Package 'PLNseq'

November 4, 2016

Type Package

Title PLNseq: A multivariate Poisson lognormal distribution for high-throughput correlated RNA-sequencing read count data

Version 1.0

Date 2014-06-20

Author Hong Zhang

Maintainer Hong Zhang <zhanghd@fudan.edu.cn>

Depends R (>= 2.10), MASS

Description PLNseq is an R package for identifying differentially expressed genes using RNA-sequencing read count data from correlated samples.

License GPL (>= 2)

LazyLoad yes

URL http://github.com/zhanghfd/PLNseq

BugReports http://github.com/zhanghfd/PLNseq/issues

RoxygenNote 5.0.1

R topics documented:

PLNseq-package *Differential expression analysis using matched read count data*

Description

This R package conducts differential expression (DE) analysis using high throughput next-generation sequencing read count data generated from correlated samples. The marginal distribution of the read count is the compounding of the Poisson distribution and the lognormal distribution ('PLN' distribution for short), and the correlation between the read counts of each matched sample set is modeled by the multivariate lognormal distribution with correlation coefficient matrix that is assumed to be common for all genes. This package provides estimates of rho (correlation coefficient matrix in multivariate lognormal distribution) and its standard error, common or genewise sigma (standard deviation of lognormal distribution), fold change (defined as the difference between log-gene expression of matched samples) and p-value for detecting differentially expressed genes.

Details

Package:	PLNseq
Type:	Package
Version:	1.0
Date:	2014-06-18
License:	GPL (>= 3)

Author(s)

Hong Zhang

Maintainer: Hong Zhang <zhanghd@fudan.edu.cn>

References

Zhang, H., Xu, J., Jiang N., Hu, X., and Luo, Z. (2015). PLNseq: A multivariate Poisson lognormal distribution for high-throughput matched RNA-sequencing read count data. Statistics in Medicine 34: 1577-1589.

commonSigma *Common sigma*

Description

Estimate 'mu' (mean parameter lognormal distribution) for each gene and condition and a common 'sigma' (standard deviation parameter of lognormal distribution).

Usage

```
commonSigma(d)
```

Arguments

d	This is a PLNseq object.

Value

d$commonSigma	A common 'sigma'

Examples

```
data(lung);
count = lung[,c(2:4,8:10)];
d = PLNobject(count,2);
d = sizeFactor(d,maxCount=2e3);
d = commonSigma(d);
```

---

correlationCoefficient

Correlation coefficient

Description

Estimate correlation coefficient parameter and its standard error in the multivariate lognormal distribution.

Usage

```
correlationCoefficient(d)
```

Arguments

d	This is a PLNseq object.

Value

d$rho	Correlation coefficient 'rho' in the multivariate lognormal distribution
d$rho.se	Standard error of estimated 'rho'

Examples

```
data(lung);
count = lung[,c(2:4,8:10)];
d = PLNobject(count,2);
d = sizeFactor(d,maxCount=2e3);
d = commonSigma(d);

## common correlation
## commonCorrelation = TRUE;
## d = correlationCoefficient(d);

## clustered correlation
```

```
## commonCorrelation = FALSE;
    ## J = nrow(count);
    ## J1 = round(J/2);
    ## d$cluster = c(rep(1,J1),rep(2,J-J1));
    ## d = correlationCoefficient(d);
```

---

genewiseSigma *Genewise sigma*

Description

Estimate genewise 'sigma' (standard deviation parameter of lognormal distribution).

Usage

```
genewiseSigma(d,w=25)
```

Arguments

d	This is a PLNseq object.
w	Shrinkage parameter.

Value

d$genewiseSigma
: Genewise 'sigma'

Examples

```
data(lung);
count = lung[,c(2:4,8:10)];
d = PLNobject(count,2);
d = sizeFactor(d,maxCount=2e3);

d = genewiseSigma(d);
```

LRtest1 *Likelihood ratio test for differential expression analysis with common correlation.*

Description

This function calculates log-fold changes, likelihood ratio test statistics, and p-values for a list of genes. This function should be called after a commom correlation matrix is returned by 'correlationCoefficient'.

Usage

```
LRtest1(d,z,use.commonSigma,id)
```

Arguments

d	This is a PLNseq object.
z	J independent samples (a matrix of dimension J by R) drown from multivariate normal distribution with expectations 0, variances 1, and a common correlation coefficient matrix estimated by 'correlationCoefficient'.
use.commonSigma	Use common 'sigma' (TRUE) or genewise 'sigma' (FALSE), with default value 'FALSE'.
id	A vector consisting of a subset of 1,...,J, with default value 1:J.

Value

LR	Estimation and test results: 'log-FC', 'LR statistic', 'p value'.

Examples

```
data(lung);
count = lung[,c(2:4,8:10)];
d = PLNobject(count,conditionNumber=2);
d = sizeFactor(d,maxCount=2e3);

##  Not run:
##  d = commonSigma(d);
##  d$commonCorrelation = TRUE;
##  d = correlationCoefficient(d);
##  d = genewiseSigma(d);
##  library(MASS);
##  z = mvrnorm(n=1e5,mu=rep(0,2),Sigma=d$rho);
##  d = LRtest1(d,z,use.commonSigma=FALSE,id=1:100);
```

LRtest2 *Likelihood ratio test for differential expression analysis with cluster-specific correlations.*

Description

This function calculates log-fold changes, likelihood ratio test statistics, and p-values for a list of genes. This function should be called after cluster-specific correlations are returned by 'correlationCoefficient'.

Usage

```
LRtest2(d,M,use.commonSigma,id)
```

Arguments

d	This is a PLNseq object.
M	The number of simulations used in Monte-Carlo method for calculating likelihood ratio test statistics.

use.commonSigma
: Use common 'sigma' (TRUE) or genewise 'sigma' (FALSE), with default value 'FALSE'.

id
: A vector consisting of a subset of 1,...,J, with default value 1:J.

Value

LR
: Estimation and test results: 'log-FC', 'LR statistic', 'p value'.

Examples

```
data(lung);
count = lung[,c(2:4,8:10)];
d = PLNobject(count,conditionNumber=2);
d = sizeFactor(d,maxCount=2e3);

##  Not run:
##  d = commonSigma(d);
##  J = nrow(count);
##  J1 = round(J/2);
##  d$commonCorrelation = FALSE;
##  d$cluster = c(rep(1,J1),rep(2,J-J1));
##  d = correlationCoefficient(d);
##  d = genewiseSigma(d);
##  d = LRtest2(d,M=3e4,use.commonSigma=FALSE,id=1:100);
```

---

lung *Lung cancer data*

Description

The data are from a study of the lung cancer. Six patients provided tissue samples and normal samples besides the lung tissues. The read counts were summarized by RefSeq transcript, and only those transcripts with at least 50 aligned reads for at least one tissue in each condition were provided in the table. RefSeq identifiers were mapped to the latest official gene symbols by following the user guide of the Bioconductor package 'edgeR' using the Bioconductor annotation package 'org.Hs.eg.db' (version 2.7.1). Those RefSeq identifiers not in the database were discarded, and each gene was represented by the RefSeq transcript with the greatest number of exons and the other transcripts were removed. Altogether 11,597 transcripts (genes) were kept.

Usage

```
data(lung)
```

Format

A data frame with 11,597 observations on the following 13 variables.

nameOfGene Gene name

N4 Read count for normal sample of patient 4

T4 Read count for normal sample of patient 4

N12 Read count for normal sample of patient 12

T12 Read count for tumor sample of patient 12

N13 Read count for normal sample of patient 13

T13 Read count for tumor sample of patient 13

N14 Read count for normal sample of patient 14

T14 Read count for tumor sample of patient 14

N15 Read count for normal sample of patient 15

T15 Read count for tumor sample of patient 15

N16 Read count for normal sample of patient 16

T16 Read count for tumor sample of patient 16

PLNobject *PLN object*

Description

Create a PLN object, a list containing a read count matrix 'count' and sample description matrix 'sample'.

Usage

```
PLNobject(count,conditionNumber)
```

Arguments

count	This is a matrix containing the read counts of R*I samples at J genes (R is the number of conditions in each matched sample set and I is the number of sample sets). Here columns 1 through I are for I independent samples from condition 1, columns I+1 through 2I are for I samples from condition 2 matched by samples 1 through I, ... , columns (R-1)*I+1 through R*I are for I samples from condition R matched by samples 1 through I.
conditionNumber	Number of conditions.

Value

d$count	Origninal read count matrix
d$conditionNumber	The number of conditions
d$sample	A matrix of sample information: 'SampleName', 'TotalCount', 'MedianCount'

Examples

```
data(lung);
count = lung[,c(2:4,8:10)];
d = PLNobject(count,2);
```

`sizeFactor` *Estimate size factor for each sample.*

Description

Estimate size factor for each sample using median normalization method.

Usage

```
sizeFactor(d,maxCount)
```

Arguments

`d`	This is a PLNseq object.
`maxCount`	The maximal count after shrinkage, with a default value NA (no shrinkage).

Value

`d$sample$sizeFactor`
Estimated size factors

Examples

```
data(lung);
count = lung[,c(2:4,8:10)];
d = PLNobject(count,conditionNumber=2);
d = sizeFactor(d,maxCount=2e3);
```

Index