Package 'dcGSA'

October 12, 2016

Type Package

Title Distance-correlation based Gene Set Analysis for longitudinal gene expression profiles

Version 1.0.1

Date 2015-11-09

Depends R (>= 3.3), Matrix

Imports BiocParallel

Suggests knitr

VignetteBuilder knitr

Author Jiehuan Sun [aut, cre], Jose Herazo-Maya [aut], Xiu Huang [aut], Naftali Kaminski [aut], and Hongyu Zhao [aut]

Maintainer Jiehuan sun <jiehuan.sun@yale.edu>

Description Distance-correlation based Gene Set Analysis for longitudinal gene expression profiles. In longitudinal studies, the gene expression profiles were collected at each visit from each subject and hence there are multiple measurements of the gene expression profiles for each subject. The dcGSA package could be used to assess the associations between gene sets and clinical outcomes of interest by fully taking advantage of the longitudinal nature of both the gene expression profiles and clinical outcomes.

License GPL-2

LazyData TRUE

RoxygenNote 5.0.1

biocViews GeneSetEnrichment,Microarray, StatisticalMethod, Sequencing, RNASeq, GeneExpression

NeedsCompilation no

R topics documented:

---

dcGSA *Perform gene set analysis for longitudinal gene expression profiles.*

Description

Perform gene set analysis for longitudinal gene expression profiles.

Usage

```
dcGSA(data = NULL, geneset = NULL, nperm = 10, c = 0,
  parallel = FALSE, BPparam = MulticoreParam(workers = 4))
```

Arguments

data	A list with ID (a character vector for subject ID), pheno (a data frame with each column being one clinical outcome), gene (a data frame with each column being one gene).
geneset	A list of gene sets of interests (the output of readGMT function).
nperm	An integer number of permutations performed to get P values.
c	An integer cutoff value for the overlapping number of genes between the data and the gene set.
parallel	A logical value indicating if parallel computing is wanted.
BPparam	Parameters to configure parallel evaluation environments if parallel is TRUE. The default value is to use 4 cores in a single machine. See BiocParallelParam object in Bioconductor package BiocParallel for more details.

Value

returns a data frame with following columns.

Geneset	Names for the gene sets.
TotalSize	The original size of each gene set.
OverlapSize	The overlapping number of genes between the data and the gene set.
Stats	Longitudinal distance covariance between the clinical outcomes and the gene set.
NormScore	Only available when permutation is performed. Normalized longitudinal distance covariance using the mean and standard deviation of permutated values.
P	Only available when permutation is performed. Permutation P values.

References

Distance-correlation based Gene Set Analysis in Longitudinal Studies. Jiehuan Sun, Jose Herazo-Maya, Xiu Huang, Naftali Kaminski, and Hongyu Zhao.

Examples

```
data(dcGSAtest)
fpath <- system.file("extdata", "sample.gmt.txt", package="dcGSA")
GS <- readGMT(file=fpath)
system.time(res <- dcGSA(data=dcGSAtest,geneset=GS,nperm=100))
head(res)
```

---

dcGSAtest *dcGSA test data*

Description

A R data object of example data to test dcGSA. This is a list comprised of ID, data (phenotypes of interest), gene (longitudinal gene expresion profiles).

Examples

```
data(dcGSAtest) # load the test dataset
```

LDcov *Calculate longitudinal distance covariance statistics.*

Description

Calculate longitudinal distance covariance statistics.

Usage

```
LDcov(x.dist = NULL, y.dist = NULL, nums = NULL, bmat = NULL)
```

Arguments

x.dist	A block-diagonal distance matrix of each block being pairwise distance matrix of genes for each subject.
y.dist	A block-diagonal distance matrix of each block being pairwise distance matrix of clinical outcomes for each subject.
nums	A vector of integer numbers indicating the number of repeated measures for each subject.
bmat	A numerical matrix with one column for each subject (binary values indicating the locations of the repeated measures for that subject).

Value

returns the longitudinal distance covariance statistics.

Examples

```
## Not run: require(Matrix)
x <- cbind(rnorm(7),rnorm(7)) ## two genes
y <- cbind(rnorm(7),rnorm(7)) ## two clinical outcomes
## Two subjects: the first one has three measures
## while the other one has four measures
ID <- c(1,1,1,2,2,2,2) ## The IDs for the two subjects.
nums <- c(3,4) ## number of repeated measures for each subjects
## prepare block-diagonal distance matrix for genes and clinical outcomes
lmat <- lapply(nums,function(x){z=matrix(1,nrow=x,ncol=x)})
mat <- as.matrix(bdiag(lmat))
lmat <- lapply(nums,function(x){z=matrix(0,nrow=x,ncol=x);z[,1]=1;z})
bmat <- as.matrix(bdiag(lmat))
ind <- apply(bmat,2,sum)
bmat <- bmat[,ind!=0]
ydist <- as.matrix(dist(y))*mat
xdist <- as.matrix(dist(x))*mat
LDcov(x.dist=xdist,y.dist=ydist,nums=nums,bmat)
```

---

readGMT *Read gene set file in gmt format*

Description

Read gene set file in gmt format

Usage

```
readGMT(file = NULL)
```

Arguments

file	filename for the gmt file

Value

a list of gene sets with each element being a vector of gene names

Examples

```
fpath <- system.file("extdata", "sample.gmt.txt", package="dcGSA")
GS <- readGMT(file=fpath)
```

Index