Package 'NBBttest'

April 23, 2020

Type Package

Title Negative binomial beta t-test

Version 1.0.0

Date 2020-02-03

Suggests BiocGenerics

Author Yuan-De Tan

Maintainer Yuan-De Tan <tanyuande@gmail.com>

Description NBBttest was constructed for doing negative binomial beta t-test for multiple null hypotheses between two conditions using RNA-seq data. This package has multiple performances in differential analysis,including data quality check; data normalization; identification of differentially expressed genes at RNA level,differential splicing events at 3'UTR, 5'UTR and intron retention, differential skipping exons, differential polyadynation events, differential CRISPR screens between two conditions; heatmap of differentially expressed genes, isoforms and pathway or gene functions; NBBplot of a specified gene and gene annotation. NBBttest is the most powerful tool for identifying genes or isoforms differentially expressed in small samples so far because it is equivalent to two-way test (for example, t-test and fold change or F-test and t-test or U-test and t-test).

License GPL-3

Depends R (>= 3.5.0), stats, gplots, gtools, graphics, base, GenomicRanges, utils, grDevices

LazyLoad yes

biocViews Sequencing, DifferentialExpression, MultipleComparison, SAGE, GeneExpression, Transcription, AlternativeSplicing,Coverage, DifferentialSplicing

NeedsCompilation no

R topics documented:

`NBBttest-package` *Negative Beta Binormail t-test package*

Description

This package consists of 19 functions, of which betaparametab.R, betaparametVP.R, betaparametw.R, gbetattest.R,betattest.R, mbetattest.R, normalize.R, oddratio.R, pratio.R, omega.R, simulat.R, and smbetattest.R are used to estimate beta, alpha, weight, t-statistics,rho and omega, p-values,and multiple tests at gene level or RNA isoform level. NBBplot.R is used to visualize count data of exon isoform in given conditions.QC.R is used to peform quality check, and myheatmap.R and myheatmap2.R, pathwayheatmap.R are used to show heatmap of differential expressions of DE genes, DE RNA isoform or pathways.Run of mbetattest.R would output beta t-test results including geneid or isoformid, gene name, the other information, t-value, p-value, rho, and omega(W).

Details

Package:	NBBttest
Type:	Package
Version:	1.0.0
Date:	2018-01-11
License:	GPL-3

Author(s)

Yuan-De Tan

Maintainer: Yuan-De Tan <tanyuande@gmail.com>

References

Baggerly KA, Deng L, Morris JS, Aldaz CM (2003) Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics*, **19**: 1477-1483.
Yuan-De Tan Anita M. Chandler, Arindam Chaudhury, and Joel R. Neilson(2015) A Powerful Statistical Approach for Large-scale Differential Transcription Analysis. *Plos One*,10.1371/journal.pone.0123658.

See Also

betaparametab, betaparametVP, betaparametw, gbetattest, betattest, mbetattest, myheatmap, pathwayHeatmap myheatmap2, normalized, oddratio, NBBplot, omega,pratio, QC,simulat, smbetattest

Examples

```
data(jkttcell)
res<-mbetattest(X=jkttcell[1:500, ], nci=7, na=3,
nb=3, alpha=0.05, norm="yes",C=0,side="both",
level="isoform")
```

---

annotat *Annotation of genes within which alternative splicing occurs*

Description

Alternative splicing is detected in any element of 3'UTR, 5'UTR, exons and introns of a gene using RNA-seq data where RNA reads are mapped to a reference genome. As an example for annotation, the RNA-seq reads derived from human samples can be mapped onto human genome reference (GRCh38) using different methods, for example, HTSeq, spladder, rMAT, cufflinks, etc. These methods can detect alternative splicing sites within genes. However, these methods do not do gene annotation for users. Therefore, our NBBttest offers a function for annotating each exon, or isoform with genes

Usage

```
annotat(infile, mfile, type="gene", colunmset = "NLL")
```

Arguments

infile	input data file with ENSGid column for annotation
mfile	reference genome file with ENSG id column.
type	has three options: "gene", "isoform" or ""isof" and "exon", see details. The default is "gene".

colunmset	EGNSid column set. If the RNA count data are made by using HTSeq and DEXseq annotation file, then some of genes have many different ENSGids. For example, ENSG00000285476+ENSG00000182230+ENSG00000251623 has three ENSG ids ENSG00000285476,ENSG00000182230, and ENSG00000251623 that share one gene, so it is splited into three columns (2,3,4)in excel. The default is 2(column 2).

Details

If type = "gene", then count data of RNA reads are obtained on gene level, annotation would be executed on gene level or if type = "isof" or "isoform", then RNA reads were mapped onto elements (for example, 3'UTR,5'UTR, exon or cassette, intron) within genes and annotation would be executed on isoform level or if type = "exon", then RNA count data were obtained by mapping RNA reads onto exome by DEXseq and annotation would be done on exon level defined by DEXSeq. Note that GRCh38 is too big so it was removed from data. User may request to get it from yuande/github.

Value

return original data with an additional column for gene.

Author(s)

Yuan-De Tan <tanyuande@gmail.com>

Examples

```
#data(DDX39_100)
#gtf<-data(GRCh38)
#DDX39_100a<-annotat(infile=DDX39_100,mfile=gtf,type="gene")
data(prime3_PRP8_50)
#prime3_PRP8_50a<-annotat(infile=prime3_PRP8_50,mfile=gtf,type="isoform")
#data(DDX39_cnt_49)
## DDX39_cnt_49a<-annotat(infile=DDX39_cnt_49,mfile=gtf,type="exon",
##colunmset=c(2,3,4,5,6,7,8,9,10))
```

---

betaparametab *Estimation of parameters alpha (α) and beta (β) of beta distribution*

Description

Parameters alpha(a) and beta (b) in betat distribution are estimated by using an iteration algorithm.

Usage

```
betaparametab(xn, w, P, V)
```

Arguments

xn	column vector, a set of library sizes.
w	column vector, a set of weights.
P	proportion of counts of a gene or an isoform.
V	variance of proportions of counts of a gene or an isoform over m replicate libraries in a condition.

Value

return parameters a and b.

Author(s)

Yuan-De Tan <tanyuande@gmail.com>

References

Baggerly KA, Deng L, Morris JS, Aldaz CM (2003) Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics* **19**: 1477-1483.

See Also

betaparametVP, betaparametw

Examples

```
XX<-c(2000,2000,2000)
p<-0.15
V=0.004
w<-c(0.3,0.3,0.3)
betaparametab(xn=XX,w=w,P=p,V=V)
#[1] 1.145868 6.493254
```

---

betaparametVP *Estimation of parameters V and P in count data of RNA reads*

Description

This function is used to estimate parameters P and V by optimalizing estimates of parameters: alpha and beta.

Usage

```
betaparametVP(X, NX)
```

Arguments

X	count dataset derived from m replicate libraries in one condition.
NX	vector of m library sizes. Library size is sum of counts over the whole library.

Details

Count data of *RNA* sequence reads are assumed to follow binomial distribution with parameters (P) and (n) or negative binomial distribution with parameters (P) and (r) , while P is frequency of a gene or an isoform in RNA sequence population and assumed to follow beta distribution with parameters alpha (a) and beta(b). Parameters P and V are estimated by using optimal estimation of parameters a and b. The optimal method is an iteration algorithm drived by weighting proportion of gene or isoform in each replicate library. This is a large-scale method for estimating these parameters. Estimation of parameters P and V is core of the multiple beta t-test method because P and V will be used to calculate t-value.

Value

return a list:

P	N proportions estimated.
V	N variances estimated.

Note

betaparametVP requres functions betaparametab and betaparametw.

Author(s)

Yuan-DE Tan <tanyuande@gmail.com>

References

Baggerly KA, Deng L, Morris JS, Aldaz CM (2003) Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics*, **19**: 1477-1483.
Yuan-De Tan, Anita M. Chandler, Arindam Chaudhury, and Joel R. Neilson(2015) A Powerful Statistical Approach for Large-scale Differential Transcription Analysis. *Plos One*,10.1371/journal.pone.0123658.

See Also

betaparametab, betaparametw

Examples

```
data(jkttcell)
X<-jkttcell[1:500,]
na<-3
nb<-3
cn<-length(X[1,])
rn<-length(X[,1])
XC<-X[,1:(cn-na-nb)]
XX<-X[,(cn-na-nb+1):cn]
n<-na+nb
XA<-XX[,1:na]
SA<-apply(XA,2,sum)
PA<-betaparametVP(XA,SA)
```

---

betaparametw *Estimation of proportion weights*

Description

Function betaparametw is used to calculate weight.

Usage

```
betaparametw(xn, a, b)
```

Arguments

xn	a vector of m library sizes. Library size is sum of counts over the whole library.
a	parameter alpha(α) in beta distribution derived from output of function betaparametab.
b	parameter beta (β)in beta distribution derived from output of function betaparametab.

Details

alpha and beta (α,β) are used to calculate weight. Then weight is in turn used to correct bias of estimation of alpha and beta in betaparametab function.

Value

return weight(W).

Author(s)

Yuan-De Tan <tanyuande@gmail.com>

References

Baggerly KA, Deng L, Morris JS, Aldaz CM (2003) Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics*, **19**: 1477-1483.
Yuan-De Tan, Anita M. Chandler, Arindam Chaudhury, and Joel R. Neilson(2015) A Powerful Statistical Approach for Large-scale Differential Transcription Analysis. *Plos One*. 2015 DOI: 10.1371/journal.pone.0123658.

See Also

betaparametab and betaparametVP.

Examples

```
XX<-c(2000,2000,2000)
a<-1.1458
b<-6.4932
betaparametw(xn=XX,a=a,b=b)
#[1] 0.3333333 0.3333333 0.3333333
```

---

betattest *Beta t-test*

Description

Beta t-test and degree of freedom for each gene or isoform are calculated in this function.

Usage

```
betattest(X, na, nb, NX=100, level)
```

Arguments

X	count data of RNA sequence reads containing N genes (or isoforms).
na	number of replicate libraries in condition A.
nb	number of replicate libraries in condition B.
NX	numeric value. It is used at level="isoform". NX=100 is default but does not used at any level.
level	string, has three options: "isoform" or "sgRNA"

Details

In beta t-test,

$$t = \frac{(P_A - P_B)}{\sqrt{(V_A + V_B)}}$$

where P_A and P_B are proportions of a gene or an isoform in conditions A and B, V_A and V_B are variances of this gene or isoform in conditions A and B, respectively. They are output of betaparametVP.

Value

return two lists:

t	t-value list.
df	df list. df is degree of freedom.

Note

In our method, pooled standard error > 0 in any case, so the t-statistics alway has definition.

Author(s)

Yuan-De Tan <tanyuande@gmail.com>

References

Baggerly KA, Deng L, Morris JS, Aldaz CM (2003) Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics*, **19**: 1477-1483.
Yuan-De Tan, Anita M. Chandler, Arindam Chaudhury, and Joel R. Neilson(2015) A Powerful Statistical Approach for Large-scale Differential Transcription Analysis. *Plos One*. 2015 DOI: 10.1371/journal.pone.0123658.

See Also

pratio, oddratio.

Examples

```
data(jkttcell)
X<-jkttcell[1:1000,]
na<-3
nb<-3
cn<-ncol(X)
rn<-nrow(X)
```

```
XC<-X[,1:(cn-na-nb)]
XX<-X[,(cn-na-nb+1):cn]
betattest<-betattest(X=XX, na=3,nb=3, level="isoform")
```

`gbetattest` *Beta t-tests within a group*

Description

Beta t-tests are conducted within a group, a gene or within library.

Usage

```
gbetattest(xx, W, nci, na, nb, level, C, side)
```

Arguments

`xx`	a datasheet consisting of n columns and m rows. Columns contain data information columns and count data columns and rows are variables (genes, isoforms, or both).
`W`	numeric value. It is omega estimated from null simulation.
`nci`	int numeric value indicating number of data information column.
`na`	int numeric value indicating number of replicates in condition a.
`nb`	int numeric value indicating number of replicates in condition b.
`level`	string value. It has 6 options: "isoform", "sgRNA", "RNA", "polyA.gene", "CRISPR.gene" and "splicing.gene".
`C`	float numeric value for specifying a multiple procedure. C=0 tells mbetattest to perform single tests, C=1.22 tells mbetattest to perform BH correction of pvalues, C>1000 tells mbetattest to perform Bonferroni correction of pvalues.
`side`	string value for specifying one-tail test or two-tail test: side="up" for left-tail test, side="down" for right-tail test and side="both" for two-tail tests.

Details

Data inputted are separated into groups as level comamd. Beta t-test will be conducted in the group. If level="RNA", then beta t-tests will be conducted within a whole library or whithin whole data. If level="isoform", then data will be sparated in two parts: gene single-isoform dataset and gene multi-isoform data. For gene single-isoforms, data are as a group and beta t-tests will be performed in the group. For the gene multi-isoforms, t-test will be performed within genes. If level="polyA.gene" or "CRISPR.gene", then t-test will be performed at gene level. If level="splicing.gene", then t-values and p-values will be chosen from gene groups with the least p-values.

Value

return a list containing dataset, t-values, corrected p-values, rhos and w.

Author(s)

Yuan-De Tan <tanyuande@gmail.com>

See Also

betattest

Examples

```
data(jkttcell)
colnames(jkttcell)[3]<-"Gene"
res.isfo<-gbetattest(xx=jkttcell[1:500,], W=1, nci=7,
na=3, nb=3, level="isoform", C=0,side="both")

res.gene<-gbetattest(xx=jkttcell[1:500,], W=1, nci=7,
na=3, nb=3, level="polyA.gene", C=0,side="both")
```

---

jkttcell *Jurkat T-cell transcritomic data*

Description

The data are transcriptomic count data of *RNA* reads generated by next generation sequencing from Jurkat T-cells.

Usage

```
data("jkttcell")
```

Format

A data frame with 13409 observations on the following 13 variables.

`tagid` : a numeric vector

`geneid` : a numeric vector

`name` : a string vector

`chr` : a string vector

`strand` : a charactor vector

`pos` : a numeric vector

`anno` : a string vector

`Jurk.NS.A` : a numeric vector

`Jurk.NS.B` : a numeric vector

`Jurk.NS.C` : a numeric vector

`Jurk.48h.A` : a numeric vector

`Jurk.48h.B` : a numeric vector

`Jurk.48h.C` : a numeric vector

Details

The data are count data generated by next generation sequencing from Jurkat T-cells. The T-cells were treated by resting and stimulating with *CD3/CD28* for 48 hours. The data have 7 columns for the information of *poly(A)* site: tagid, geneid, gene name, chromosome, strand, *poly(A)* site position, *poly(A)* site annotation and 6 columns for count data: `Jurk.NS.A`, `Jurk.NS.B`, `Jurk.NS.C`, `Jurk.48h.A`, `Jurk.48h.B`, `Jurk.48h.C`. where NS means Normal state and 48h means 48 hours after *CD3/CD28* stimulatuin of T-cells. 13409 *RNA* isoforms were detected to have alternative *poly(A)* sites.

Value

ID, information, count data of RNA reads

Source

Real transcriptomic count data

References

Yuan-De Tan Anita M. Chandler, Arindam Chaudhury, and Joel R. Neilson(2015) A Powerful Statistical Approach for Large-scale Differential Transcription Analysis. *Plos One*. DOI: 10.1371/journal.pone.0123658.

Examples

```
data(jkttcell)
## maybe str(jkttcell) ; plot(jkttcell) ...
```

---

mbetattest *Performance of multiple beta t-test on count data*

Description

This function is used to peform multiple beta t-test method on real count data. The result lists geneid or isoformid, gene name, the other information, t-value, p-value, rho, and w (ω).

Usage

```
mbetattest(X, nci, na, nb, alpha=0.05, norm="no",
C=0, side="both", level="sgRNA")
```

Arguments

`X`	count data of RNA sequence reads with na replicates in condition A and nb replicates in condition B.
`nci`	int numeric value: number of columns for data information, such as geneID, isoformID, gene name etc.
`na`	int numeric value: number of replicate libraries in condition A.
`nb`	int numeric value: number of replicate libraries in condition B.
`alpha`	float numeric value, a probabilistic threshold. The value must be in [0,1]. User can set alpha=0.05 or 0.01 or the other values. Defalt value is 0.05

norm	logistic value:"yes" or "no". If norm="yes", the count data will be normalized and mbetattest will work on normalized data, if norm="no", then mbetattest will work on unnormalized data.
C	real numeric value for specifying a multiple procedure.
side	string value for specifying tail(s) of t-distribution. side="up" for left tail, side="down" for right tail.
level	string value for specifying which level mbetattest work on. In the current version, level has 6 options: "isoform", "sgRNA", "RNA", "splicing.gene","polyA.gene", and "CRISPR.gene".

Details

see MBttest2-manual.

Value

return a data and result list: data columns, t-values, rho.

Author(s)

Yuan-De Tan <tanyuande@gmail.com>

See Also

smbetattest, mtpvadjust, normalized,omega.

Examples

```
data(jkttcell)
res<-mbetattest(X=jkttcell[1:500, ], nci=7, na=3,
nb=3, alpha=0.05, norm="yes", C=0,side="both",
level="isoform")
```

---

mtpvadjust *Adjust p-values for multiple comparisons*

Description

Given a set of p-values and chosen a C-value, returns a set of adjusted p-values

Usage

```
mtpvadjust(pv, C)
```

Arguments

pv	numeric vector of p-values (possibly with NAs). Any other R object is coerced by as.numeric.
C	real numeric value for specifying a multiple procedure.

Details

C=0 indicates that p-values are not adjusted, C=1.22 indicates that p-values are adjusted with Benjamini & Hochberg (1995) ("BH"). The adjusted p-values are called "fdr". When C>=1000, p-values are adjusted with the Bonferroni method. C<1.22 indicates that p-values are adjusted by a relaxed BH method while C>1.22, p-values are adjusted by a more strict BH method.

Value

A numeric vector of corrected p-values (of the same length as p, with names copied from p)

Author(s)

Yuan-De Tan
<tanyuande@gmaail.com>

References

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B, 57, 289 - 300. jstor.
Yuan-De Tan, Hongyan Xu; A general method for accurate estimation of false discovery rates in identification of differentially expressed genes, Bioinformatics, Volume 30, Issue 14, 15 July 2014, Pages 2018 - 2025, doi.

See Also

p.adjust

Examples

```
set.seed(123)
x <- rnorm(50, mean = c(rep(0, 25), rep(3, 25)))
p <- 2*pnorm(sort(-abs(x)))
p.1.22<-mtpvadjust(pv=p, C=1.22)
p.0<-mtpvadjust(pv=p, C=1000)
```

---

myheatmap *Heatmap*

Description

This function is used to display heatmap of differential expressions of genes or isoforms detected by MBttest in the real count data.

Usage

```
myheatmap(dat, IDcol, nci, r, r1, r2,
colrs="greenred", rwcex=2.0,clcex=2.0,x=10, tree="both",
method="euclidean", ky=1.5, rowBarColor=NULL,
colBarColor=NULL,  labrow="yes", labcol="yes", rsort="yes",
adjrow=c(0.3, 0.0 ), adjcol = c(1, 1) , rwangle=30,
clangle=30,  maptitle)
```

Arguments

dat	MBttest object that is outputted by mbetattest, includes information columns, data columns, t-value column, pvalue column.
IDcol	numeric integer indicating which column is ID column where IDs will be shown in heatmap rows, is required.
nci	numeric integer indicating number of informaction (gene or isoform annotation) columns in the data, is required.
r	numeric integer indicating column number of count data including empty columns if there are multiple datasets with the same row names.
r1	numeric integer indicating replicate number in sample 1.
r2	numeric integer indicating replicate number in sample 2.
colrs	heatmap colors. User has 8 options: "redgreen", "greenred", "redblue", "bluered", "cm.colors", "terrain.colors","topo.colors", and"heat.colors".
rwcex	numeric argument: nonnegative value, used as cex.axis in for the row axis labeling. The default=1.8.
clcex	numeric argument: nonnegative value, used as cex.axis for the column axis labeling. The default=1.8.
x	numeric argument: nonnegative value, used as argument of cm.colors(x), terrain.colors(x) and topo.colors(x), the default value is 10.
tree	tree(s) drawn on row or colum or both. User has four options: "both" for drawing trees on both row and column,"row" for drawing tree only on row,"column" for drawing tree only on column, and "none" for no tree specified on rows and columns. Default is "both".
method	method to be chosen to calculate distance between columns or rows. It has four options: "euclidean", "pearson","spearman" and "kendall". The latter three are d=1-cc where cc is correlation coefficients. Default is "euclidean".
ky	numeric argument: nonnegative value is used to determine key size. The defalut =1.5
rowBarColor	(optional) character vector for RowSideColors and colRow. Length of rowBarColor equals to gene or isoform number or row number. rowBarColor contains the color names of classes or typpes or groups of row names and show row color side bars and color row names. If row names, for example, genes in row are not classified, then we suggest user use its default value: rowBarColor is NULL.
colBarColor	(optional) character vector for colSideColors and colCol. Length of colBarColor equals to sample number. colBarColor contains the color names of classes or typpes or groups of coloumn names and show column color side bars and color column names. If column names, for example, samples in column are just two types, then we suggest user use its default value: colBarColor=NULL.
labrow	a string, logistic value. Rows on heatmap are labeled with genes or targets if labrow="yes", otherwise, the rows are not labeled.
labcol	a string, logistical value. columns on heatmap are labeled with samples or treatments if labcol="yes", otherwise, the columns are not labeled.
rsort	logistical value. If choose tree="both" or "row", then rsort does not work. However, if tree ="none" or "columnn", then rsort ="yes" will force rows to be sorted in descent way. The default is "yes"

adjrow	two numeric values. The first value is used to adjust left or right position (x-axis) of labels, and the second value is used to adjust up or down position (y-axis) of labels. The fadault values are c(0.4, 1).
adjcol	two numeric values. The first value is used to adjust left or right position (x-axis) of labels, and the second value is used to adjust up or down position (y-axis) of labels. The fadault values are c(1, 1).
rwangle	angle of xlab under heatmap. The Default value is 30.
clangle	angle of ylab. The default value is 30
maptitle	string argument for giving heatmap title, default is set to be empty string ' '.

Details

This function uses sigificance to choose data and then to normalize the selected data by using z-scale. This function has multiple options to select map color, distance, cluster and x- and y-lab angles. If r>(r1+r2), then data dat are multiple datasets.

Value

no return value but create a heatmap.

warning

For multiple datasets, R console would show Warning message: In image.default(z = matrix(z, ncol = 1), col = col, breaks = tmpbreaks, : unsorted 'breaks' will be sorted before use. But it doesn't matter, it works.

Note

requres gplots and grDevices.

Author(s)

Yuan-De Tan <tanyuande@gmail.com>

See Also

heatmap.2,myheatmap2

Examples

```
require(gplots)
data(result)
colclass=c("1","1","1","2","2","2")

myheatmap (dat=result, IDcol=1, nci=7, r=6, r1=3,
r2=3, colrs="bluered", rowBarColor=NULL,
colBarColor=colclass, labrow="yes", labcol="yes",
rsort="yes", adjrow=c(0.3, 0.0 ),
adjcol = c(1, 1) , maptitle="My heatmap")
```

myheatmap2 *Heatmap2*

Description

This function is used to display heatmap of differential expressions of genes or isoforms detected by MBttest in the real count data.

Usage

```
myheatmap2(dat, IDcol=1, nci=NULL, r, colrs="greenred", rwcex=2.8,
clcex=2.8, x=10, tree="both", method="euclidean", ky=1.5, rowBarColor=NULL,
colBarColor=NULL, labrow="yes", labcol="yes", rsort="yes", adjrow=c(0.2, 0.0 ),
adjcol = c(1, 1) , rwangle=30, clangle=30, maptitle="")
```

Arguments

dat	dat is object of MBttest, including information columns, data columns, tvalue column, pvalue column.
nci	numeric integer, number of columns for information of genes or isoforms, like geneid, tagetid, straind, annotation etc.
IDcol	numeric integer, indicating which column is ID column where IDs will be shown in heatmap rows.
r	numeric integer, indicate column number of count data including empty columns if there are multiple datasets with the same row names.
colrs	heatmap colors. User has 8 options: "redgreen", "greenred", "redblue", "bluered", "cm.colors","terrain.colors","topo.colors", and"heat.colors". Default colrs is "redgreen".
rwcex	numeric argument: nonnegative number, used as cex.axis for the row axis labeling. The default value is 1.8.
clcex	numeric argument: nonnegative number, used as cex.axis for the column axis labeling. The default value is 1.8.
x	numeric argument: nonnegative number, used as argument of cm.colors(x), terrain.colors(x) and topo.colors(x), the default value is 10.
tree	tree(s) drawn on row or column or both. User has four options: "both" for drawing trees on both row and column,"row" for drawing tree only on row,"column" for drawing tree on only column, and "none" for no tree specified on rows and columns. If tree =both, then colomns and rows are sorted by trees. If tree = "row", the colmns are not sorted. If tree ="column", then rows are not sorted. Default is "both".
method	method to be chosen to calculate distance between columns or rows. It has four options: "euclidean", "pearson","spearman" and "kendall". The latter three are d=1-cc where cc is correlation coefficients. Default is "euclidean".
ky	numeric argument: nonnegative number, is used to determine key size. The defalut =1.5.

rowBarColor	(optional) character vector for RowSideColors and colRow. Length of rowBarColor equals to gene or isoform number or row number. rowBarColor contains the color names of classes or typpes or groups of row names and show row color side bars and color row names. If row names, for example, genes in row are not classified, then we suggest user use its default value: rowBarColor=NULL.
colBarColor	(optional) character vector for colSideColors and colCol. Length of colBarColor equals to sample number. colBarColor contains the color names of classes or typpes or groups of coloumn names and show column color side bars and color column names. If column names, for example, samples in column are just two types, then we suggest user use its default value: colBarColor=NULL.
labrow	a string, logistic value. Rows on heatmap are labeled with genes or targets if labrow="yes", otherwise, the rows are not labeled.
labcol	a string, logistical value. Columns on heatmap are labeled with samples or treatments if labcol="yes", otherwise, the columns are not labeled.
adjrow	two numeric values. The first value used to adjust left or right position (x-axis) of labels and the second value is used to adjust up or down position (y-axis) of labels. The fadault values are c(0.3, 0.0).
adjcol	two numeric values. The first value used to adjust left or right position (x-axis) of labels, and the second value is used to adjust up or down position (y-axis) of labels. The fadault values are c(1, 1).
rsort	logistical value. If choose tree="both" or "row", then rsort does not work. However, if tree ="none" or "columnn", then rsort ="yes" will force rows to be sorted in descent way. The default is "yes"
rwangle	heatmap object: angle of xlab. Thed default value is 30.
clangle	heatmap object: angle of ylab. The default value is 30.
maptitle	string argument for giving heatmap title. The default value is set to be empty string ' '.

Details

This function uses pvalue to choose genes or isoforms in the data and then to normalize the selected data by using n-scale. Different from z-score, n-score does not follow standard normal distribution with mean = 0 and variance =1 for all rows but it has the same largest count in all rows and shows multiple colors for numeric difference between two conditions. This function has multiple options to select map color, distance, cluster and x- and y-lab angles. This function can be able to display multiple datasets in two ways: if multiple datasets have the same row names or features, the these datasets are put onto the different columns separated with empty column named with dataset names. If multiple datasets have the same column names of the datasets, then put them on different rows saparated with empty rows named with dataset names or whatever names user specifies.

Value

not return value but create a heatmap.

warning

For multiple datasets, R console would show 1: In FUN(newX[, i], ...) : no non-missing arguments to max; returning -Inf. But it doesn't matter, it works.

Note

requires gplots and grDevices. If the data for heatmap are multiple datasets, then tree="none" and sort="no", otherwise, myheatmap2 will get error. So before performing myheatmap2, user should sort the data in excel.

Author(s)

<tanyuande@gmail.com>

See Also

heatmap.2, grDevices, and myheatmap.

Examples

```
require(gplots)
data(result)
colclass=c("1","1","1","2","2","2")

myheatmap2(dat=result, IDcol=1, nci=7, r=6,
colrs="greenred", rwcex=1.8, clcex=1.8, x=10, tree="both",
method="euclidean", ky=1.5, rowBarColor=NULL,
colBarColor=colclass, labrow="yes", labcol="yes",
adjrow=c(0.2, 0.0 ), adjcol = c(1, 1) ,
rwangle=0, clangle=30,maptitle="My heatmap2")
```

NBBplot *Plot differential expression of exons within a specified gene using result outputed by NBBttest.*

Description

After performing NBBttest, NBBplot can be used to show differential expression of exons within a specified gene in na and nb replicates between conditions A and B.

Usage

```
NBBplot(res, gene, nci, na, nb, C1, C2)
```

Arguments

res	object of NBBttest containing information of genes including gene name, strand, chromosome,exons, introns,data, and t-value, p-value, significance etc.
gene	gene name or symbol specified by user.
nci	number of columns for gene information.
na	replicate number in condition A
nb	replicate number in condition B
C1	name for condition 1(A)
C2	name for condition 2(B)

Details

NBBplot consists of two parts: top is expression value of each exon in each replicate in two conditions marked in red and blue and bottom is boxes for exon and solid lines for introns. Differential expression of an exon is marked in red.

Value

output NBBplot figure for given gene.

Author(s)

Yuan-De Tan <tanyuande@gmail.com>

See Also

plotDEXSeq

Examples

```
#data(TAIR10)

#NBBplot(res=TAIR10, gene="AT1G01060", nci=7, na=3, nb=3, C1="control", C2="heat shock")
```

---

normalized *Normalization of data*

Description

Function normalize makes all libraries in dataset have the same library size.

Usage

```
normalized(dat, nci, m=0, lg2="no")
```

Arguments

dat	count data of RNA reads.
nci	number of columns for the information of genes or isoforms in dataset.
m	numeric value for choosing genes or isoforms. If user wants to discard genes or isoforms with mean < 5, then m = 5. The default value is 0.
lg2	logistic value. lg2="yes"indicates that data are transformed in logarithm of 2.

Details

Due to difference in RNA abstraction between libraries or cell samples or tissues, PCR amounts of RNA libraries would have difference. This difference is not due to biological effects. To correctly compare differential expressisons of genes between conditions or samples, one must should give the same RNA abstraction in all given samples. This is impossible. To address this problem, only one way is to normalize these count data in all given samples so that all experimental samples (libraries) have the same total counts. DESeq and DESeq2 can also perform normalization on data, but the normalized data cannot have the same library size among the replicate samples. So DESeq and DESeq2 normalization is not correct.

Value

output a standard datasheet.

Author(s)

Yuan-De Tan <tanyuande@gmail.com>

References

Yuan-De Tan Anita M. Chandler, Arindam Chaudhury, and Joel R. Neilson(2015) A Powerful Statistical Approach for Large-scale Differential Transcription Analysis. *Plos One*, 10.1371/journal.pone.0123658.
Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol*, 11: R106.

Examples

```
data(jkttcell)
njkttcell<-normalized(dat=jkttcell,nci=7)
```

`oddratio` *Calculation of zeta (ζ)*

Description

ζ is used to measure homogeneity intensity of two subdatasets. If ζ is larger than 1, these two subdatasets have good homogeneity; otherwise, $\zeta < 1$ indicates that two subdatasets have poor homogeneity (big noise).

Usage

```
oddratio(XX, na, nb)
```

Arguments

`XX`	nonnegative count data of RNA reads generated by next generation sequencing.
`na`	number of replicate libraries in condition A.
`nb`	number of replicate libraries in condition B.

Details

Zeta (ζ) is defined as

$$\zeta = ln\left(1 + \frac{\bar{X} \times \sigma^2 + 1}{\bar{X}_A \times \sigma_A^2 + \bar{X}_B \times \sigma_B^2 + 1}\right)$$

where ζ is different from ψ. If two subdatasets have big a gap and good homogeneity,then seta value has much larger than 1.

Value

zeta vector.

Author(s)

Yuan-De Tan <tanyuande@gmail.com>

References

Yuan-De Tan Anita M. Chandler, Arindam Chaudhury, and Joel R. Neilson(2015) A Powerful Statistical Approach for Large-scale Differential Transcription Analysis. *Plos One*. 2015 DOI: 10.1371/journal.pone.0123658.

See Also

pratio, mbetattest.

Examples

```
XX<-matrix(NA,2,8)
XX[1,]<-c(112,122, 108,127,302, 314, 322, 328)
XX[2,]<-c(511, 230, 754, 335,771, 842, 1014,798)
#XX
#     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
#[1,]  112  122  108  127  302  314  322  328
#[2,]  511  230  754  335  771  842 1014  798
oddratio(XX=XX,na=4,nb=4)

#[1] 3.9432676 0.8762017

# see example in mbetattest
```

---

omega *Omega calcularion*

Description

Omega (ω) function is a function that is used to estimate omega using simulate null data from negative bionomial distribution. Omege is a null rho that is used as a threshold for real rho. Simulation is dependent on the original data.

Usage

```
omega(XX, nci, r1, r2, sn, alpha = 0.05)
```

Arguments

XX	MBttest object, the real dataset.
nci	number of culomns for information of data, like gene id, isoform id, gene name, etc.
r1	size of sample 1 or number of replicates in condition 1.
r2	size of sample 2 or number of replicates in condition 2.
sn	number of simulations specified.
alpha	significance level of test. Default is 0.05.

Details

This function is to use null data to calculate omega value with rho = 1.

Value

return a numeric value, omega.

Author(s)

Yuan-De Tan <tanyuande@gmail.com>

References

Yuan-De Tan Anita M. Chandler, Arindam Chaudhury, and Joel R. Neilson(2015) A Powerful Statistical Approach for Large-scale Differential Transcription Analysis. *Plos One*. 2015 DOI: 10.1371/journal.pone.0123658.

See Also

pratio and oddratio

Examples

```
data(jkttcell)

w<-omega(XX=jkttcell[1:1000,],nci=7,r1=3,r2=3,sn=2,alpha=0.05)
```

---

pathwayHeatmap *Heatmap for pathways found by gene ontology analysis*

Description

This function is used to show differential expressions of pathways or functions between conditions. These pathways or functions were detected by function annotation or gene ontology methods such as David function analysis tools or Ingenuity pathway analysis. Pathway score or pathway value is weighted mean of expression amount of genes in a pathway or a function. The weigths of genes are given by p-values of enrichment or hit in function analysis.

Usage

```
pathwayHeatmap(dat, pathway, nci, r1, r2, colclass, rowclass,
colrs, maptitle)
```

Arguments

dat	count dataset that contains a column for gene name and count data columns in two conditions.
pathway	a list that lists a function colomn and at least two gene columns. Row is function name.
nci	number of columns for gene information such as gene name, strand, chromosome, etc in dataset dat.

r1	number of replicates in condition 1 in dataset dat.
r2	number of replicates in condition 2 in dataset dat.
colclass	colonm class, equivalent to replicates in two conditions, such as c(1,1,1,1,2,2,2,2),meaning that condition A has 4 replicates and condition B has 4 replicates. Default =NULL.
rowclass	row class, pathway class, or pathway module. Default = NULL.
colrs	heatmap colors. User has 8 options: "redgreen", "greenred", "redblue", "bluered", "cm.colors", "terrain.colors","topo.colors", and"heat.colors".
maptitle	title, default = "".

Note

requres gplots and grDevices.

Author(s)

Yuan-De Tan <tanyuande@gmail.com>

See Also

heatmap.2,myheatmap, myheatmap2

Examples

```
require(gplots)
data(upGAm)
data(pathwy.A.up)
pathwayup<-pathwy.A.up
colclass=c("1","1","1","1","2","2","2","2","2","2")
pathwayHeatmap(dat=upGAm,pathway=pathwayup,nci=1,r1=4,r2=6,
colclass=colclass,rowclass=NULL,colrs="greenred",
maptitle="pathway up-expression in Group A")
```

---

pathwy.A.up *Pathway or function data*

Description

pathway data are functions or pathways of up-regulated genes given by gene annotation tool such as DAVID gene function annotion.

Usage

```
data("pathwy.A.up")
```

Format

A data frame with 39 observations on the following 159 variables.

`V1` a factor with levels

`V2` a factor with levels

`V3` a factor with levels

`V4` a factor with levels

`V5` a factor with levels

`V6` a factor with levels

`V7` a factor with levels

`V8` a factor with levels

`V9` a factor with levels

`V10` a factor with levels

`V11` a factor with levels

`V12` a factor with levels

`V13` a factor with levels

`V14` a factor with levels

`V15` a factor with levels

`V16` a factor with levels

`V17` a factor with levels

`V18` a factor with levels

`V19` a factor with levels

`V20` a factor with levels

`V21` a factor with levels

`V22` a factor with levels

`V23` a factor with levels

`V24` a factor with levels

`V25` a factor with levels

`V26` a factor with levels

`V27` a factor with levels

`V28` a factor with levels

`V29` a factor with levels

`V30` a factor with levels

`V31` a factor with levels

`V32` a factor with levels

`V33` a factor with levels

`V34` a factor with levels

`V35` a factor with levels

`V36` a factor with levels

`V37` a factor with levels

`V38` a factor with levels

`V39` a factor with levels
`V40` a factor with levels
`V41` a factor with levels
`V42` a factor with levels
`V43` a factor with levels
`V44` a factor with levels
`V45` a factor with levels
`V46` a factor with levels
`V47` a factor with levels
`V48` a factor with levels
`V49` a factor with levels
`V50` a factor with levels
`V51` a factor with levels
`V52` a factor with levels
`V53` a factor with levels
`V54` a factor with levels
`V55` a factor with levels
`V56` a factor with levels
`V57` a factor with levels
`V58` a factor with levels
`V59` a factor with levels
`V60` a factor with levels
`V61` a factor with levels
`V62` a factor with levels
`V63` a factor with levels
`V64` a factor with levels
`V65` a factor with levels
`V66` a factor with levels
`V67` a factor with levels
`V68` a factor with levels
`V69` a factor with levels
`V70` a factor with levels
`V71` a factor with levels
`V72` a factor with levels
`V73` a factor with levels
`V74` a factor with levels
`V75` a factor with levels
`V76` a factor with levels
`V77` a factor with levels
`V78` a factor with levels

V79 a factor with levels
V80 a factor with levels
V81 a factor with levels
V82 a factor with levels
V83 a factor with levels
V84 a factor with levels
V85 a factor with levels
V86 a factor with levels
V87 a factor with levels
V88 a factor with levels
V89 a factor with levels
V90 a factor with levels
V91 a factor with levels
V92 a factor with levels
V93 a factor with levels
V94 a factor with levels
V95 a factor with levels
V96 a factor with levels
V97 a factor with levels
V98 a factor with levels
V99 a factor with levels
V100 a factor with levels
V101 a factor with levels
V102 a factor with levels
V103 a factor with levels
V104 a factor with levels
V105 a factor with levels
V106 a factor with levels
V107 a factor with levels
V108 a factor with levels
V109 a factor with levels
V110 a factor with levels
V111 a factor with levels
V112 a factor with levels
V113 a factor with levels
V114 a factor with levels
V115 a factor with levels
V116 a factor with levels
V117 a factor with levels
V118 a factor with levels

`V119` a factor with levels
`V120` a factor with levels
`V121` a factor with levels
`V122` a factor with levels
`V123` a factor with levels
`V124` a factor with levels
`V125` a factor with levels
`V126` a factor with levels
`V127` a factor with levels
`V128` a factor with levels
`V129` a factor with levels
`V130` a factor with levels
`V131` a factor with levels
`V132` a factor with levels
`V133` a factor with levels
`V134` a factor with levels
`V135` a factor with levels
`V136` a factor with levels
`V137` a factor with levels
`V138` a factor with levels
`V139` a factor with levels
`V140` a factor with levels
`V141` a factor with levels
`V142` a factor with levels
`V143` a factor with levels
`V144` a factor with levels
`V145` a factor with levels
`V146` a factor with levels
`V147` a factor with levels
`V148` a factor with levels
`V149` a factor with levels
`V150` a factor with levels
`V151` a factor with levels
`V152` a factor with levels
`V153` a factor with levels
`V154` a factor with levels
`V155` a factor with levels
`V156` a factor with levels
`V157` a factor with levels
`V158` a factor with levels
`V159` a factor with levels

Details

The 39 observations are 39 functions or pathways listed each being constited with a set up-regulated genes found by NBBttest in knockdowned DDX39 cell line. Log10 of pvalue of gene in differential expression detection is used to weight conpoment of this gene playing role in this function. Then heatmap of pathways or functions are made using myheatmap. V1-V159 are genes of function or pathways

Examples

```
data(pathwy.A.up)
```

`pratio` *Calculation of psi (ψ)*

Description

ψ is also called polar ratio. It is used to measure overlap of two subdatasets. If $\psi > 1$, these two subdatasets have a gap, while $\psi < 1$ indicates that two subdatasets overlap.

Usage

```
pratio(xx, na, nb)
```

Arguments

`xx`	count data of RNA reads generated by next generation sequencing.
`na`	number of replicate libraries in condition A.
`nb`	number of replicate libraries in condition B.

Details

Psi (ψ) is defined as

$$\psi = max\left(\frac{min(X_A)}{max(X_B)}, \frac{min(X_B)}{max(X_A)}\right)$$

Value

`pratio`	pratio list

Author(s)

Yuan-De Tan <tanyuande@gmail.com>

References

Yuan-De Tan Anita M. Chandler, Arindam Chaudhury, and Joel R. Neilson(2015) A Powerful Statistical Approach for Large-scale Differential Transcription Analysis. *Plos One*. 2015 DOI: 10.1371/journal.pone.0123658.

See Also

`omega`, and `oddratio`

Examples

```
XX<-matrix(NA,2,8)
XX[1,]<-c(112,122, 108,127,302, 314, 322, 328)
XX[2,]<-c(511, 230, 754, 335,771, 842, 1014,798)
#XX
#     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
#[1,]  112  122  108  127  302  314  322  328
#[2,]  511  230  754  335  771  842 1014  798
pratio(xx=XX,na=4,nb=4)
```

---

prime3_PRP8_50 *3'UTR splicing data of 50 genes detected in the knockdowned PRP8 cell line*

Description

Splicing events occurring in 3'UTR of 50 genes were detected in the knockdowned PRP8 cell line.

Usage

```
data("prime3_PRP8_50")
```

Format

A data frame with 50 observations on the following 24 variables.

`chr` a factor with levels

`strand` a factor with levels

`isoform` a factor with levels

`Gene` a factor with levels

`exon_const_start` a numeric vector

`exon_const_end` a numeric vector

`exon_alt1_start` a numeric vector

`exon_alt1_end` a numeric vector

`exon_alt2_start` a numeric vector

`exon_alt2_end` a numeric vector

`Siha_Ctl_I_S3` a numeric vector

`Siha_Ctl_II_S7` a numeric vector

`Siha_Ctl_III_S11` a numeric vector

`PRP8_I_S2` a numeric vector

`PRP8_II_S6` a numeric vector

`PRP8_III_S10` a numeric vector

`tvalue` a numeric vector

`rho` a numeric vector

`pvalue` a numeric vector

`adjp` a numeric vector

w a numeric vector

X a numeric vector

X.1 a numeric vector

X.2 a numeric vector

Details

This dataset is an example for annotation of genes within which splicing events occurring in 3'UTR were detected by spladder-NBBttest in knockdowned PRP8 cell line. Gene information are assigned in Gene, isoform,Exon_const_start and exon_const_end that are exon constitution start and end positions and exon_alt1_start and end that are exon alternative start and end positions.Count data of reads are assigned to three replicate control columns Siha_Ctl_I_S3,II-S7,III_S11 and three replicate knokdowned PRP8 columns PRP8_I_S2,II_S6 and III_S10. The tvalue is t-statistic, rho is gene-wise variable, pvalue is p-value for t-statistic, w is threshold for rho, FDR (X1.1)is false discovery rate, selection (X.2)= 1 if pvalue < FDR, 0, otherwise.

Examples

```
data(prime3_PRP8_50)
```

QC *Count data quality check*

Description

QC is used to check data quality in two ways: scatter plot of two replicated samples and correlation heatmap of all samples.

Usage

```
QC(dat, nci, S1="NULL", S2="NULL", method="plot", colrs="greenred", rwcex=1.8,
clcex=1.8, x=10, tree="none", log="none", col="blue", pch=19, labsize=1.5, axis=1.5)
```

Arguments

dat	count matric dataset.
nci	number of columns containing data information suach as gene id, library id, target id, gene name, strand etc.
S1	numeric integer value, indicates which column in data matrix is specified in x-axis. S1 > nci.
S2	numeric integer value, indicate which column in data matrix is specified in y-axis. S2> nci.
method	string. Here two methods are given for choice: "plot" and "heatmap". Default is "plot".
colrs	colrs is a string. 8 color sets are given for choice in this program: "redgreen", "heat.colors", "redblue", "greenred", "bluered", "cm.colors", "terrain.colors", "topo.colors". The default color set is "redgreen".
rwcex	positive numbers, used as cex.axis for the row axis labeling. The default value is 1.8.

clcex	positive numbers, used as cex.axis for the column axis labeling. The default value is 1.8.
x	numeric argument: positive number, used as argument of cm.colors(x), terrain.colors(x) and topo.colors(x), the default value is 10.
tree	tree(s) drwan on row or column or both. User has four options: "both" for drawing trees on both row and column,"row" for drawing treeonly on row,"column" for drawing tree on only column, and "none" for no tree specified. Default is "none".
log	a string. Two options are given for choice: "none" and "log". log="log" indicates that data value would be transfomed with log2. The default is "none".
col	a string, used to specify scatter plot dot color.
pch	numeric value, used to specify dot type.
labsize	numeric value for size of xlabel and ylabel
axis	numeric value for axis scale.

Details

S1, S2 and nci must be given numeric integer values for plot and heatmap. However, when method is chosen to be "heatmap", then S1 and S2 are not specified. Columns of information should be left of matrix and count data should be after columns of information.

Value

not return values but create scatter plot or heatmap plot.

Note

requires gplots and grDevices.

Author(s)

Yuan-De Tan <tanyuande@gmail.com>

See Also

heatmap.2 and grDevices

Examples

```
require(gplots)
data(jkttcell)
QC(dat=jkttcell, nci=7, S1=8, S2=9, method = "plot", log = "log", col = "blue", pch = 19)
QC(dat=jkttcell, nci=7, S1=8, S2=9, method = "plot", log = "log", col = "blue", pch = 19)
QC(dat=jkttcell, nci=7, method = "heatmap", log = "log")
```

result *Jurkat T-cell Transcritomic Data with isoforms selected by MBttest*

Description

A data consist of 7 information columns and 6 data (numeric) columns and 1953 RNA isoforms selected by MBttest are used to make heatmaps.

Usage

```
data("result")
```

Format

A data frame with 1953 observations on the following 14 variables.

`tagid` : a numeric vector

`geneid` : a numeric vector

`name` : a string vector

`chr` : a string vector, a set of 24 chrosomomes

`strand` : a factor with levels `-` `+`

`pos` : a numeric vector

`anno` : string vector, poly(A) site types

`Jurk.NS.A` : a numeric vector

`Jurk.NS.B` : a numeric vector

`Jurk.NS.C` : a numeric vector

`Jurk.48h.A` : a numeric vector

`Jurk.48h.B` : a numeric vector

`Jurk.48h.C` : a numeric vector

`tvalue` : a numeric vector

Details

The original data are count data generated by next generation sequencing from Jurkat T-cells. This dataset is a short dataset contains 1953 isoforms that show differential expression between rest status and stimulation status. The T-cells were treated by resting and stimulating with *CD3/CD28* for 48 hours. The data have 7 columns for the information of *poly(A)* site: tagid, geneid, gene name, chromosome, strand,*poly(A)* site position, *poly(A)* site annotation and 6 columns for data: Jurk.NS.A, Jurk.NS.B, Jurk.NS.C, Jurk.48h.A, Jurk.48h.B, Jurk.48h.C. where NS means Normal state and 48h means 48 hours after *CD3/CD28* stimulatuin of T-cells. 13409 *RNA* isoforms were detected to have alternative *poly(A)* sites.

Value

datasheet containd ID, information, count data of RNA reads.

References

Yuan-De Tan Anita M. Chandler, Arindam Chaudhury, and Joel R. Neilson(2015) A Powerful Statistical Approach for Large-scale Differential Transcription Analysis.*Plos One*. DOI: 10.1371/journal.pone.0123658.

Examples

```
data(result)
## maybe str(result) ; plot(result) ...
```

---

sgRNA *sgRNA dataset*

Description

This dataset was created by simulating single guide RNAs to edit genes.

Usage

```
data("sgRNA")
```

Format

A data frame with 1000 observations on the following 11 variables.

`gene` a factor with levels

`sgRNA` a factor with levels

`class` a factor with levels

`L1` a numeric vector

`L2` a numeric vector

`L3` a numeric vector

`L4` a numeric vector

`H1` a numeric vector

`H2` a numeric vector

`H3` a numeric vector

`H4` a numeric vector

Details

The dataset is CRISPR screening data. Each gene was edited by10 sgRNAs that contains a targeting sequence (crRNA sequence) and a Cas9 nuclease-recruiting sequence (tracrRNA). L1-L4 samples were lowly targeted by sgRNAs and H1-H4 were highly targeted by sgRNAs.

Examples

```
data(sgRNA)
```

simSplicing *Simulated alternative splicing*

Description

This alternative splicing count data were created by simulating dorsal and ventral RNA-sequence count data.

Usage

```
data("simSplicing")
```

Format

A data frame with 5000 observations on the following 9 variables.

`Isoform` a factor with levels

`geneid` a factor with levels

`label` a numeric vector

`D1` a numeric vector

`D2` a numeric vector

`D3` a numeric vector

`V1` a numeric vector

`V2` a numeric vector

`V3` a numeric vector

Details

D1-D3 are samples from mouse dorsal tissue and V1-V3 are samples from mouse ventral tissue. This dataset was simulated with negative binomial ditribution with 10 percent of isoforms of being differentially spliced and differential effect of 500U where U is uniform variable.

Examples

```
data(simSplicing)
```

simulat *Simulation*

Description

This function uses negative binomial (NB) pseudorandom generator to create count datasets of RNA isoform reads based on real data.

Usage

```
simulat(yy, nci, r1, r2, p, q, A)
```

Arguments

yy	real count data
nci	numeric argument: column number of information related to genes or isoforms.
r1	numeric argument: number of replicate libraries in condition 1.
r2	numeric argument: number of replicate libraries in condition 2.
p	numeric argument: proportion of genes or isoforms differentially expressed. The value is in range of 0~1. Default value is 0.
q	numeric argument: proportion of genes or isoforms artificially noised. The value is in range of 0~1. Default value is 0.
A	numeric argument: conditional effect value. The value is larger than or equal to 0. Default value is 0.

Details

Null count data are created by using R negative binomial pseudorandom generator rnbinom with mu and size. Parameters mu and size are given by mean and variance drawn from real read counts of a gene set or an isoform set in a condition. Condition (or treatment) effect on differential transcription of isoforms is linearly and randomly assigned to genes or isoforms. The conditional effect = AU where U is uniform variable and A is input constant. P percent of genes or isoforms is set to be differentially expressed or differentially spliced. Q percent of genes or isoforms has technical noise. If P = 0, then simulation is null simulation, the data are null data or baseline data.

Value

Return count data.

Author(s)

Yuan-De Tan <tanyuande@gmail.com>

References

Yuan-De Tan Anita M. Chandler, Arindam Chaudhury, and Joel R. Neilson(2015) A Powerful Statistical Approach for Large-scale Differential Transcription Analysis.*Plos One*, 10.1371/journal.pone.0123658.

See Also

NegBinomial

Examples

```
data(jkttcell)
jknull<-simulat(yy=jkttcell[1:500,],nci=7,r1=3,
r2=3,p=0,q=0.2,A=0)
```

skjt *Simulated Null Transcriptomic data*

Description

The dataset skjt generated by using R negative binomial pseudorandom generator rnbinom is used as an example for calculating omega.

Usage

```
data("skjt")
```

Format

A data frame with 13409 observations on the following 14 variables.

`geneid` : a string vector

`tagid` : a numeric vector

`geneid.1` : a numeric vector

`name` : a string vector

`chr` : a string vector

`strand` : a character vector

`pos` : a numeric vector

`anno` : a string vector

`Jurk.NS.A` : a numeric vector

`Jurk.NS.B` : a numeric vector

`Jurk.NS.C` : a numeric vector

`Jurk.48h.A` : a numeric vector

`Jurk.48h.B` : a numeric vector

`Jurk.48h.C` : a numeric vector

Details

The dataset skjt was generated by using R negative binomial pseudorandom generator rnbinom with mu and size. Parameters mu and size are given by mean and variance drawn from real Jurkat T cell transcriptomic count data . Condition (or treatment) effect on differential transcription of isoforms was set to zero. The data have 13409 genes and 7 information columns: geneid tagid name `chr`,`strand`,`pos`,`anno`, and 6 data columns.

Value

datasheet containd ID, information, count data of RNA reads.

Source

Simulation.

References

Yuan-De Tan Anita M. Chandler, Arindam Chaudhury, and Joel R. Neilson(2015) A Powerful Statistical Approach for Large-scale Differential Transcription Analysis. *Plos One*. DOI: 10.1371/journal.pone.0123658.

Examples

```
data(skjt)
## maybe str(skjt) ; plot(skjt) ...
```

`smbetattest` *Performance of multiple Beta t-test on simulated data*

Description

This function is to peform mBeta t-test with rho=1 and omega=1 on simulated data. The result lists differentially expressed genes or isoforms and their rho values. The rho values are used to calculate omega value for performance of mBeta t-tests on the real data.

Usage

```
smbetattest(X, na, nb, alpha)
```

Arguments

`X`	simulated count data with N genes or isoforms.
`na`	number of replicate libraries in condition A.
`nb`	number of replicate libraries in condition B.
`alpha`	statistical probabilistic threshold, default value is 0.05.

Details

Before performing mbeta t-test on real data, user needs omega value for the threshold of rho. To determine omega value, user is requred to simulate null data having the same gene or isoform number and the same numbers of replicate libraries in two conditions and then performs mbeta t-test on the simulated null data by setting rho =1 and omega =1. To calculate accurately omega value, user needs such performance on 4-6 simulated null datasets. Manual provides method for omega calculation.

Value

Return results from multple beta t-tests on simulated data.

Author(s)

Yuan-De Tan <tanyuande@gmail.com>

References

Yuan-De Tan Anita M. Chandler, Arindam Chaudhury, and Joel R. Neilson(2015) A Powerful Statistical Approach for Large-scale Differential Transcription Analysis. *Plos One*. DOI: 10.1371/journal.pone.0123658.

See Also

See Also as mbetattest

Examples

```
data(skjt)
nrho<-smbetattest(X=skjt[1:1000,],na=3,nb=3,alpha=0.05)
```

---

subdata *Splite data into two subsets*

Description

Data are splited into two subsets: gene single-isoform data and gene multi-isoform data.

Usage

```
subdata(xx, sg)
```

Arguments

xx	datasheet with gene single- and multi-isoform data.
sg	integer numeric

Details

sg=1 for gene single-isoform data, sg=2 for gene multi-isoform data.

Value

return dataset with sg=1 or sg=2

Author(s)

Yuan-De Tan <tanyuande@gmail.com>

Examples

```
data(jkttcell)
colnames(jkttcell)[3]<-"Gene"
jk.sg<-subdata(xx=jkttcell, sg=1)
jk.mg<-subdata(xx=jkttcell, sg=2)
```

upGAm *Count data of group A treated breast cancer in mice*

Description

Data of RNA-seq reads were obtained from drug treated breast cancer in mice and mapped by using STAR onto mm10 to create count data. Dataset upGAm is output of NBBttest.

Usage

```
data("upGAm")
```

Format

A data frame with 263 observations on the following 19 variables.

`gene` a factor with 263 DE genes

`A.2_S35S` a numeric vector

`A.4_S36S` a numeric vector

`A.42_S45S` a numeric vector

`A.39_S41S` a numeric vector

`A.9_S44R` a numeric vector

`A.12_S40R` a numeric vector

`A.18_S37R` a numeric vector

`A.29_S39R` a numeric vector

`A.31_S38R` a numeric vector

`A.38_S34R` a numeric vector

`tvalue` a numeric vector

`rho` a numeric vector

`pvalue` a numeric vector

`adjp` a numeric vector

`w` a numeric vector

`order` a numeric vector

`FDR` a numeric vector

`significance` a numeric vector

Details

This is dataset with 263 DE genes detected by NBBttest. The dataset consists of two tomur cell groups: sensitive to drug(S) and resistant to drug (R). 4 sensitive cells and 6 resistant cells are available for differential analysis. Tvalue is t-statistic, rho is gene-wise variable, pvalue is p-value for t-statistic, w is threshold for rho, FDR is false discovery rate, significance = 1 if pvalue < FDR, 0, otherwise.

Examples

```
data(upGAm)
```

Index