Package 'Linnorm'

April 14, 2017

Type Package

Title Linear model and normality based transformation method (Linnorm)

Version 1.2.11

Date 2016-03-10

Author Shun Hang Yip <shunyip@bu.edu>, Panwen Wang <pwwang@pwwang.com>, Jean-Pierre Kocher <Kocher.JeanPierre@mayo.edu>, Pak Chung Sham <pcsham@hku.hk>, Junwen Wang <junwen@uw.edu>

Maintainer Ken Shun Hang Yip <shunyip@bu.edu>

Description Please note that significant updates to Linnorm are available in version 1.99.x +, we strongly suggest using the newest version.
Linnorm is an R package for the analysis of RNA-seq, scRNA-seq, ChIP-seq count data or any large scale count data. It transforms such datasets for parametric tests.
In addition to the transformtion function, the following pipelines are implemented:
1. Cell subpopluation analysis and visualization using PCA clustering,
2. Differential expression analysis or differential peak detection using limma,
3. Highly variable gene discovery and visualization,
4. Gene correlation network analysis and visualization.
5. Hierarchical clustering and plotting.
Linnorm can work with raw count, CPM, RPKM, FPKM and TPM. Additionally, Linnorm provides the RnaXSim function for the simulation of RNA-seq raw counts for the evaluation of differential expression analysis methods. RnaXSim can simulate RNA-seq dataset in Gamma, Log Normal, Negative Binomial or Poisson distributions.

Depends R(>= 3.3)

License MIT + file LICENSE

Imports Rcpp (>= 0.12.2), RcppArmadillo, fpc, vegan, mclust, apcluster, ggplot2, ellipse, limma, utils, statmod, MASS, igraph, grDevices, graphics, fastcluster, ggdendro, zoo, stats, amap

LinkingTo Rcpp, RcppArmadillo

Suggests BiocStyle, knitr, rmarkdown, gplots, RColorBrewer

VignetteBuilder knitr

biocViews Sequencing, ChIPSeq, RNASeq, DifferentialExpression, GeneExpression, Genetics, Normalization, Software, Transcription, BatchEffect, PeakDetection, Clustering, Network

NeedsCompilation yes

LazyData false

URL http://www.jjwanglab.org/Linnorm/

RoxygenNote 5.0.1

R topics documented:

`Islam2011` *scRNA-seq data from Islam et al. 2011*

Description

GEO accession GSE29087: 92 single cells (48 mouse embryonic stem cells, 44 mouse embryonic fibroblasts and 4 negative controls) were analyzed by single-cell tagged reverse transcription (STRT).

Usage

```
data(Islam2011)
```

Format

A matrix with 22936 rows (genes) and 96 columns (samples). The first 48 columns are ES cells, the following 44 columns are mouse embryonic fibroblasts and the remaining 4 columns and negative controls. Data is in raw counts format.

References

Islam et al. (2011) Genome Res 2011 Jul;21(7):1160-7. PMID: 21543516

LIHC — *Partial RNA-seq data from TCGA LIHC (Liver Hepatocellular Carcinoma)*

Description

TPM Expression data

Usage

```
data(LIHC)
```

Format

A matrix with 25914 rows (genes) and 20 columns (samples). The first 10 columns are Tumor samples, the remaining 10 columns are adjacent Normal samples. They are paired samples from 10 individuals. Data is in TPM format.

References

https://tcga-data.nci.nih.gov/

Linnorm — *Linnorm Transformation Function*

Description

This function performs the Linear model and normality based transformation method (Linnorm) for RNA-seq expression data or large scale count data.

Usage

```
Linnorm(datamatrix, showinfo = FALSE, method = "default",
  perturbation = 10, minZeroPortion = 2/3, keepAll = TRUE)
```

Arguments

datamatrix	The matrix or data frame that contains your dataset. Each row is a feature (or Gene) and each column is a sample (or replicate). Raw Counts, CPM, RPKM, FPKM or TPM are supported. Undefined values such as NA are not supported. It is not compatible with log transformed datasets.
showinfo	Logical. Show lambda value calculated. Defaults to FALSE.
method	Character. "default" or "lambda" The program will output the transformed matrix if the method is "default". If the method is "lambda", the program will output a lambda value.

perturbation	Integer >=2. To search for an optimal minimal deviation parameter (please see the article), Linnorm uses the iterated local search algorithm which perturbs away from the initial local minimum. The range of the area searched in each perturbation is exponentially increased as the area get further away from the initial local minimum, which is determined by their index. This range is calculated by 10 * (perturbation ^ index).
minZeroPortion	Double >=0, <= 1. For example, setting minZeroPortion as 0.5 will remove genes with more than half data values being zero in the calculation of normalizing parameter. It is strongly suggested to change this to 0.5 for single cell RNA-seq data. Defaults to 2/3.
keepAll	Logical. After applying minZeroPortion filtering, should Linnorm keep all genes in the results? Defualts to TRUE.

Details

If method is default, Linnorm outputs a transformed expression matrix. For users who wish to work with lambda instead, the output is a single lambda value. Please note that users with the lambda value can obtain a transformed Linnorm dataset by: log1p(lambda * datamatrix). There is no need to rerun the program if a lambda is already calculated.

Value

This function returns either a transformed data matrix or a lambda value.

Examples

```
#Obtain example matrix:
data(LIHC)
#Transformation:
transformedExp <- Linnorm(LIHC)
transformedExp <- Linnorm(LIHC, method = "lambda")
```

Linnorm.Cor *Linnorm-gene correlation network analysis.*

Description

This function first performs Linnorm transformation on the dataset. Then, it will perform correlation network analysis on the dataset.

Usage

```
Linnorm.Cor(datamatrix, input = "Raw", method = "pearson",
  showinfo = FALSE, perturbation = 10, minZeroPortion = 2/3,
  sig.q = 0.05, plotNetwork = TRUE, plotNumPairs = 5000, plotdegree = 0,
  plotname = "networkplot", plotformat = "png", plotVertexSize = 1,
  plotFontSize = 1, plot.Pos.cor.col = "red", plot.Neg.cor.col = "green",
  vertex.col = "cluster", plotlayout = "kk",
  clusterMethod = "cluster_edge_betweenness")
```

Arguments

datamatrix	The matrix or data frame that contains your dataset. Each row is a feature (or Gene) and each column is a sample (or replicate). Raw Counts, CPM, RPKM, FPKM or TPM are supported. Undefined values such as NA are not supported. It is not compatible with log transformed datasets. If a Linnorm transfored dataset is being used, please set the "input" argument into "Linnorm".
input	Character. "Raw" or "Linnorm". In case you have already transformed your dataset with Linnorm, set input into "Linnorm" so that you can input the Linnorm transformed dataset into the "datamatrix" argument. Defaults to "Raw".
method	Character. "pearson", "kendall" or "spearman". Method for the calculation of correlation coefficients. Defaults to "pearson"
showinfo	Logical. Show lambda value calculated. Defaults to FALSE.
perturbation	Integer >=2. To search for an optimal minimal deviation parameter (please see the article), Linnorm uses the iterated local search algorithm which perturbs away from the initial local minimum. The range of the area searched in each perturbation is exponentially increased as the area get further away from the initial local minimum, which is determined by their index. This range is calculated by 10 * (perturbation ^ index).
minZeroPortion	Double >=0, <= 1. For example, setting minZeroPortion as 0.5 will remove genes with more than half data values being zero in the calculation of normalizing parameter. Since this test is based on correlation coefficient, which requires more non-zero values, it is suggested to set it to a larger value. Defaults to 2/3.
sig.q	Double >=0, <= 1. Only gene pairs with q values less than this threshold will be included in the "Results" data frame. Defaults to 0.05.
plotNetwork	Logical. Should the program output the network plot to a file? An "igraph" object will be included in the output regardless. Defaults to TRUE.
plotNumPairs	Integer >= 50. Number of gene pairs to be used in the network plot. Defaults to 5000.
plotdegree	Integer >= 0. In the network plot, genes (vertices) without at least this number of degree will be removed. Defaults to 0.
plotname	Character. Name of the network plot. File extension will be appended to it. Defaults to "networkplot".
plotformat	Character. "pdf" or "png". Network plot output format. Defaults to "png".
plotVertexSize	Double >0. Controls vertex Size in the network plot. Defaults to 1.
plotFontSize	Double >0. Controls font Size in the network plot. Defaults to 1.
plot.Pos.cor.col	Character. Color of the edges of positively correlated gene pairs. Defaults to "red".
plot.Neg.cor.col	Character. Color of the edges of negatively correlated gene pairs. Defaults to "green".
vertex.col	Character. "cluster" or a color. This controls the color of the vertices. Defaults to "cluster".
plotlayout	Character. "kk" or "fr". "kk" uses Kamada-Kawai algorithm in igraph to assign vertex and edges. It scales edge length with correlation strength. However, it can cause overlaps between vertices. "fr" uses Fruchterman-Reingold algorithm in igraph to assign vertex and edges. It prevents overlatps between vertices better than "kk", but edge lengths are not scaled to correlation strength. Defaults to "kk".

`clusterMethod` Character. "cluster_edge_betweenness", "cluster_fast_greedy", "cluster_infomap", "cluster_label_prop", "cluster_leading_eigen", "cluster_louvain", "cluster_optimal", "cluster_spinglass" or "cluster_walktrap". These are clustering functions from the igraph package. Defaults to "cluster_edge_betweenness".

Details

This function performed gene correlated study in the dataset by using Linnorm transformation.

Value

This function will output a list with the following objects:

- Results: A data frame containing the results of the analysis, showing only the significant results determined by "sig.q" (see below).
- Cor.Matrix: The resulting correlation matrix between each gene.
- q.Matrix: A matrix of q values of each of the correlation coefficient from Cor.Matrix.
- Cluster: A data frame that shows which gene belongs to which cluster.
- igraph: The igraph object for users who want to draw the network plot manually.
- Linnorm: Linnorm transformed and filtered data matrix.

The "Results" data frame has the following columns:

- Gene1: Name of gene 1.
- Gene2: Name of gene 2.
- XPM1: Gene 1 average expression level in XPM. If input is raw counts or CPM, this column is in CPM unit. If input is RPKM, FPKM or TPM, this column is in the TPM unit.
- XPM2: Gene 2 average expression level in XPM. If input is raw counts or CPM, this column is in CPM unit. If input is RPKM, FPKM or TPM, this column is in the TPM unit.
- Cor: Correlation coefficient between the two genes.
- p.value: p value of the correlation coefficient.
- q.value: q value of the correlation coefficient.

Examples

```
data(Islam2011)
#Analysis on Islam2011 embryonic stem cells
results <- Linnorm.Cor(Islam2011[,1:48])
```

Linnorm.HClust *Linnorm-hierarchical clustering analysis.*

Description

This function first performs Linnorm transformation on the dataset. Then, it will perform hierarchical clustering analysis.

Usage

```
Linnorm.HClust(datamatrix, showinfo = FALSE, input = "Raw",
  perturbation = 10, minZeroPortion = 0, keepAll = TRUE,
  method_hclust = "ward.D2", method_dist = "maximum", Group = NULL,
  num_Clust = 4, ClustRect = TRUE, RectColor = "red", fontsize = 0.5,
  linethickness = 0.5)
```

Arguments

datamatrix	The matrix or data frame that contains your dataset. Each row is a feature (or Gene) and each column is a sample (or replicate). Raw Counts, CPM, RPKM, FPKM or TPM are supported. Undefined values such as NA are not supported. It is not compatible with log transformed datasets. If a Linnorm transfored dataset is being used, please set the "input" argument into "Linnorm".
showinfo	Logical. Show information about the computing process. Defaults to FALSE.
input	Character. "Raw" or "Linnorm". In case you have already transformed your dataset with Linnorm, set input into "Linnorm" so that you can input the Linnorm transformed dataset into the "datamatrix" argument. Defaults to "Raw".
perturbation	Integer >=2. To search for an optimal minimal deviation parameter (please see the article), Linnorm uses the iterated local search algorithm which perturbs away from the initial local minimum. The range of the area searched in each perturbation is exponentially increased as the area get further away from the initial local minimum, which is determined by their index. This range is calculated by 10 * (perturbation ^ index).
minZeroPortion	Double >=0, <= 1. For example, setting minZeroPortion as 0.5 will remove genes with more than half data values being zero in the calculation of normalizing parameter. Defaults to 0.
keepAll	Logical. After applying minZeroPortion filtering, should Linnorm keep all genes in the results? Defualts to TRUE.
method_hclust	Charcter. Method to be used in hierarchical clustering. (From hclust fastcluster: the agglomeration method to be used. This should be (an unambiguous abbreviation of) one of "ward.D", "ward.D2", "single", "complete", "average", "mcquitty", "median" or "centroid".) Defaults to "ward.D2".
method_dist	Charcter. Method to be used in hierarchical clustering. (From Dist amap: the distance measure to be used. This must be one of "euclidean", "maximum", "manhattan", "canberra", "binary", "pearson", "correlation", "spearman" or "kendall". Any unambiguous substring can be given.) Defaults to "maximum".
Group	Character vector with length equals to sample size. Each character in this vector corresponds to each of the columns (samples) in the datamatrix. If this is provided, sample names will be colored according to their group. Defaults to NULL.
num_Clust	Integer >= 0. Number of clusters in hierarchical clustering. No cluster will be highlighted if this is set to 0. Defaults to 4.
ClustRect	Logical. If num_Clust > 0, should a rectangle be used to highlight the clusters? Defaults to TRUE.
RectColor	Character. If ClustRect is TRUE, this controls the color of the rectangle. Defaults to "red".
fontsize	Numeric. Font size of the texts in the figure. Defualts to 0.5.
linethickness	Numeric. Controls the thickness of the lines in the figure. Defaults to 0.5.

Details

This function performs PCA clustering using Linnorm transformation.

Value

It returns a list with the following objects:

- Results: If num_Clust > 0, this outputs a named vector that contains the cluster assignment information of each sample. Else, this outputs a number 0.
- plot: Plot of hierarchical clustering.
- Linnorm: Linnorm transformed and filtered data matrix.

Examples

```
#Obtain example matrix:
data(Islam2011)
#Example:
HClust.results <- Linnorm.HClust(Islam2011, Group=c(rep("ESC",48), rep("EF",44), rep("NegCtrl",4)), num_Clus
```

Linnorm.HVar *Linnorm-Hvar pipeline for highly variable gene discovery.*

Description

This function first performs Linnorm transformation on the dataset. Then, it will perform highly variable gene discovery.

Usage

```
Linnorm.HVar(datamatrix, input = "Raw", method = "SD", spikein = NULL,
  showinfo = FALSE, perturbation = 10, minZeroPortion = 2/3,
  keepAll = FALSE, log.p = FALSE, sig.value = "p", sig = 0.05)
```

Arguments

datamatrix	The matrix or data frame that contains your dataset. Each row is a feature (or Gene) and each column is a sample (or replicate). Raw Counts, CPM, RPKM, FPKM or TPM are supported. Undefined values such as NA are not supported. It is not compatible with log transformed datasets. If a Linnorm transfored dataset is being used, please set the "input" argument into "Linnorm".
input	Character. "Raw" or "Linnorm". In case you have already transformed your dataset with Linnorm, set input into "Linnorm" so that you can input the Linnorm transformed dataset into the "datamatrix" argument. Defaults to "Raw".
method	Character. "SE" or "SD". Use Standard Error (SE) or Standard Deviation (SD) to calculate p values. Defaults to SD.
spikein	character vector. Row names of the spike-in genes in the datamatrix. If this is provided, test of significance will be performed against the spike in genes. Defaults to NULL.
showinfo	Logical. Show lambda value calculated. Defaults to FALSE.

perturbation	Integer >=2. To search for an optimal minimal deviation parameter (please see the article), Linnorm uses the iterated local search algorithm which perturbs away from the initial local minimum. The range of the area searched in each perturbation is exponentially increased as the area get further away from the initial local minimum, which is determined by their index. This range is calculated by 10 * (perturbation ^ index).
minZeroPortion	Double >=0, <= 1. For example, setting minZeroPortion as 0.5 will remove genes with more than half data values being zero in the calculation of normalizing parameter. Since this test is based on variance, which requires more non-zero values, it is suggested to set it to a larger value. Defaults to 2/3.
keepAll	Logical. After applying minZeroPortion filtering, should Linnorm keep all genes in the results? Defualts to FALSE.
log.p	Logical. Output p/q values in log scale. Defaults to FALSE.
sig.value	Character. "p" or "q". Use p or q value for highlighting significant genes. Defaults to "p".
sig	Double >0, <= 1. Significant level of p or q value for plotting. Defaults to 0.05.

Details

This function discovers highly variable gene in the dataset using Linnorm transformation.

Value

This function will output a list with the following objects:

- Results: A matrix with the results.
- plot: Mean vs Standard Deviation Plot which highlights significant genes.
- Linnorm: Linnorm transformed and filtered data matrix.

The Results matrix has the following columns:

- XPM: Average non-zero expression level in XPM. If input is raw coutns or CPM, this column is in CPM unit. If input is RPKM, FPKM or TPM, this column is in the TPM unit.
- XPM.SD: Standard deviation of average non-zero expression.
- Transformed.Avg.Exp: Average expression of non-zero Linnorm transformed data.
- Transformed.SD: Standard deviation of non-zero Linnorm transformed data.
- Normalized.Log2.SD.Fold.Change: Normalized log2 fold change of the gene's standard deviation.
- p.value: p value of the statistical test.
- q.value: q value/false discovery rate/adjusted p value of the statistical test.

Examples

```
data(Islam2011)
results <- Linnorm.HVar(Islam2011)
```

`Linnorm.limma` *Linnorm-limma pipeline for Differentially Expression Analysis*

Description

This function first performs Linnorm transformation on the dataset. Then, it will perform limma for DEG analysis. Please cite both Linnorm and limma when you use this function for publications.

Usage

```
Linnorm.limma(datamatrix, design = NULL, input = "Raw",
  output = "DEResults", noINF = TRUE, showinfo = FALSE,
  perturbation = 10, minZeroPortion = 2/3, keepAll = TRUE,
  robust = TRUE)
```

Arguments

`datamatrix`	The matrix or data frame that contains your dataset. Each row is a feature (or Gene) and each column is a sample (or replicate). Raw Counts, CPM, RPKM, FPKM or TPM are supported. Undefined values such as NA are not supported. It is not compatible with log transformed datasets. If a Linnorm transfored dataset is being used, please set the "input" argument into "Linnorm".
`design`	A design matrix required for limma. Please see limma's documentation or our vignettes for more detail.
`input`	Character. "Raw" or "Linnorm". In case you have already transformed your dataset with Linnorm, set input into "Linnorm" so that you can input the Linnorm transformed dataset into the "datamatrix" argument. Defaults to "Raw".
`output`	Character. "DEResults" or "Both". Set to "DEResults" to output a matrix that contains Differential Expression Analysis Results. Set to "Both" to output a list that contains both Differential Expression Analysis Results and the transformed data matrix.
`noINF`	Logical. Prevent generating INF in the fold change column by using Linnorm's lambda and adding one. If it is set to FALSE, INF will be generated if one of the conditions has zero expression. Defaults to TRUE.
`showinfo`	Logical. Show lambda value calculated. Defaults to FALSE.
`perturbation`	Integer >=2. To search for an optimal minimal deviation parameter (please see the article), Linnorm uses the iterated local search algorithm which perturbs away from the initial local minimum. The range of the area searched in each perturbation is exponentially increased as the area get further away from the initial local minimum, which is determined by their index. This range is calculated by 10 * (perturbation ^ index).
`minZeroPortion`	Double >=0, <= 1. For example, setting minZeroPortion as 0.5 will remove genes with more than half data values being zero in the calculation of normalizing parameter. It is strongly suggested to change this to 0.5 for single cell RNA-seq data. Defaults to 2/3.
`keepAll`	Logical. After applying minZeroPortion filtering, should Linnorm keep all genes in the results? Defualts to TRUE.
`robust`	Logical. In the eBayes function of Limma, run with robust setting with TRUE or FALSE. Defaults to TRUE.

Details

This function performs both Linnorm and limma for users who are interested in differential expression analysis. Please note that if you directly use a Linnorm Nomralized dataset with limma, the output fold change and average expression with be wrong. (p values and adj.pvalues will be fine.) This is because the voom-limma pipeline assumes input to be in raw counts. This function is written to fix this problem.

Value

If output is set to "DEResults", this function will output a matrix with Differntial Expression Analysis Results with the following columns:

- logFC: Log 2 Fold Change
- XPM: Average Expression. If input is raw count or CPM, this column has the CPM unit. If input is RPKM, FPKM or TPM, this column has the TPM unit.
- t: moderated t-statistic
- P.Value: p value
- adj.P.Val: Adjusted p value. This is also called False Discovery Rate or q value.
- B: log odds that the feature is differential

If output is set to Both, this function will output a list with the following objects:

- DEResults: Differntial Expression Analysis Results as described above.
- Linnorm: Linnorm transformed and filtered data matrix.

Examples

```
#Obtain example matrix:
data(LIHC)
#Create limma design matrix (first 10 columns are tumor, last 10 columns are normal)
designmatrix <- c(rep(1,10),rep(2,10))
designmatrix <- model.matrix(~ 0+factor(designmatrix))
colnames(designmatrix) <- c("group1", "group2")
rownames(designmatrix) <- colnames(LIHC)
#DEG analysis
DEGResults <- Linnorm.limma(LIHC, designmatrix)
```

Linnorm.PCA *Linnorm-PCA Clustering pipeline for subpopulation Analysis*

Description

This function first performs Linnorm transformation on the dataset. Then, it will perform Principal component analysis on the dataset and use k-means clustering to identify subpopulations of cells.

Usage

```
Linnorm.PCA(datamatrix, showinfo = FALSE, input = "Raw",
  perturbation = 10, minZeroPortion = 0, keepAll = TRUE, num_PC = 2,
  num_center = c(1:20), Group = NULL, Coloring = "Group",
  pca.scale = FALSE, kmeans.iter = 2000)
```

Arguments

datamatrix	The matrix or data frame that contains your dataset. Each row is a feature (or Gene) and each column is a sample (or replicate). Raw Counts, CPM, RPKM, FPKM or TPM are supported. Undefined values such as NA are not supported. It is not compatible with log transformed datasets. If a Linnorm transfored dataset is being used, please set the "input" argument into "Linnorm".
showinfo	Logical. Show information about the computing process. Defaults to FALSE.
input	Character. "Raw" or "Linnorm". In case you have already transformed your dataset with Linnorm, set input into "Linnorm" so that you can input the Linnorm transformed dataset into the "datamatrix" argument. Defaults to "Raw".
perturbation	Integer >=2. To search for an optimal minimal deviation parameter (please see the article), Linnorm uses the iterated local search algorithm which perturbs away from the initial local minimum. The range of the area searched in each perturbation is exponentially increased as the area get further away from the initial local minimum, which is determined by their index. This range is calculated by 10 * (perturbation ^ index).
minZeroPortion	Double >=0, <= 1. For example, setting minZeroPortion as 0.5 will remove genes with more than half data values being zero in the calculation of normalizing parameter. Defaults to 0.
keepAll	Logical. After applying minZeroPortion filtering, should Linnorm keep all genes in the results? Defualts to TRUE.
num_PC	Integer >= 2. Number of principal componenets to be used in K-means clustering. Defaults to 3.
num_center	Numeric vector. Number of clusters to be tested for k-means clustering. fpc, vegan, mclust and apcluster packages are used to determine the number of clusters needed. If only one number is supplied, it will be used and this test will be skipped. Defaults to c(1:20).
Group	Character vector with length equals to sample size. Each character in this vector corresponds to each of the columns (samples) in the datamatrix. This is for plotting purposes only. In the plot, the shape of the points that represent each sample will be indicated by their group assignment. Defaults to NULL.
Coloring	Character. "kmeans" or "Group". If Group is not NA, coloring in the PCA plot will reflect each sample's group. Otherwise, coloring will reflect k means clustering results. Defaults to "Group".
pca.scale	Logical. In the prcomp(for Principal component analysis) function, set the "scale." parameter. It signals the function to scale unit variances in the variables before the analysis takes place. Defaults to FALSE.
kmeans.iter	Numeric. Number of iterations in k-means clustering. Defaults to 2000.

Details

This function performs PCA clustering using Linnorm transformation.

Value

It returns a list with the following objects:

- k_means: Output of kmeans(for K-means clustering) from the stat package. Note: It contains a "cluster" object that indicates each sample's cluster assignment.

- PCA: Output of prcomp(for Principal component analysis) from the stat package.
- plot: Plot of PCA clustering.
- Linnorm: Linnorm transformed and filtered data matrix.

Examples

```
#Obtain example matrix:
data(Islam2011)
#Example:
PCA.results <- Linnorm.PCA(Islam2011)
```

RnaXSim *This function simulates a RNA-seq dataset based on a given distribution.*

Description

This function simulates a RNA-seq dataset based on a given distribution.

Usage

```
RnaXSim(datamatrix, distribution = "Poisson", NumRep = 3, NumDiff = 2000,
  NumFea = 20000, showinfo = FALSE, DEGlog2FC = "Auto",
  MaxLibSizelog2FC = 0.5)
```

Arguments

datamatrix	Matrix. The matrix or data frame that contains your dataset. Each row is a feature (or Gene) and each column is a sample (or replicate). Raw Counts, CPM, RPKM, FPKM or TPM are supported. Undefined values such as NA are not supported. It is not compatible with log transformed datasets.This program assumes that all columns are replicates of the same sample.
distribution	Character: Defaults to "Poisson". This parameter controls the output distribution of the simulated RNA-seq dataset. It can be one of "Gamma" (Gamma distribution), "Poisson" (Poisson distribution), "LogNorm" (Log Normal distribution) or "NB" (Negative Binomial distribution).
NumRep	Integer: The number of replicates. This is half of the number of output samples. Defaults to 3.
NumDiff	Integer: The number of Differentially Changed Features. Defaults to 2000.
NumFea	Integer: The number of Total Features. Defaults to 20000.
showinfo	Logical: should we show data information on the console? Defaults to FALSE.
DEGlog2FC	"Auto" or Double: log 2 fold change threshold that defines differentially expressed genes. If set to "Auto," DEGlog2FC is defined at the level where ANOVA can get a q value of 0.05 with the average expression, where the data values are log1p transformed. Defaults to "Auto".
MaxLibSizelog2FC	Double: The maximum library size difference from the mean that is allowed, in terms of log 2 fold change. Set to 0 to prevent program from generating library size differences. Defaults to 0.5.

Value

This function returns a list that contains a matrix of count data in integer raw count and a vector that shows which genes are differentially expressed. In the matrix, each row is a gene and each column is a replicate. The first NumRep (see parameter) of the columns belong to sample 1, and the last NumRep (see parameter) of the columns belong to sample 2. There will be NumFea (see parameter) number of rows. The top NumCorr of genes will be positively or negatively correlated with each other (randomly); and they are evenly separated into groups. Each group is not intended to be correlated to each other, but, by chance, it can happen.

Examples

```
#Obtain example matrix:
data(SEQC)
SampleA <- SEQC
#Extract a portion of the matrix for an example
expMatrix <- SampleA[,1:10]
#Example for Negative Binomial distribution
simulateddata <- RnaXSim(expMatrix, distribution="NB", NumRep=3, NumDiff = 200, NumFea = 2000)
#Example for Poisson distribution
simulateddata <- RnaXSim(expMatrix, distribution="Poisson", NumRep=3, NumDiff = 200, NumFea = 2000)
#Example for Log Normal distribution
simulateddata <- RnaXSim(expMatrix, distribution="LogNorm", NumRep=3, NumDiff = 200, NumFea = 2000)
#Example for Gamma distribution
simulateddata <- RnaXSim(expMatrix, distribution="Gamma", NumRep=3, NumDiff = 200, NumFea = 2000)
```

SEQC *Partial RNA-seq data from SEQC/MAQC-III Sample A*

Description

Raw Count data

Usage

```
data(SEQC)
```

Format

A matrix with 50227 rows (genes) and 10 columns (samples).

References

SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nature biotechnology 32.9 (2014): 903-914.

Index