Package 'discordant'

October 17, 2017

Version 1.0.0

Date 2016-10-21

Title The Discordant Method: A Novel Approach for Differential Correlation

Author Charlotte Siska [cre,aut], Katerina Kechris [aut]

Depends R (>= 3.4)

Maintainer Charlotte Siska <siska.charlotte@gmail.com>

Description Discordant is a method to determine differential correlation of molecular feature pairs from -omics data using mixture models. Algorithm is explained further in Siska et al.

Encoding latin1

biocViews BiologicalQuestion, StatisticalMethod, mRNAMicroarray, Microarray, Genetics, RNASeq

Suggests BiocStyle, knitr

Imports Biobase, stats, biwt, gtools, MASS, tools

License GPL (>= 2)

URL https://github.com/siskac/discordant

NeedsCompilation yes

VignetteBuilder knitr

R topics documented:

createVectors *Create Pearson's correlation coefficient vectors based on bivariate data*

Description

Calculates correlation coefficients based on two groups of omics bivariate data. Currently, only two groups of samples can be specified. Used to make input for discordantRun().

Usage

```
createVectors(x, y = NULL, groups, cor.method = c("spearman"))
```

Arguments

x	ExpressionSet of -omics data
y	optional second ExpressionSet of -omics data, induces dual -omics analysis
groups	n-length vector of 1s and 2s matching samples belonging to groups 1 and 2
cor.method	correlation method to measure association. Options are "spearman", "pearson", "bwmc" and "sparcc"

Details

Creates vectors of correlation coefficents based on feature pairs within x or between x and y. The names of the vectors are the feature pairs taken from x and y.

Value

v1	List of correlation coefficients for group 1
v2	List of correlation coefficients for group 2

Author(s)

Charlotte Siska <siska.charlotte@gmail.com>

References

Siska C, Bowler R and Kechris K. The Discordant Method: A Novel Approach for Differential Correlation. (2015) Bioinformatics. 32(5): 690-696. Friedman J and Alm EJ. Inferring Correlation Networks from Genomic Survey Data. (2012) PLoS Computational Biology. 8:9, e1002687.

Examples

```
## load data
data("TCGA_GBM_miRNA_microarray") # loads matrix called TCGA_GBM_miRNA_microarray
data("TCGA_GBM_transcript_microarray") # loads matrix called TCGA_GBM_transcript_microarray
print(colnames(TCGA_GBM_transcript_microarray)) # look at groups

groups <- c(rep(1,10), rep(2,20))

# transcript-transcript pairs
```

```
vectors <- createVectors(TCGA_GBM_transcript_microarray, groups = groups, cor.method = c("pearson"))

# miRNA-transcript pairs

vectors <- createVectors(TCGA_GBM_transcript_microarray, TCGA_GBM_miRNA_microarray, groups = groups)
```

---

discordantRun *Run Discordant Algorithm*

Description

Runs discordant algorithm on two vectors of correlation coefficients.

Usage

```
discordantRun(v1, v2, x, y = NULL, transform = TRUE, subsampling = FALSE, subSize = dim(x)[1], iter
```

Arguments

v1	Vector of Pearson correlation coefficients in group 1
v2	Vector of Pearson correlation coefficients in group 2
x	ExpressionSet of -omics data
y	ExpressionSet of -omics data, induces dual -omics analysis
transform	If TRUE v1 and v2 will be Fisher transformed
subsampling	If TRUE subsampling will be run
subSize	Indicates how many feature pairs to be used for subsampling. Default is the feature size in x
iter	Number of iterations for subsampling. Default is 100
components	Number of components in mixture model.

Details

The discordant algorithm is based on a Gaussian mixture model. If there are three components, correlation coefficients are clustered into negative correlations (-), positive correlations (+) and no correlation (0). If there are five components, then there are two more classes for very negative correlation (–) and very positive correlations (++). All possible combinations for these components are made into classes. If there are three components, there are 9 classes. If there are five components, there are 25 classes.

The posterior probabilities for each class are generated and outputted into the value probMatrix. The value probMatrix is a matrix where each column is a class and each row is a feature pair. The values discordPPVector and discordPPMatrix are the summed differential correlation posterior probability for each feature pair. The values classVector and classMatrix are the class with the highest posterior probability for each feature pair.

Value

`discordPPVector`
: Vector of differentially correlated posterior probabilities.

`discordPPMatrix`
: Matrix of differentially correlated posterior probabilities where rows and columns reflect features

`classVector`
: Vector of classes that have the highest posterior probability

`classMatrix`
: Matrix of classes that have hte highest posterior probability where rows and columns reflect features

`probMatrix`
: Matrix of posterior probabilities where rows are each molecular feature pair and columns are nine different classes

`loglik`
: Final log likelihood

Author(s)

Charlotte Siska <siska.charlotte@gmail.com>

References

Siska C, Bowler R and Kechris K. The Discordant Method: A Novel Approach for Differential Correlation (2015), Bioinformatics. 32 (5): 690-696. Lai Y, Zhang F, Nayak TK, Modarres R, Lee NH and McCaffrey TA. Concordant integrative gene set enrichment analysis of multiple large-scale two-sample expression data sets. (2014) BMC Genomics 15, S6. Lai Y, Adam B-l, Podolsky R, She J-X. A mixture model approach to the tests of concordance and discordancd between two large-scale experiments with two sample groups. (2007) Bioinformatics 23, 1243-1250.

Examples

```
## load Data

data(TCGA_GBM_miRNA_microarray) # loads matrix called TCGA_GBM_miRNA_microarray
data(TCGA_GBM_transcript_microarray) # loads matrix called TCGA_GBM_transcript_microarray
print(colnames(TCGA_GBM_transcript_microarray)) # look at groups
groups <- c(rep(1,10), rep(2,20))

## DC analysis on only transcripts pairs

vectors <- createVectors(TCGA_GBM_transcript_microarray, groups = groups)
result <- discordantRun(vectors$v1, vectors$v2, TCGA_GBM_transcript_microarray)

## DC analysis on miRNA-transcript pairs

vectors <- createVectors(TCGA_GBM_transcript_microarray, TCGA_GBM_miRNA_microarray, groups = groups, cor.met
result <- discordantRun(vectors$v1, vectors$v2, TCGA_GBM_transcript_microarray, TCGA_GBM_miRNA_microarray)
```

fishersTrans *Fisher Transformation of Pearson Correlation Coefficients to Z Scores*

Description

Transforms Pearsons correlation coefficients into z scores using Fishers method.

Usage

```
fishersTrans(rho)
```

Arguments

rho	Integer or numeric vector of Pearson's correlation coefficients

Details

Fisher's transformation is when correlation coefficients are transformed into a z score. These z scores have an approximately normal distribution.

Value

Returns Fisher-transformed correlation coefficients

References

Fisher, R.A. (1915). "Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population". Biometrika (Biometrika Trust) 10 (4).

Examples

```
## Create integer or list of Pearson's correlation coefficients.

library(MASS)
rhoV <- as.vector(cor(t(mvrnorm(10,rep(3,100),diag(100)))))

## Determine Fisher-Transformed z scores of rho
zV <- fishersTrans(rhoV)
```

splitMADOutlier *Outliers using left and right MAD*

Description

Identify features with outliers using left and right median absolute deviation (MAD).

Usage

```
splitMADOutlier(mat, filter0 = TRUE, threshold = 2)
```

Arguments

mat	mxn matrix of -omics data, where rows are features and columns samples.
filter0	Option to filter out features if they have at least one 0 value. Default is TRUE.
threshold	Threshold of how many MADs outside the left or right median is used to determine features with outliers.

Details

The purpose of this function is to determine outliers in non-symmetric distributions. The distribution is split by the median. Outliers are identifed by being however many median absolute deviations (MAD) from either split distribution.

Value

mat.filtered	Input matrix where features with outliers filtered out.
index	Index of features that have no outliers.

References

Leys C, Klein O, Bernard P and Licata L. "Detecting Outliers: Do Not Use Standard Deviation Around the Mean, Use Absolute Deivation Around the Median." Journal of Experimental Social Psychology, 2013. 49(4), 764-766. Magwene, PM, Willis JH, Kelly JK and Siepel A. "The Statistics of Bulk Segregant Analysis Using Next Generation Sequencing." PLoS Computational Biology, 2011. 7(11), e1002255.

Examples

```
## Simulate matrix of continuous -omics data.
data(TCGA_Breast_miRNASeq)

## Filter matrix based on outliers.
mat.filtered <- splitMADOutlier(TCGA_Breast_miRNASeq)$mat.filtered
```

TCGA_Breast_miRNASeq *TCGA Breast Cancer miRNASeq Sample Dataset*

Description

This dataset contains TMM normalized miRNA count values from miRNASeq that was taken from the Cancer Genome Atlas, or TCGA. The dataset has 100 miRNA and 57 samples. The original dataset has 212 miRNA and 57 samples.

Usage

```
TCGA_Breast_miRNASeq
```

Format

A matrix of miRNA count values

Value

Breast miRNA-Seq count data with 100 features and 57 samples.

Author(s)

Charlotte Siska <siska.charlotte@gmail.com>

References

National Institues of Health. The Cancer Genome Atlas. http://cancergenome.nih.gov/

`TCGA_Breast_miRNASeq_voom`

TCGA Breast Cancer miRNASeq Sample Dataset

Description

This dataset contains TMM normalized voom-transformed miRNA count values from miRNASeq that was taken from the Cancer Genome Atlas, or TCGA. The dataset has 100 miRNA and 57 samples. The original dataset has 212 miRNA and 57 samples.

Usage

```
TCGA_Breast_miRNASeq_voom
```

Format

A matrix of miRNA count values

Value

Breast miRNA-Seq voom-transformed count data with 100 features and 57 samples.

Author(s)

Charlotte Siska <siska.charlotte@gmail.com>

References

Charity W Law, Yunshun Chen, Wei Shi, Gordon K Smyth. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. 2014. Genome Biology, 15:R29. National Institues of Health. The Cancer Genome Atlas. http://cancergenome.nih.gov/

TCGA_Breast_RNASeq *TCGA Breast Cancer RNASeq Sample Dataset*

Description

This dataset contains TMM normalized RNA count values from RNASeq that was taken from the Cancer Genome Atlas, or TCGA. It has 100 features and 57 samples. The original dataset had 17972 features and 57 samples.

Usage

```
TCGA_Breast_RNASeq
```

Format

A matrix of RNA count values

Value

Breast RNA-Seq count data with 100 features and 57 samples.

Author(s)

Charlotte Siska <siska.charlotte@gmail.com>

References

National Institues of Health. The Cancer Genome Atlas. http://cancergenome.nih.gov/

TCGA_Breast_RNASeq_voom *TCGA Breast Cancer RNASeq Sample Dataset*

Description

This dataset contains TMM normalized voom-transformed RNA count values from RNASeq that was taken from the Cancer Genome Atlas, or TCGA.

Usage

```
TCGA_Breast_RNASeq_voom
```

Format

A matrix of RNA count values

Value

Breast RNA-Seq voom-transformed count data with 100 features and 57 samples.

Author(s)

Charlotte Siska <siska.charlotte@gmail.com>

References

Charity W Law, Yunshun Chen, Wei Shi, Gordon K Smyth. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. 2014. Genome Biology, 15:R29. National Institues of Health. The Cancer Genome Atlas. http://cancergenome.nih.gov/

`TCGA_GBM_miRNA_microarray`

TCGA Glioblastoma Multiforme miRNA Sample Dataset

Description

This dataset contains miRNA expression values from a microarray that was taken from the Cancer Genome Atlas, or TCGA. It has 10 features and 30 samples. The original dataset had 331 features and 30 samples.

Usage

```
TCGA_GBM_miRNASample
```

Format

A matrix of miRNA expression values

Value

GBM miRNA microarray data with 10 features and 30 samples.

Author(s)

Charlotte Siska <siska.charlotte@gmail.com>

References

National Institues of Health. The Cancer Genome Atlas. http://cancergenome.nih.gov/

`TCGA_GBM_transcript_microarray` *TCGA Glioblastoma Multiforme Transcript Sample Dataset*

Description

This dataset contains transcript expression values from a microarray that was taken from the Cancer Genome Atlas, or TCGA. It has 10 features and 30 samples. The original dataset had 72656 features and 30 samples.

Usage

```
TCGA_GBM_transcript_microarray
```

Format

A matrix of transcript expression values

Value

GBM transcript microarray data with 10 features and 30 samples.

Author(s)

Charlotte Siska <siska.charlotte@gmail.com>

References

National Institues of Health. The Cancer Genome Atlas. http://cancergenome.nih.gov/

Index