the NB model is calculated, and used to test the hypothesis that the expected expression of a gene in two conditions are not different.

Details

Package:	sSeq
Type:	Package
Version:	1.0
Date:	2013-02-25

Author(s)

Danni Yu <dyu@purdue.edu>, Wolfgang Huber <whuber@embl.de> and Olga Vitek <ovitek@purdue.edu>

References

Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size

Examples

```
#load a simulated data that includes a count table
data("countsTable")
```

```
#calculate the p-values using the shrinkage approach.
conds <- c("A", "B")
resJS <- nbTestSH( countsTable, conds, "A", "B")
```

---

countsTable *An Example Simulation Data*

Description

A subset of simulated data. It is used as an example for running some functions in this package.

Usage

```
data(countsTable)
```

Format

The format is: num [1:10000, 1:2] 90 155 13347 254 228 ... - attr(*, "dimnames")=List of 2 ..$: chr [1:10000] "1_FALSE" "2_FALSE" "3_TRUE" "4_FALSE" ... ..$: chr [1:2] "A1" "B1"

Details

A simulation counts table.

Examples

```
data(countsTable)
head(countsTable)
```

---

drawMA_vol *Draw MA Plot and Volcano Plot*

Description

Based on the count table and the p-values, this function can be used to draw a MA plot of the log2 ratios versus the log2 averages upon means of gene expression in condition A and B, and a volcano plot of negative log2 p-values versus the log2 ratios.

Usage

```
drawMA_vol(y, groups2, pv, cutoff=NULL, xlab1="(log2(A)+log2(B))/2",
           ylab1="log2(A)-log2(B)", tt1="MA plot", tt2="volcano plot",
           log2FoldChange =NULL, col1=c("black","red"))
```

Arguments

y	A count table in which row represents genes and column represents samples.
groups2	A vector indicates the two groups information of samples. It must match to the column in the count table, which is the input for y. For example, groups2=c("A","A","B","B") when the first two columns in the count table are the two samples from condition A, and the second two columns in the count table are the two samples from condition B.
pv	A vector of per-gene p-values based on the count table. The order of genes in pv does matter. It must be the same as the order of genes in the count table.
cutoff	A value used as a threshold for per-gene p-values to decide the genes that are differentially expressed between two conditions. If NULL, the cutoff value is calculated so that the red dots in the MA plot and volcano plot represent the first 5
xlab1	A character indicating the label of x axis in MA plot.
ylab1	A character indicating the label of y axis in MA plot.
tt1	A character indicating the title of the MA plot.
tt2	A character indicating the title of the volcano plot.
log2FoldChange	A vector of fold changes in log2 scale. It will be calculated automatically when "log2FoldChange=NULL".
col1	A vector with two values including the colors of points. The first color in "col1" is the color for the points that are non-differentially changed. The second value in "col1" is the color for the points that are differentially changed. The default is c("black", "red").

Examples

```
x <- matrix(rnorm(4000, 10), ncol=4)
px <- apply(x, 1, function(y){t.test(y[1:2], y[3:4])$p.value})
drawMA_vol(x, c("A","A","B","B"), px, cutoff=0.05)
```

---

ecdfAUC *Draw Empirical Cumulative Density Function (ECDF) plot*

Description

This function is used to draw Empirical CDF plot. It relies on the trapz function in the caTools package. A user needs to install the caTool library first.

Usage

```
ecdfAUC(dd, col.line=NULL, main="ECDF", cex.leg=1, drawRef=FALSE,
       rm1=FALSE, lineType=NULL, addLeg=TRUE, xlab="p-value",
       ylab="ECDF", cex.axis=1.5, cex.main=1.8, cex.lab=1.2,
      axis.padj=c(-1, 1), lab.padj=c(-1.5, 1), lwd=1, box.lwd=1.2)
```

Arguments

dd	A data frame of p-values in which a column represents the p-values or posterior probabilities resulted by a method.
col.line	A vector of color characters. The default is NULL and this function automatically assigns the color for each cover shown in the ECDF plot.
main	The title of the plot.
cex.leg	An integer specifying size of the legend in the the plot.
drawRef	If TRUE, then a gray 45 degree line will be added in the plot.
rm1	If users believe that the p-values equal to 1 belong to the different group of the others, and want to exclude them from the calculation of empirical CDF, then use rm1=TRUE.
lineType	A vector of integers indicating the type of lines used for the methods.
addLeg	If "TRUE" then a legend box with legend is added to the figure.
xlab	Label of x axis.
ylab	Label of y axis.
cex.axis	The size of labels on the axes.
cex.main	A characteristic string indicating the size of the main.
cex.lab	The size for the labels on x and y.
axis.padj	The perpendicular adjustment of ticks.
lab.padj	The perpendicular adjustment of labels for an axix.
lwd	The width of the line shown in a figure.
box.lwd	The width of the box line in a figure.

Examples

```
x<-data.frame(A=runif(100), B=rbeta(100, 0.5, 1.2))
ecdfAUC(x);
```

---

equalSpace *Calculate Grouped Shrinkage Estimates*

Description

This is an internal function. When the local mean-dispersion dependence is present, data can be separated into groups based on the means. The windows used to partition groups have equal width upon each other. The shrinkage (SH) estimates for dispersion will be calculated within each group. For example, when range of the per-gene mean is 1 and 3000, if data will be separated into 3 groups, then group 1 includes the genes having mean values between 1 and 1000, group 2 includes the genes having mean values between 1001 and 2000, and group 3 includes the genes having mean values between 2001 and 3000. The SH estimates will be calculated within each of the 3 groups, respectively.

Usage

```
equalSpace(y, x, numcls=1, propForSigma=c(0, 1), shrinkTarget=NULL,
          shrinkQuantile=0.975, vb=TRUE)
```

Arguments

y	A vector including the initial values that will be regularized. For example, it can be the per-gene method of moment (MM) estimates for dispersion based on the Negative Binomial distribution for the counts table.
x	A vector that will be used to separate data into groups. For example, it can be the per-gene averages for the counts table.
numcls	An integer that indicates the number of groups to be considered. The default value is 1.
propForSigma	A range vector between 0 and 1 that is used to select a subset of data. It helps users to make a flexible choice on the subset of data when they believe only part of data should be used to estimate the variation among per-gene dispersion. A default input propForSigma=c(0, 1) is recommended. It means that we want to use all the data to estimate the variation.
shrinkTarget	A value that represents the targeted point of stabilization for shrinkage estimates on dispersion. When shrinkTarget=NULL, the point of stabilization will be calculated according to the input of shrinkQuantile. If a numeric value is input for shrinkTarget, the shrinkQuantile argument will be ignored.
shrinkQuantile	A value between 0 and 1 that represents the target quantile point of stabilization for shrinkage estimates on dispersion. When a numeric value is not provided for shrinkTarget, the shrinkQuantile argument is used. The default value is NULL and means that the function will automatically estimate the point of stabilization based on the pattern of the average squared difference (ASD) between the initial method of moment (MM) estimates and the shrinkage (SH) estimates on dispersion.
vb	A logic value. When verbose=TRUE, the detail information will be printed in the console.

Value

This function returns a vector of shrinkage estimate on the basis of y.

Author(s)

Danni Yu

Examples

```
data("countsTable");

#calculate the row means;
rM <- rowMeans(countsTable);

#calculate the row variances;
rV <- rowVars(countsTable);

#calculate the method-of-moment estimation on dispersions;
disp <- (rV - rM)/rM^2;

#calculate SH estimates in 3 groups;
disp3 <- equalSpace(disp, rM, 3);
head(disp3);
```

exactNBtest1 *Perform only one exact test under the Negative Binomial modeling.*

Description

One exact test for only one gene.

Usage

```
exactNBtest1(kA, kB, mu, disp, sA=1, sB=1, rA=0.5, rB=0.5)
```

Arguments

kA	An integer matrix under condition A.
kB	An integer under matrix condition B.
mu	The expectation.
disp	The dispersion.
sA	The size factors under condition A.
sB	The size factors under condition B.
rA	Proportion of samples that are under condition A.
rB	Proportion of samples that are under condition B.

Value

pval	P-value.

Examples

```
exactNBtest1(100, 150, 125, 1.1)
```

getAdjustDisp *Calculate Shrinkage (SH) Estimates for Dispersion*

Description

In this shrinkage approach, the per-gene dispersion is considered as a variable in large dimensions. For example, if sequences of 30,000 genes are read in a RNA-seq experiment, then the dispersion variable is distributed in 30,000 dimensions. Firstly method-of-moment (MM) estimates on dispersion are calculated under the Negative Binomial (NB) modeling respectively for each gene. Those initial estimates are independently obtained in each dimension. Since RNA-seq experiments typically includes small number of samples (such as 1,2,3,4), the per-gene MM estimates are not reliable due to the limitation of sample size. We believe that there is a common variation shared across genes. The shrinkage approach regularizes per-gene dispersion estimates toward the common variation and produces robust estimates. Therefore in the second step, the MM estimates are shrunk towards an estimated target that minimizes the average squared difference (ASD) between the initial estimates and the shrinkage estimates.

Usage

```
getAdjustDisp(obs, propForSigma=c(0.5, 1), shrinkTarget=NULL,
              shrinkQuantile=NULL, verbose=TRUE)
```

Arguments

obs	A vector of initial estimates that are used to obtain the shrinkage (SH) estimates. The length of this vector must equal to the number of rows in the counts table. For example, the method-of-moment (MM) estimates for dispersion based on the Negative Binomial (NB) distribution are the initial estimates.
propForSigma	A range of percentiles that is used to identify a subset of data. It helps users to make a flexible choice on the subset of data when calculating variance of initial estimates among per-gene dispersion. A default input propForSigma=c(0, 1) is recommended. It means that we want to use all the data to estimate the variance.
shrinkTarget	A value that represents the targeted point of stabilization for shrinkage estimates on dispersion. When shrinkTarget=NULL, the point of stabilization will be calculated according to the input of shrinkQuantile. If a numeric value is input for shrinkTarget, the shrinkQuantile argument will be ignored.
shrinkQuantile	A value between 0 and 1 that represents the target quantile point of stabilization for shrinkage estimates on dispersion. When a numeric value is not provided for shrinkTarget, the shrinkQuantile argument is used. The default value is NULL and means that the function will automatically estimate the point of stabilization based on the pattern of the average squared difference (ASD) between the initial method of moment (MM) estimates and the shrinkage (SH) estimates on dispersion.
verbose	A logic value. When verbose=TRUE, the detail information will be printed in the console.

Value

adj	The SH estimates that shrink the input vector of obs toward the common information.
cpm	A data.frame that includes several summary statistics, such as the average and the variance of values in obs based on the subset controlled by the propForSigma argument.

Examples

```
data("countsTable");

#calculate the row means;
rM <- rowMeans(countsTable);

#calculate the row variances;
rV <- rowVars(countsTable);

#obtain an initial estimates;
disp <- (rV - rM)/rM^2;

#calculate the shrinkage estimates that shrink the initial estimates toward the common information;
dispSH <- getAdjustDisp(disp);
head(dispSH);
```

getNormFactor *Estimate size factors*

Description

Calculate the size factor.

Usage

```
getNormFactor(countsTable1)
```

Arguments

countsTable1	A data.frame or a matrix of counts in which a row represents for a gene and a column represents for a sample. There must be at least two columns in countsTable.

References

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biology, 11, R106.

Examples

```
#load a simulated data that includes a count table
data("countsTable");
getNormFactor(countsTable);
```

getQ *Estimate the shrinkage target based on the quantiles of initial targets*

Description

The shrinkage target is estimated.

Usage

```
getQ(countsTable, sizeFactors=NULL, q.vec=NULL, plotASD=FALSE,
    numPart=1, propForSigma=c(0, 1), verbose=TRUE, shrinkTarget=NULL,
    shrinkQuantile=NULL)
```

Arguments

countsTable	A data.frame or a matrix of counts in which a row represents for a gene and a column represents for a sample. There must be at least two columns in countsTable.
sizeFactors	A vector of values around 1 which are used to normalize between samples or libraries. The length of this vector equals to the number of columns in countsTable.
q.vec	A vector of sequence defines the quantiles. When q.vec=NULL, this function will generate a sequence for q.vec using seq(0.05, 0.995, 0.005).

plotASD — A logic value. If plotASD=TRUE, then the plot of ASD versus target points will be drawn. The SH estimates are obtained by shrinking the MM estimates toward a target point. Different SH estimates are generated using different target points. The target point that helps produce a small and stable averaged squared difference (ASD) between the MM estimates and the SH estimates is the point that approximates the common information across per- gene dispersion.

numPart — An integer indicates the number of groups for dispersion estimation. 'numPart=1' is the default value. It assumes that most of the genes share one point of stabilization (POS), and calculates the SH estimates without separating data into groups. When we assumes that genes can share different targets, the grouped SH estimates on dispersion can be be utilized. In this situation, users need to provide a number indicating the number of POS.

propForSigma — A range vector between 0 and 1 that is used to select a subset of data. It helps users to make a flexible choice on the subset of data when they believe only part of data should be used to estimate the variation among per-gene dispersion. A default input propForSigma=c(0, 1) is recommended. It means that we want to use all the data to estimate the variation.

verbose — A logic value. When verbose=TRUE, the detail information will be printed in the console.

shrinkTarget — A value for the shrinkage target of dispersion estimates. If "shrinkTarget=NULL" and "shrinkQuantile" is a value instead of NULL, then the quantile value for "shrinkQuantile" is converted into the scale of dispersion estimates and used as the target. If both of them are NULL, then a value that is small and minimizes the average squared difference is automatically used as the target value. If both of them are not NULL, then the value of "shrinkTarget" is used as the target.

shrinkQuantile — A quantile value for the shrinkage target of dispersion estimates. If "shrinkTarget=NULL" and "shrinkQuantile" is a value instead of NULL, then the quantile value for "shrinkQuantile" is converted into the scale of dispersion estimates and used as the target. If both of them are NULL, then a value that is small and minimizes the average squared difference is automatically used as the target value. If both of them are not NULL, then the value of "shrinkTarget" is used as the target.

Value

target — The estimated point for stabilization that represents the common in formation across per-gene dispersion.

q — A value that shows the quantile of the target value across per-gene dispersion.

Examples

```
#load a simulated data that includes a count table
data("countsTable")
conds <- c("A",  "B")
getQ(countsTable, plotASD=TRUE)
```

getT *Estimate the shrinkage target based on the initial estimates*

Description

This function is recommended to estimate the shrinkage target.

Usage

```
getT(countsTable, sizeFactors = NULL, q.vec = NULL, plotASD = FALSE,
    numPart = 1, propForSigma = c(0, 1), verbose = TRUE,
    shrinkTarget = NULL, shrinkQuantile = NULL, shrinkVar = FALSE,
    eSlope = 0.05, disp = NULL, dispXX = NULL, normalize = FALSE,
    lwd1 = 4.5, cexlab1 = 1.2)
```

Arguments

countsTable	A data.frame or a matrix of counts in which a row represents for a gene and a column represents for a sample. There must be at least two columns in countsTable.
sizeFactors	A vector of values around 1 which are used to normalize between samples or libraries. The length of this vector equals to the number of columns in countsTable.
q.vec	A vector of sequence defines the quantiles. When q.vec=NULL, this function will generate a sequence for q.vec using seq(0.05, 0.995, 0.005).
plotASD	A logic value. If plotASD=TRUE, then the plot of ASD versus target points will be drawn. The SH estimates are obtained by shrinking the MM estimates toward a target point. Different SH estimates are generated using different target points. The target point that helps produce a small and stable averaged squared difference (ASD) between the MM estimates and the SH estimates is the point that approximates the common information across per- gene dispersion. This target point is termed as the point of stabilization.
numPart	An integer indicates the number of groups for dispersion estimation. 'numPart=1' is the default value. It assumes that most of the genes share one point of stabilization (POS), and calculates the SH estimates without separating data into groups. When we assumes that genes can share different points of stabilization, the grouped SH estimates on dispersion can be be utilized. In this situation, users need to provide a number indicating the number of POS.
propForSigma	A range vector between 0 and 1 that is used to select a subset of data. It helps users to make a flexible choice on the subset of data when they believe only part of data should be used to estimate the variation among per-gene dispersion. A default input propForSigma=c(0, 1) is recommended. It means that we want to use all the data to estimate the variation.
verbose	A logic value. When verbose=TRUE, the detail information will be printed in the console.
shrinkTarget	A value for the shrinkage target of dispersion estimates. If "shrinkTarget=NULL" and "shrinkQuantile" is a value instead of NULL, then the quantile value for "shrinkQuantile" is converted into the scale of dispersion estimates and used as the target. If both of them are NULL, then a value that is small and minimizes the average squared difference is automatically used as the target value. If both of them are not NULL, then the value of "shrinkTarget" is used as the target.

shrinkQuantile	A quantile value for the shrinkage target of dispersion estimates. If "shrinkTarget=NULL" and "shrinkQuantile" is a value instead of NULL, then the quantile value for "shrinkQuantile" is converted into the scale of dispersion estimates and used as the target. If both of them are NULL, then a value that is small and minimizes the average squared difference is automatically used as the target value. If both of them are not NULL, then the value of "shrinkTarget" is used as the target.
shrinkVar	A logic value. When "shrinkVariance=TRUE", the testing is based on the shrinkage estimates for variance instead of dispersion.
eSlope	A positive value near to zero. When selecting the shrinkage target that is small and minimizing the average squared difference (ASD), the value of "elope" is a threshold to stop the selection steps if the absolute value of a local slope for the ASD is less than the threshold. The default value is 0.05.
disp	A vector of initial estimates of dispersions. The length of this vector equals to the number of rows in countsTable.
dispXX	A vector of normalized mean expression. The length of this vector equals to the number of rows in countsTable.
normalize	A logic value. When estimating the shrinkage target based on the average squared difference (ASD) between the shrinkage estimates and the initial estimates, the initial estimates and ASD are normalized when "normalize=TRUE".
lwd1	A value specifying the width of the curve shown in the plot for the average squared difference when "plotASD=TRUE". The default value is 4.5.
cexlab1	A value specifying the size of label text shown in the plot for the average squared difference when "plotASD=TRUE". The default value is 1.2.

Value

target	The estimated point for stabilization that represents the common in formation across per-gene dispersion.
q	A value that shows the quantile of the target value across per- gene dispersion.

Examples

```
#load a simulated data that includes a count table
data("countsTable")
conds <- c("A",  "B")
getT(countsTable, plotASD=TRUE)
```

---

getTgroup *This is an internal function used to calculate the shrinkage estimation when multiple shrinkage targets are considered.*

Description

Internal function where there are multiple shrinkage targets.

Usage

```
getTgroup
```

See Also

nbTestSH.

Examples

```
data("countsTable")
conds <- c("A",  "B")
resSH <- nbTestSH( countsTable, conds, "A", "B", numPart=10)
```

Hammer2months *An example of real experiment.*

Description

A subset of the real experiment Hammer et al. It is used as an example for running some functions in this package.

Usage

```
data(Hammer2months)
```

Format

A data.frame containing 4 columns and 29516 rows.

Details

It compares gene expression in rat strains Sprague Dawley and L5 SNL Sprague Dawley 2 at the end of two months in a factorial design. Two distinct biological libraries per condition were quantified using the Illumina platform.

Source

http://bowtie-bio.sourceforge.net/recount/

References

Hammer, P. et al. (2010). mRNA-seq with agnostic splice site discovery for nervous system transcriptomics tested in chronic pain. Genome Res., 20, 847-860.

Frazee, A. et al. (2011). ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets. BMC Bioinformatics, 12, 449.

Examples

```
data(Hammer2months);
head(countsTable);
```

nbinomTestForMatricesSH *Exact test under Negative Binomial Test with Shrinkage Estimates on Dispersions*

Description

This is an internal function used by nbTestSH. It calculates the exact per-gene probabilities for p-values, and tests the null hypothesis that the expected expression of a gene under two conditions are not different.

Usage

```
nbinomTestForMatricesSH(countsA, countsB, sizeFactorsA, sizeFactorsB,
    numPart=1, SHonly=FALSE, propForSigma=c(0, 1), shrinkTarget=NULL,
    shrinkQuantile=NULL, cLA, cLB, contrast=NULL,
    keepLevelsConsistant=TRUE,
    useMMdisp=FALSE, shrinkVariance=FALSE, pairedDesign=FALSE,
    pairedDesign.dispMethod="per-pair", useFisher=FALSE, Dispersions=NULL,
    eSlope=NULL, plotASD=FALSE, lwd_ASD=4.5, cex_ASD=1.2)
```

Arguments

countsA	A counts table under condition “condA".
countsB	A counts table under condition “condB".
sizeFactorsA	A vector of size factors under condition “condA".
sizeFactorsB	A vector of size factors under condition “condB".
numPart	An integer indicating the number of targets for the shrinkage dispersion estimates. “numPart=1" is the default value. It assumes that all the genes share one common target, and then the method of moment estimates are shrunk toward one single target. When it is assumed that the genes share multiple targets, the value for “numPart" is the number of targets and the grouped shrinkage estimates for dispersion are calculated.
SHonly	If ‘SHonly’ is TRUE, then the function outputs the shrinkage estimates for dispersion without testing the differentiation between conditions. If FALSE, then the function outputs a data frame including the per-gene p-values of tests.
propForSigma	A range vector between 0 and 1 that is used to select a subset of data. It helps users to make a flexible choice on the subset of data when they believe only part of data should be used to estimate the variation among per-gene dispersion. An input “propForSigma=c(0.1, 0.9)" means that the genes having method of moment estimates for dispersion greater than the 10th quantile and less than the 90th quantile are used to estimate the dispersion variation. The default input “propForSigma=c(0, 1)" is recommended. It means that we want to use all the data to estimate the dispersion variation.
shrinkTarget	A value for the shrinkage target of dispersion estimates. If “shrinkTarget=NULL" and “shrinkQuantile" is a value instead of NULL, then the quantile value for “shrinkQuantile" is converted into the scale of dispersion estimates and used as the target. If both of them are NULL, then a value that is small and minimizes the average squared difference is automatically used as the target value. If both of them are not NULL, then the value of “shrinkTarget" is used as the target.

`shrinkQuantile` A quantile value for the shrinkage target of dispersion estimates. If "shrinkTarget=NULL" and "shrinkQuantile" is a value instead of NULL, then the quantile value for "shrinkQuantile" is converted into the scale of dispersion estimates and used as the target. If both of them are NULL, then a value that is small and minimizes the average squared difference is automatically used as the target value. If both of them are not NULL, then the value of "shrinkTarget" is used as the target.

`cLA` A data.frame indicating the levels or extra factors under condition "condA".

`cLB` A data.frame indicating the levels or extra factors under condition "condB".

`contrast` A contrast vector for testing in complex experiments. The length of this vector equals to the number of columns in countsTable.

`keepLevelsConsistant`
A logic TRUE/FALSE value. When "coLevels" is used to indicate a paired design experiment, "keepLevelsConsistant=TRUE" silences the genes that have different changing directions (i.e. positive and negative test statistics) among individual samples by setting their p-values as 1.

`useMMdisp` A logic value. When "useMMdisp=TRUE" the method of moment (MM) estimates for dispersion without any shrinkage approach are used for testing the differentiation of genes between two conditions.

`shrinkVariance` A logic value. When "shrinkVariance=TRUE", the testing is based on the shrinkage estimates for variance instead of dispersion.

`pairedDesign` A logic value. When pairedDesign=TRUE is specified, the tests are performed specifically for the paired design experiment. The Null hypotheses $\sum_l(\mu_{gA,l} - \mu_{gB,l}) = 0$ will be tested.

`pairedDesign.dispMethod`
A character specifying the method of selecting data used for the paired design experiment. When the input is "per-pair" (the default input), the dispersion estimates are shrunk within each pair of samples. The shrinkage target is different in different pair of samples. When the input is "pooled", firstly method of moment estimates for dispersion are obtained within each pair of samples, and then the average estimates across all pairs of samples are shrunk toward a common targets among genes.

`useFisher` A logic value specifying whether Fisher's method of combining multiple p-values for a gene is used in the paired design experiment. In detail the formula of calculating the Fisher's combined p-values is $\text{pval}_g = \chi^2_{df=2k}(X > x)$ where k is the number of pairs and $x = -2 * \sum_{l=1}^{k} log_e(p_l)$. The default input is FALSE and the formulae $\text{pval}_g = exp(\sum_{l=1}^{k} log_e(p_l))$ is used.

`Dispersions` If it is not null, then the input is a vector of known dispersion values. The length of the vector equals to the number of genes in the counts table. The default value is "NULL".

`eSlope` A positive value near to zero. When selecting the shrinkage target that is small and minimizing the average squared difference (ASD), the value of "elope" is a threshold to stop the selection steps if the absolute value of a local slope for the ASD is less than the threshold. The default value is 0.05.

`plotASD` A logic value. If plotASD=TRUE, then the plot of average squared difference (ASD) versus target points is produced. The shrinkage (SH) estimates are obtained by shrinking the method of moment (MM) estimates toward a target. In the figure, the vertical axis are ASD values when the shrinkage target (represented by the horizontal axis) varies within the range of dispersion estimates. The selected target is a small value minimizing ASD.

lwd_ASD	A value specifying the width of the curve shown in the plot for the average squared difference when “plotASD=TRUE". The default value is 4.5.
cex_ASD	A value specifying the size of label text shown in the plot for the average squared difference when “plotASD=TRUE". The default value is 1.2.

See Also

nbTestSH.

---

nbTestSH *Differential Analysis based on RNA-seq experiments using Negative Binomial (NB) Model with Shrinkage Approach of Dispersion Estimation.*

Description

This is the main function calculating the exact per-gene probabilities for p-values. It tests the null hypothesis that the expected expression of a gene under two conditions are the same.

Usage

```
nbTestSH(countsTable, conds, condA = "A", condB = "B",
    numPart = 1, SHonly = FALSE, propForSigma = c(0, 1),
    shrinkTarget = NULL, shrinkQuantile = NULL, plotASD = FALSE,
    coLevels = NULL, contrast = NULL, keepLevelsConsistant = FALSE,
    useMMdisp = FALSE, addRawData = FALSE, shrinkVariance = FALSE,
    pairedDesign = FALSE, pairedDesign.dispMethod = "per-pair",
    useFisher = FALSE, Dispersions = NULL, eSlope = 0.05, lwd_ASD = 4.5,
    cex_ASD = 1.2)
```

Arguments

countsTable	A data.frame or a matrix of counts in which a row represents for a gene and a column represents for a sample. There must be at least two columns in countsTable.
conds	A vector of characters representing the two conditions (or two groups). It must be matchable to the columns in countsTable. For example, c("A", "A", "B", "B") matches to a countsTable that has four columns (or samples) in which the first two columns are samples under condition A and the last two columns are samples under condition B.
condA	A character specifying the first condition in countsTable, e.g. condA="A".
condB	A character specifying the second condition in countsTable, e.g. condB="B".
numPart	An integer indicating the number of targets for the shrinkage dispersion estimates. “numPart=1" is the default value. It assumes that all the genes share one common target, and then the method of moment estimates are shrunk toward one single target. When it is assumed that the genes share multiple targets, the value for “numPart" is the number of targets and the grouped shrinkage estimates for dispersion are calculated.
SHonly	If ‘SHonly’ is TRUE, then the function outputs the shrinkage estimates for dispersion without testing the differentiation between conditions. If FALSE, then the function outputs a data frame including the per-gene p- values of tests.

`propForSigma` A range vector between 0 and 1 that is used to select a subset of data. It helps users to make a flexible choice on the subset of data when they believe only part of data should be used to estimate the variation among per-gene dispersion. An input "propForSigma=c(0.1, 0.9)" means that the genes having method of moment estimates for dispersion greater than the 10th quantile and less than the 90th quantile are used to estimate the dispersion variation. The default input "propForSigma=c(0, 1)" is recommended. It means that we want to use all the data to estimate the dispersion variation.

`shrinkTarget` A value for the shrinkage target of dispersion estimates. If "shrinkTarget=NULL" and "shrinkQuantile" is a value instead of NULL, then the quantile value for "shrinkQuantile" is converted into the scale of dispersion estimates and used as the target. If both of them are NULL, then a value that is small and minimizes the average squared difference is automatically used as the target value. If both of them are not NULL, then the value of "shrinkTarget" is used as the target.

`shrinkQuantile` A quantile value for the shrinkage target of dispersion estimates. If "shrinkTarget=NULL" and "shrinkQuantile" is a value instead of NULL, then the quantile value for "shrinkQuantile" is converted into the scale of dispersion estimates and used as the target. If both of them are NULL, then a value that is small and minimizes the average squared difference is automatically used as the target value. If both of them are not NULL, then the value of "shrinkTarget" is used as the target.

`plotASD` A logic value. If plotASD=TRUE, then the plot of average squared difference (ASD) versus target points is produced. The shrinkage (SH) estimates are obtained by shrinking the method of moment (MM) estimates toward a point target. In the figure, the vertical axis are ASD values when the shrinkage target (represented by the horizontal axis) varies within the range of dispersion estimates. The selected target is a small value minimizing ASD.

`coLevels` A data.frame specifying the additional factors for testing in complex experiments. The number of row in "coLevels" matches the number of columns in countsTable. It describes the extra features or factors other than the two basic conditions. For example, "conds=c("A","A","B","B")" and "coLevels=data.frame(sample=c(1,2,1,2)) indicate a paired design experiment. Column 1 and 3 in countsTable are a paired observations for sample 1 in two different conditions.

`contrast` A contrast vector for testing in complex experiments. The length of this vector equals to the number of columns in countsTable.

`keepLevelsConsistant`
A logic TRUE/FALSE value. When "coLevels" is used to indicate a paired design experiment, "keepLevelsConsistant=TRUE" silences the genes that have different changing directions (i.e. positive and negative test statistics) among individual samples by setting their p-values as 1.

`useMMdisp` A logic value. When "useMMdisp=TRUE" the method of moment (MM) estimates for dispersion without any shrinkage approach are used for testing the differentiation of genes between two conditions.

`addRawData` A logic value. When "addRawData=TRUE", this function also outputs the original values of countsTable.

`shrinkVariance` A logic value. When "shrinkVariance=TRUE", the testing is based on the shrinkage estimates for variance instead of dispersion.

`pairedDesign` A logic value. When pairedDesign=TRUE is specified, the tests are performed specifically for the paired design experiment. The Null hypotheses $\sum_l(\mu_{gA,l} - \mu_{gB,l}) = 0$ will be tested.

`pairedDesign.dispMethod`
: A character specifying the method of selecting data used for the paired design experiment. When the input is “per-pair" (the default input), the dispersion estimates are shrunk within each pair of samples. The shrinkage target is different in different pair of samples. When the input is "pooled", firstly method of moment estimates for dispersion are obtained within each pair of samples, and then the average estimates across all pairs of samples are shrunk toward a common targets among genes.

`useFisher`
: A logic value specifying whether Fisher's method of combining multiple p- values for a gene is used in the paired design experiment. In detail the formula of calculating the Fisher's combined p-values is $\text{pval}_g = \chi^2_{df=2k}(X > x)$ where k is the number of pairs and $x = -2 * \sum_{l=1}^{k} log_e(p_l)$. The default input is FALSE and the formulae $\text{pval}_g = exp(\sum_{l=1}^{k} log_e(p_l))$ is used.

`Dispersions`
: If it is not null, then the input is a vector of known dispersion values. The length of the vector equals to the number of genes in the counts table. The default value is “NULL".

`eSlope`
: A positive value near to zero. When selecting the shrinkage target that is small and minimizing the average squared difference (ASD), the value of “elope" is a threshold to stop the selection steps if the absolute value of a local slope for the ASD is less than the threshold. The default value is 0.05.

`lwd_ASD`
: A value specifying the width of the curve shown in the plot for the average squared difference when “plotASD=TRUE". The default value is 4.5.

`cex_ASD`
: A value specifying the size of label text shown in the plot for the average squared difference when “plotASD=TRUE". The default value is 1.2.

Value

`Mean`
: The row per-gene averages over the values in countsTable.

`log2FoldChange`
: The per-gene fold Changes between condition A and B in the log2 scale.

`dispMM`
: The per-gene method of moment (MM) estimates on dispersion.

`dispSH`
: The per-gene shrinkage (SH) estimates on dispersion.

`pval`
: The per-gene p-values based on the exact tests. Smaller p-value indicates a higher chance of rejecting the null hypothesis that the expected gene expression distributes identically between the two conditions.

References

Yu, D., Huber, W. and Vitek O. (2013). Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size. Bioinformatics.

Examples

```
#load a simulated data that includes a count table
data("countsTable")

#Differential analysis in sSeq.
conds <- c("A",  "B")
resSH <- nbTestSH( countsTable, conds, "A", "B")

#If users only want to calculate the SH dispersion estimates and
```

```
#draw a mean-dispersion plot, the following scripts can be used.
library('RColorBrewer')
dispSH <- nbTestSH( countsTable, conds, "A", "B", SHonly=TRUE)
plotDispersion(dispSH)
```

`plotDispersion` *Drawing Dispersion-Mean plot.*

Description

This function is used to draw a scatter plot of dispersion versus mean of count table. It helps to visually inspect the dependence between the dispersion estimates and the mean estimates.

Usage

```
plotDispersion(DispSH, extraOutput=NULL, plotMethod="logDisp",
    ylim1=NULL, legPos="topleft", myCol=brewer.pal(9, "Set1"),
    tt=NULL)
```

Arguments

`DispSH`	A data frame includes 'SH', 'raw', and 'mus'. They are the shrinkage estimates of dispersion, the method of moment estimates of dispersion, and the estimates of mean. This data frame is obtained using the function 'nbTestSH' and specifying 'SHonly=TRUE'.
`extraOutput`	A data.frame including dispersion estimates and expectation estimates using another method. When users want to compare the dispersion estimates using two different method, this argument can be used to include the result from the second method. The default value is NULL. This means that no extra method is compared.
`plotMethod`	If plotMethod="logDisp" which is the default, then both dispersion and mean estimates are shown in the log scale. If plotMethod="Disp", then only mean estimates are shown in the log scale.
`ylim1`	A vector of two values that specifies the minimum and maximum values of the vertical y axis in the plot. It is used to limit the presenting range of y axis in the plot. If ylime1=NULL then the range of the shrinkage estimates of dispersion is used.
`legPos`	A character indicating the position of legend in the plot. The value of this argument can be "topleft", "topright", "bottomleft" and "bottomright".
`myCol`	A vector of colors corresponding to the dispersions estimated using different methods.
`tt`	A character representing the title of the plot, which is shown on the top in the plot.

Examples

```
data("countsTable")
conds <- c("A", "B")
dispSH <- nbTestSH( countsTable, conds, "A", "B", SHonly=TRUE)

library('RColorBrewer')
plotDispersion(dispSH, legPos="topleft")
```

rnbinomMV *Randomly Generate Negative Binomial Variable with parameters mean and variance.*

Description

This function is based on the re-parameterized Negative Binomial distribution to generate random observations.

Usage

```
rnbinomMV(n, mu, v)
```

Arguments

n	The number of values that will be randomly generated.
mu	The expectation of the Negative Binomial distribution.
v	The variance of the Negative Binomial distribution.

Examples

```
x <- rnbinomMV(50, 10, 15)
hist(x)
```

rowVars *Calculating the sample variance within each row of A matrix*

Description

This function helps to obtain row-wise estimation across columns.

Usage

```
rowVars(x)
```

Arguments

x	A matrix or data.frame that includes multiple columns.

Value

A vector showing the per-row variance estimates for the matrix or data.frame.

Examples

```
x <- matrix(rnorm(10), 5)
rowVars(x)
```

sim *Generating Simulated Data*

Description

This function is used to approximate the real experiment and to generate simulated counts table based on Negative Binomial distribution.

Usage

```
sim(ngenes, true_mean1, conds,
  alpha = function(m) {rep(0.1, length(m))},
  mean_DE = 0, sd_DE = 2, s0 = NULL, s0_mean = 2, s0_sd = 3,
  true_isDE_proportion = 0.3)
```

Arguments

ngenes	The total number of genes or rows in the simulated counts table.
true_mean1	The expected gene expression (μ_g) in a library or a sample. The length of this vector equals 'ngenes'. It can generated from either random distributions or averages of counts table from a real experiment.
conds	A vector of characters representing the two conditions (or two groups). It must be matchable to the columns in countsTable, e.g., c("A", "A", "B", "B") matches to a countsTable that has four columns (or samples) in which the first two columns are samples under condition A and the last two columns are samples under condition B.
alpha	A function used to generate the true dispersion values. The default function generates a constant 0.1 for all the genes. It can also be a function specifying the dependence between dispersion and mean.
mean_DE	A true mean value of ϵ in $\mu_{gB} = \mu_g / exp(\epsilon)$ where ϵ follows a Normal distribution.
sd_DE	A true standard deviation of ϵ in $\mu_{gB} = \epsilon \mu_g$ where ϵ follows a Normal distribution.
s0	The true size factors for samples. The length of this vector equals to the length of the vector 'conds'.
s0_mean	If the true size factors for samples are not defined for 's0', then the true size factors are assumed to follow a Normal distribution with mean as the value for 's0_mean'.
s0_sd	If the true size factors for samples are not defined for 's0', then the true size factors are assumed to follow a Normal distribution with standard deviation as the value for 's0_sd'.
true_isDE_proportion	The proportion of genes that are truly different. The default value is 0.3.

Value

The function outputs a list including the simulated counts table, a vector with TRUE of FALSE values indicating the truly differentiating genes, the true mean values, the true variance values, and the true dispersion values.

Note

We acknowledge Dr. Simon Anders since he provided the details for simulation in the manual of DESeq package.

References

Yu, D., Huber, W. and Vitek O. (2013). Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size. Bioinformatics.

Examples

```
ng = 10000;
sim1 <- sim(ngenes=ng, conds=c("A","A","B","B"),
   true_mean1=round(rexp(ng, rate=1/200)), alpha=function(m){1/(m+100)},
   mean_DE=2, sd_DE=1, s0=runif(4, 0, 2) );
true_isDE <- sim1$true_isDE;
countsTable <- sim1$countsTable;
```

Sultan *An example of real experiment.*

Description

A subset of the real experiment Sultan et al. It is used as an example for running some functions in this package.

Usage

```
data(Sultan)
```

Format

A data.frame containing 4 columns and 52580 rows.

Details

It compares two biological replicates of human cell lines Ramos B and HEK293T with the Illumina platform.

Source

http://bowtie-bio.sourceforge.net/recount/

References

Sultan, M. et al. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science, 321, 956.

Frazee, A. et al. (2011). ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets. BMC Bioinformatics, 12, 449.

Examples

```
data(Sultan);
head(countsTable);
```

Tuch *An example of real experiment.*

Description

A subset of the real experiment Tuch et al. It is used as an example for running some functions in this package.

Usage

```
data(Tuch)
```

Format

A data.frame containing 6 columns and 10453 rows.

Details

It compares the expression of genes in normal human tissues and in tissues with oral squamous cell carcinoma. The experiment had a paired design in that pairs of normal and tumor samples were obtained from three patient. The six libraries were sequenced using the SOLiD platform.

Source

The table of read counts was downloaded from GEO (accession GSE20116).

References

Tuch, B. et al. (2010). Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. PloS One, 5, e9317.

Examples

```
data(Tuch);
head(countsTable);
```

Index