Package 'deGPS'

February 26, 2014

Type Package

Title Differential Expression Tests Based on Generalized Poisson Statistic

Version 1.0

Date 2014-02-14

Author Chen Chu

Maintainer Chen Chu <chuchen.blueblues@gmail.com>

Depends R (>= 2.15.3), foreach, doMC

Suggests LPE, limma, edgeR

Description Use methods based on Generalized Poisson Distribution to do RNA-seq differential expression tests.

License GPL-2

Encoding latin1

R topics documented:

deGPS-package *Differential Expression Test for RNA-seq Data*

Description

To do normalization using multiple methods, including those based on Generalized Poisson Distribution, at the same time and calculate the corresponding pvalues according to the empirical distribution of T-statistics of RNAs.

Details

Package:	deGPS
Type:	Package
Version:	1.0
Date:	2014-02-24
License:	GPL-2

To do DE test for miRNA, call GPSmle. To do DE test for mRNA, call deGPS_mRNA.

TIPS

1. If SampleSize > 10, Specify maxIter. The larger it is, more time the function will take.

2. paired = TRUE is not proven yet. Better to use unpaired test.

3. If paired = TRUE, careful with the sample order in yourData and group specify. For example, there are two pairs: X1 & Y1, X2 & Y2, you'd better:

group = c(1, 1, 2, 2)

yourData <- yourData[, c("X1", "X2", "Y1", "Y2"]

The program takes the first "1" and the first "2" in "group" specification as the first pair and so on.

4. pvalue = "bind" is the most powerful choice in MLE2.

5. MLE2 refers to GP-MLE2L in the our article, while GP refers to GP-Quantile and GP2 refers to GP-Theta.

6. nSubcore is used to split your RNA. Suppose there are 2000 RNAs in your data. And nSubcore = 2. In each core, 1000 RNAs' T-stats are calculated and therefore the pvalues after combining the T-stats. When multiple methods are specified or method = "MLE2", ncore > nSubcore is recommended for the parallel computing.

7. It is suggested to follow the exact specification of the examples below (except data/group)

Examples

deGPS_mRNA(dataSim = yourData, group = rep(1:2, dim(yourData)[2] / 2), method = "MLE2", pvalue = "bind", paired = FALSE, nSubcore = 8, ncore = 16)

GPSmle(data = yourData, group = rep(1:2, dim(yourData)[2] / 2), type = "pvalue", method = "MLE2", pvalueType = "bind", maxIter = 500, paired = FALSE, ncpu = 4, resampling = "random")

Author(s)

Chen Chu

Maintainer: Chen Chu <chuchenblueblues@gmail.com>

deGPS_mRNA *Differential Expression Test for mRNA-seq Data*

Description

Differential Expression Test for mRNA-seq Data

Usage

```
deGPS_mRNA(dataSim = dataSim, dataNormal = dataNormal,
group = rep(1:2, each = 5), method = "MLE2", nSubcore = 3, ncore = 15,
pvalue = "bind", paired = FALSE, maxIter = 150, MLEIter = 1)
```

Arguments

dataSim	mRNA-seq Data. The column represents samples while the row represents mRNAs.
dataNormal	not used anymore.
group	specify which group the sample belongs to. Its values are 1 or 2. Must have the equal length to the dataSim column.
method	the methods of normalization. It can be a single charactor or a vector, the values can be "global", "Lowess", "GP", "Quantile", "TMM", "GP2" or "MLE2".
nSubcore	the cpu used for splitted computing of mRNAs.
ncore	the total cpus used for the calculations. It is better to make round(MLEIter*nrow(dataSim)) * nSubcore/ncore and sum(method != MLE2) * nSubcore/ncore integer.
pvalue	the value is "bind", "overlap" or "ave", which represents the minimum, maximum and average of pvalue for MLE2 normalized data.
paired	paired test or not. The default is FALSE.
maxIter	the default value of maxIter is 150. When sample size is large. Instead of transverse every possible resample, randomly resampling with a maximum iteration times is used to get empirical distribution. The larger maxIter is, the more time the algorithm takes. Be careful with the choice.
MLEIter	the algorithm defaultly takes every sample as reference to do normalization under method MLE2. If sample size is large, there may be no need to do that. Instead, take part of the samples as reference is enough. The default is 1. round(MLEIter * nrow(dataSim)) samples are used as reference.

Details

To do nomalization using multiple methods at one time and calculate the corresponding empirical distribution of T-statistic and pvalues.

Value

StatAllArray	A list of the empirical T-stats of mRNAs of different normalization methods.
pvalue	The resulted pvalues

Author(s)

Chen Chu

See Also

GPSmle

GPSmle *Generalized Poisson Statistical Maximum Likelihood Estimation.*

Description

Use MLE of normalization factor to normalize microRNA read count data, assuming every replicate in data obey a GP distribution. Then use empirical distribution of the T-stats to evaluate the differential expressions of microRNAs.

Usage

```
GPSmle(data = dataSim, group = rep(1:2, each = 5),
type = c("normalization", "ecdf", "pvalue"),
method = c("global", "Lowess", "GP", "Quantile", "TMM", "GP2", "MLE2"),
pvalueType = c("bind", "overlap", "ave"), maxIter = 500, paired = FALSE,
set.seed = NULL, ncpu = 1, resampling = c("random", "uniformed"), MLEIter = 1)
```

Arguments

data	a data frame of microRNA read counts.
group	a numeric vector specified the group. Its values are 1 or 2. Must have the equal length to the dataSim column.
type	is one value in "normalization", "ecdf" or "pvalue". The function is processing from "normalization" to "ecdf", and at last returns the "pvalue". The type means at which step the user would like the function to stop.
method	the methods of normalization. It can be a single charactor or a vector, the values can be "global", "Lowess", "GP", "Quantile", "TMM", "GP2" or "MLE2".
pvalueType	the value is "bind", "overlap" or "ave", which represents the minimum, maximum and average of pvalue for MLE2 normalized data.
maxIter	the default value of maxIter is 500. when sample size is large. Instead of transverse every possible resample, randomly resampling with a maximum iteration times is used to get empirical distribution. The larger maxIter is, the more time the algorithm takes. Be careful with the choice.
paired	paired test or not. The default is FALSE.
set.seed	for larger sample size, it is better to set a seed for the randomly resampling. If set.seed is NULL, a random seed generated from the current time of the system is used.
ncpu	number of cores for the parallel computing
resampling	the value is "uniformed" or "random". "uniformed" means that the resampling ensures the almost equal size of the original two groups in the groups after resampling, while "random" means the resampling is conducted randomly. It is better to use "random", because "uniformed" may result in unexpected bias sometimes, especcially when the two groups are not with equal size.

MLEIter	the algorithm defaultly takes every sample as reference to do normalization under method MLE2. If sample size is large, there may be no need to do that. Instead, take part of the samples as reference is enough. The default is 1. round(MLEIter * nrow(dataSim)) samples are used as reference.

Details

Differential Expression Tests for miRNA-seq data. There are multiple choices of the normalization methods. More than one method can be specified at one time. For "MLE2", the pvalueType must be specified. Because the unexpected bias, it is suggested to specify resampling as "random", whatever the sample size is. When sample size is large, maxIter must be specified and MLEIter is suggested to specify less than 1. And if ncpu larger than 1, parallel computing is used in the calculations. More details about MLE2(GP-MLE2L), GP(GP-Quantile), GP2(GP-Theta) can be found in our thesis.

Value

normalization	the data after normalization.
ecdf	empirical cumulative distribution function of all normalization methods.
pvalue	p value derived from ecdf.
...	the rest values are the settings of this funciton

Author(s)

Chen Chu

See Also

GPSmle.default

GPSmle.default *Generalized Poisson Statistical Maximum Likelihood Estimation (default)*

Description

the default method for the function GPSmle.

Usage

```
## Default S3 method:
GPSmle(data = dataSim, group = rep(1:2, each = 5),
type = c("normalization", "ecdf", "pvalue"),
method = c("global", "Lowess", "GP", "Quantile", "TMM", "GP2", "MLE2"),
pvalueType = c("bind", "overlap", "ave"), maxIter = 500,
paired = FALSE, set.seed = NULL, ncpu = 1, resampling = c("random", "uniformed"),
MLEIter = 1)
```

Arguments

data	a data frame of microRNA read counts.
group	a numeric vector specified the group. Its values are 1 or 2. Must have the equal length to the dataSim column.
type	is one value in "normalization", "ecdf" or "pvalue". The function is processing from "normalization" to "ecdf", and at last returns the "pvalue". The type means at which step the user would like the function to stop.
method	the methods of normalization. It can be a single charactor or a vector, the values can be "global", "Lowess", "GP", "Quantile", "TMM", "GP2" or "MLE2".
pvalueType	the value is "bind", "overlap" or "ave", which represents the minimum, maximum and average of pvalue for MLE2 normalized data.
maxIter	the default value of maxIter is 500. when sample size is large. Instead of transverse every possible resample, randomly resampling with a maximum iteration times is used to get empirical distribution. The larger maxIter is, the more time the algorithm takes. Be careful with the choice.
paired	paired test or not. The default is FALSE.
set.seed	for larger sample size, it is better to set a seed for the randomly resampling. If set.seed is NULL, a random seed generated from the current time of the system is used.
ncpu	number of cores for the parallel computing
resampling	the value is "uniformed" or "random". "uniformed" means that the resampling ensures the almost equal size of the original two groups in the groups after resampling, while "random" means the resampling is conducted randomly. It is better to use "random", because "uniformed" may result in unexpected bias sometimes, especcially when the two groups are not with equal size.
MLEIter	the algorithm defaultly takes every sample as reference to do normalization under method MLE2. If sample size is large, there may be no need to do that. Instead, take part of the samples as reference is enough. The default is 1. round(MLEIter * nrow(dataSim)) samples are used as reference.

Details

Differential Expression Tests for miRNA-seq data. There are multiple choices of the normalization methods. More than one method can be specified at one time. For "MLE2", the pvalueType must be specified. Because the unexpected bias, it is suggested to specify resampling as "random", whatever the sample size is. When sample size is large, maxIter must be specified and MLEIter is suggested to specify less than 1. And if ncpu larger than 1, parallel computing is used in the calculations. More details about MLE2(GP-MLE2L), GP(GP-Quantile), GP2(GP-Theta) can be found in our thesis.

Value

normalization	the data after normalization.
ecdf	empirical cumulative distribution function of all normalization methods.
pvalue	p value derived from ecdf.
...	the rest values are the settings of this funciton

See Also

GPSmle

GPSmleEst *GPSmle Estimation Function*

Description

see also GPSmle

Usage

```
GPSmleEst(data = dataSim, group = rep(1:2, each = 5),
type = c("normalization", "ecdf", "pvalue"), dataNormal = NULL,
ecdf = NULL, method = c("global", "Lowess", "GP", "Quantile", "TMM", "GP2", "MLE2"),
pvalueType = c("bind", "overlap", "ave"), maxIter = 500, paired = FALSE,
set.seed = NULL, ncpu = 1, resampling = c("random", "uniformed"), MLEIter = 1)
```

Arguments

data	a data frame of microRNA read counts.
group	a numeric vector specified the group. Its values are 1 or 2. Must have the equal length to the dataSim column.
type	is one value in "normalization", "ecdf" or "pvalue". The function is processing from "normalization" to "ecdf", and at last returns the "pvalue". The type means at which step the user would like the function to stop.
dataNormal	The default is NULL. Specify the data after normalization here, if there is any.
ecdf	The default is NULL. Specify the empirical values of T-stats here, if there is any.
method	the methods of normalization. It can be a single charactor or a vector, the values can be "global", "Lowess", "GP", "Quantile", "TMM", "GP2" or "MLE2".
pvalueType	the value is "bind", "overlap" or "ave", which represents the minimum, maximum and average of pvalue for MLE2 normalized data.
maxIter	the default value of maxIter is 500. when sample size is large. Instead of transverse every possible resample, randomly resampling with a maximum iteration times is used to get empirical distribution. The larger maxIter is, the more time the algorithm takes. Be careful with the choice.
paired	paired test or not. The default is FALSE.
set.seed	for larger sample size, it is better to set a seed for the randomly resampling. If set.seed is NULL, a random seed generated from the current time of the system is used.
ncpu	number of cores for the parallel computing
resampling	the value is "uniformed" or "random". "uniformed" means that the resampling ensures the almost equal size of the original two groups in the groups after resampling, while "random" means the resampling is conducted randomly. It is better to use "random", because "uniformed" may result in unexpected bias sometimes, especcially when the two groups are not with equal size.
MLEIter	the algorithm defaultly takes every sample as reference to do normalization under method MLE2. If sample size is large, there may be no need to do that. Instead, take part of the samples as reference is enough. The default is 1. round(MLEIter * nrow(dataSim)) samples are used as reference.

Details

see also GPSmle

Value

normalization	the data after normalization.
ecdf	empirical cumulative distribution function of all normalization methods.
pvalue	p value derived from ecdf.
...	the rest values are the settings of this funciton

Author(s)

Chen Chu

See Also

GPSmle.default

---

plot.GPSmle *Plot GPSmle*

Description

Plot the histogram of GPSmle results.

Usage

```
## S3 method for class GPSmle
plot(x, ...)
```

Arguments

x	the object returned by GPSmle.
...	see the plot.default

Value

the histogram depends on the parameters of GPSmle. If the type equals "normalization", the histograms of normalized data are returned. So are the "ecdf" and "pvalue".

See Also

GPSmle

summary.GPSmle *summary of GPSmle*

Description

summary of GPSmle

Usage

```
## S3 method for class GPSmle
summary(object, ...)
```

Arguments

object	the object returned by GPSmle.
...	see the summary.default

Details

summary of the GPSmle

Value

summary of the GPSmle

Index