Package 'TIN'

March 17, 2015

Type Package

Title Transcriptome instability analysis

Version 0.99.4

Date 2014-07-14

Author Bjarne Johannessen, Anita Sveen and Rolf I. Skotheim

Maintainer Bjarne Johannessen <bjajoh@rr-research.no>

VignetteBuilder knitr

Depends R (>= 2.12.0), data.table, impute, aroma.affymetrix

Imports WGCNA, squash, stringr

Suggests knitr, aroma.light, affxparser, RUnit, BiocGenerics

biocViews ExonArray, Microarray, GeneExpression, AlternativeSplicing, Genetics, DifferentialSplicing

Description The TIN package implements a set of tools for transcriptome instability analysis based on exon expression profiles. Deviating exon usage is studied in the context of splicing factors to analyse to what degree transcriptome instability is correlated to splicing factor expression. In the transcriptome instability correlation analysis, the data is compared to both random permutations of alternative splicing scores and expression of random gene sets.

License Artistic-2.0

R topics documented:

`aberrantExonUsage` *Calculations of relative aberrant exon usage amounts per sample (based on FIRMA scores)*

Description

The function takes in the data.frame from 'firmaAnalysis' (containing log2 FIRMA scores for all probe sets/exons (rows) in all samples (columns)), and a number indicating which percentile value of global FIRMA scores to be used as threshold for denoting aberrant exon usage (default value '1', calculating the lower and upper 1st percentiles, indicating aberrant exon skipping and inclusion, respectively). Lower and upper percentile values are calculated and stored in the global list object 'quantiles'. Also, the total number of exons per sample denoted with aberrant exon usage (having FIRMA scores outside the indicated threshold values) is calculated and stored in the global list object 'aberrantExons'. The function returns a vector with these total sample-wise amounts of aberrant exon usage (sum of aberrant skipping and inclusion amounts) relative to the average sample-wise amount in the dataset (log2-transformed).

Usage

```
aberrantExonUsage(percentile, fs)
```

Arguments

`percentile`	This number indicates the percentile value of the global FIRMA scores to be used as threshold for denoting aberrant exon usage. Default value '1' calculates the lower and upper 1st percentiles, indicating aberrant exon skipping and inclusion, respectively.
`fs`	Data.frame consisting of log2 FIRMA scores for all probe sets/exons (rows) in all samples (columns). This data.frame is the output from 'firmaAnalysis'.

Value

A numeric vector with log2-transformed sample-wise amounts of aberrant exon usage relative to the average sample-wise amount in the dataset. In addition, the quantiles list object is created, which contains the threshold values for the lower and upper percentiles.

Examples

```
# Calculate aberrant exon usage for each sample in the data set:
    fs <- firmaAnalysis(useToyData=TRUE)
    tra <- aberrantExonUsage(1.0, fs)
#   The aberrantExonUsage function also creates the 'quantiles' object with
#   upper and lower threshold values for accepting aberrant exon usage, and
#   the list object 'aberrantExons' with the sample-wise number of exons
#   outside the threshold values.
```

clusterPlot *clusterPlot*

Description

Create plot from hierarchical clustering analysis of the samples, based on splicing factor expression levels.

Usage

```
clusterPlot(geneSummaries, tra, distmethod, clustermethod, fileName)
```

Arguments

geneSummaries	The data.frame with gene-level expression values for each sample, returned from the function 'readGeneSummaries'.
tra	The list returned from the function 'aberrantExonUsage', containing sample-wise total relative amounts of aberrant exon usage.
distmethod	Which distance measure to be used. Possible options are "euclidean", "maximum", "manhattan", "canberra", "binary" or "minkowski".
clustermethod	Which clustering algorithm to be used. Possible options are "ward", "single", "complete", "average", "mcquitty", "median" or "centroid".
fileName	Output filename. File format is optional, but must be one of png, jpg, eps or pdf.

Value

clusterPlot is used for the side-effect of producing a hierarchical clustering plot showing how the samples are separated based on expression levels for the splicing factors in each sample.

Examples

```
    fs <- firmaAnalysis(useToyData=TRUE)
    gs <- readGeneSummaries()
    tra <- aberrantExonUsage(1.0, fs)
# create cluster plot with the samples
    clusterPlot(gs, tra, "euclidean", "complete", "cluster.png")
```

---

correlation *Calculates the correlation between sample-wise amounts of aberrant exon usage and splicing factor expression levels*

Description

The function makes use of the corAndPValue function from the WGCNA package to calculate sample-wise Pearson correlation between relative amounts of aberrant exon usage and splicing factor expression levels.

Usage

```
correlation(splicingFactors, geneSummaries, tra)
```

Arguments

splicingFactors	A data.frame with a list of splicing factor genes (Affymetrix transcript cluster id's and gene symbols) to be included in the correlation analysis. The list can include any set of genes included in the data.frame returned from the function 'readGeneSummaries'. An example set of 280 genes is obtained by issuing the command 'data(splicingFactors)'.
geneSummaries	A data.frame with gene-level expression values for all samples, returned from the 'readGeneSummaries' function.
tra	List of sample-wise total relative amounts of aberrant exon usage, returned from the 'aberrantExonUsage' function.

Value

A list with sample-wise Pearson correlation values between relative amounts of aberrant exon usage and splicing factor expression levels.

Examples

```
    data(sampleSetGeneSummaries)
    gs <- sampleSetGeneSummaries
    fs <- firmaAnalysis(useToyData=TRUE)
    tra <- aberrantExonUsage(1.0, fs)
# calculate correlation between splicing factor expression and aberrant
# exon usage
    data(splicingFactors)
    corr <- correlation(splicingFactors, gs, tra)
```

---

correlationPlot *correlationPlot*

Description

Function for creating a plot that visualizes the number of splicing factor genes with expression levels significantly correlated with the sample-wise total relative amounts of aberrant exon usage (red). To compare this value with correlation values obtained from random sample permutations, the function performs two types of iterative sample calculations as control experiments. First, expression levels of the splicing factor gene set are correlated with permutations of the relative aberrant exon usage (dark blue). Second, expression levels of randomly generated gene sets are correlated with the relative aberrant exon usage amounts in the data (pale blue).

Usage

```
correlationPlot(fileName, tra, geneSummaries, splicingFactors,
randomGeneSets, traPermutations)
```

Arguments

fileName	Output filename. File format is optional, but must be one of png, jpg, eps or pdf.
tra	List of sample-wise total relative amounts of aberrant exon usage, obtained using the 'aberrantExonUsage' function.
geneSummaries	Data.frame with gene-level expression data for each sample, returned from the function 'readGeneSummaries'.
splicingFactors	List with Affymetrix transcript cluster id's and gene symbols for a set of genes involved in pre-mRNA splicing. An example set with 280 genes is obtained by issuing the command 'data(splicingFactors)', but the list may well include any set of genes included in the data.frame returned from the function 'readGeneSummaries'.
randomGeneSets	Number of random gene sets of 280 genes to be created and included in the analysis.
traPermutations	Number of permutations of the sample-wise amounts of aberrant exon usage to be performed and included in the analysis.

Value

correlationPlot is used for the side-effect of producing a plot showing the number of splicing factor genes with expression levels significantly correlated with the sample-wise total relative amounts of aberrant exon usage.

Examples

```
    data(splicingFactors)
    fs <- firmaAnalysis(useToyData=TRUE)
    gs <- readGeneSummaries()
    tra <- aberrantExonUsage(1.0, fs)

# Create a plot that visualizes the number of splicing factor
# genes with expression levels significantly correlated with the
# sample-wise total relative amounts of aberrant exon usage.
    correlationPlot("c.png", tra, gs, splicingFactors, 1000, 1000)
```

---

firmaAnalysis *Read CEL files and perform FIRMA analysis*

Description

The function makes use of the aroma.affymetrix package to analyze Affymetrix Human Exon 1.0 ST Arrays. The function reads CEL files, and performs background correction, normalization (customized RMA approach), and alternative splicing analysis according to the FIRMA method (http://www.aroma-project.org/vignettes/FIRMA-HumanExonArrayAnalysis). The function returns a data.frame with log2 FIRMA (alternative splicing) scores for each probeset/sample combination.

Usage

```
firmaAnalysis(useToyData, aromaPath, dataSetName)
```

Arguments

useToyData	Boolean argument to indicate whether sample data sets included in the TIN package should be used in the analysis.
aromaPath	Absolute or relative path to the aroma.affymetrix directory. Requires custom CDF annotation file (please refer to the FIRMA vignette for download and setup).
dataSetName	Name of folder in the 'aromaPath' containing raw data (CEL files; please refer to the FIRMA vignette for setup).

Value

A data.frame with expression level values after the FIRMA analysis has been applied. The data.frame consists of one column for each sample and one row for each probeset.

References

E. Purdom, K. Simpson, M. Robinson, J. Conboy, A. Lapuk & T.P. Speed, FIRMA: a method for detection of alternative splicing from exon array data. Bioinformatics, 2008.

Examples

```
# Perform FIRMA analysis on the raw expression data
# To use sample data sets included in the TIN package as input:
    fs <- firmaAnalysis(useToyData=TRUE)

# To use your own data, provide path to aroma.affymetrix root directory and
# name of data set as arguments:
#   fs <- firmaAnalysis(useToyData=FALSE, "/tmp/path/to/aroma.affymetrix",
#   "sampleSet")
```

---

geneAnnotation *geneAnnotation*

Description

The object contains a data.frame with Affymetrix transcript cluster id and gene symbol for the 22,011 genes defined by the Affymetrix core gene set.

Usage

```
data(geneAnnotation)
```

Format

data.frame with two columns; Affymetrix transcript cluster id and gene symbol.

Value

A data.frame with Affymetrix transcript cluster id and gene symbol for 22,011 genes.

`geneSetCorrelation` *Correlation between aberrant exon usage and expression levels for a number of gene sets.*

Description

The function makes use of the corAndPValue function from the WGCNA package to calculate the Pearson correlation between sample-wise aberrant exon usage amounts and expression levels of all genes for all gene sets defined by the input parameter list geneSets.

Usage

```
geneSetCorrelation(geneSets, geneAnnotation, geneSummaries, tra,
    noGeneSets)
```

Arguments

`geneSets`	A data.frame with a number of gene lists to be included in the correlation analysis. An example data.frame of 1,454 lists of genes (specified by Affymetrix transcript cluster id's and gene symbols) is included in the package, and will be accessible by issuing the command 'data(genesets)'. The example data.frame contains a complete collection of all Gene Ontology gene sets included in the Molecular Signatures Database v3.1.
`geneAnnotation`	A data.frame with Affymetrix transcript cluster id's and gene symbols for all 22,011 genes included in the Affymetrix 'core' set.
`geneSummaries`	A data.frame with gene-level expression values for all samples, returned from the 'readGeneSummaries' function.
`tra`	List with total relative amounts of aberrant exon usage for all samples obtained using the 'aberrantExonUsage' function.
`noGeneSets`	Optional argument specifying how many gene sets to include in the analysis.

Value

A data.frame with one row for each data set used as input, and columns for name of set, number of genes, number of significant positively/negatively correlated genes in the set, and median correlation strength.

Examples

```
# Load data
    data(geneSets)
    data(geneAnnotation)
    fs <- firmaAnalysis(useToyData=TRUE)
    gs <- readGeneSummaries()
    tra <- aberrantExonUsage(1.0, fs)
# Calculate correlation in other gene sets
    crs <- geneSetCorrelation(geneSets, geneAnnotation, gs, tra, 50)
```

geneSets *geneSets*

Description

A data.frame with 1,454 lists of genes (specified by Affymetrix transcript cluster id's and gene symbols) to be included in the correlation analysis. The data.frame contains a complete collection of all Gene Ontology gene sets included in the Molecular Signatures Database v3.1.

Usage

```
data(geneSets)
```

Format

data.frame that contains 1,454 lists of genes, specified by Affymetrix transcript cluster id's and gene symbols.

Value

A data.frame with 1,454 Gene Ontology gene lists specified by Affymetrix transcript cluster id's and gene symbols.

posNegCorrPlot *posNegCorrPlot*

Description

The posNegCorrPlot is a scatterPlot that compares the amount of splicing factor genes (red) for which expression levels are significant positively (vertical axis) and negatively (horizontal axis) correlated with the total relative amounts of aberrant exon usage per sample. The plot can also include results from permutations of the sample-wise aberrant exon usage amounts (dark blue), and randomly constructed gene sets of 280 genes.

Usage

```
posNegCorrPlot(fileName, tra, geneSummaries, splicingFactors,
    randomGeneSets, traPermutations)
```

Arguments

`fileName`	Output filename. File format is optional, but must be one of png, jpg, eps or pdf.
`tra`	List object with total relative amounts of aberrant exon usage per sample. The object is returned by the function 'aberrantExonUsage'.
`geneSummaries`	The data.frame with gene-level expression values for each sample, returned from the function 'readGeneSummaries'.

`splicingFactors`
: List of genes (Affymetrix transcript cluster id's and gene symbols) involved in pre-mRNA splicing. An example set of 280 genes is obtained by issuing the command 'data(splicingFactors)', but the input list can include any set of genes included in the data.frame returned from the function 'readGeneSummaries'.

`randomGeneSets`
: Number of random gene sets of 280 genes to be created and included in the analysis.

`traPermutations`
: Number of permutations of the sample-wise amounts of aberrant exon usage to be performed and included in the analysis.

Value

posNegCorrPlot is used for the side-effect of producing a scatter plot that compares the amount of splicing factor genes (red) for which expression levels are significant positively (vertical axis) and negatively (horizontal axis) correlated with the total relative amounts of aberrant exon usage per sample. In addition, the plot can also include results from permutations of the sample-wise aberrant exon usage amounts (dark blue), and randomly constructed gene sets of 280 genes.

Examples

```
    data(splicingFactors)
    fs <- firmaAnalysis(useToyData=TRUE)
    gs <- readGeneSummaries()
    tra <- aberrantExonUsage(1.0, fs)

# Create plot that compares the amount of splicing factor genes for which
# expression levels are significant positively and negatively correlated
# with the total relative amounts of aberrant exon usage per sample
    posNegCorrPlot("cg.png", tra, gs, splicingFactors, 1000, 1000)
```

`probesetPermutations` *Permutations of the samples at each probeset*

Description

The function takes in the data.frame from 'firmaAnalysis' (containing log2 FIRMA scores for all probe sets/exons (rows) in all samples (columns)), along with the list 'percentiles' from 'aberrantExonUsage' (containing the lower and upper percentile values of FIRMA scores used as thresholds for denoting aberrant exon usage), and makes permutations of the FIRMA scores for each probe set/exon across all samples. Based on the permutation, random relative amounts of aberrant exon skipping and inclusion per sample is calculated and returned.

Usage

```
    probesetPermutations(fs, percentiles)
```

Arguments

fs	FIRMA scores for each probe set/sample combination (the data.frame returned from the function 'firmaAnalysis').
percentiles	List containing two FIRMA score percentile values, i.e., the lower and upper percentiles used as thresholds for denoting aberrant exon usage (the list object 'percentiles' returned from the function 'aberrantExonUsage').

Value

A list with two vectors, with number of exon skipping and inclusion events, respectively, for each sample after permutations of the expression levels at each probeset.

Examples

```
# Set up data set with FIRMA scores and calculate relative aberrant
# exon usage for each sample
    fs <- firmaAnalysis(useToyData=TRUE)
    tra <- aberrantExonUsage(1.0, fs)
# Make permutations of the expression data at each probeset
    perms <- probesetPermutations(fs, quantiles)
```

---

readGeneSummaries *Read gene-level expression summaries*

Description

The function reads pre-pocessed and summarized gene-level expression data from file, and builds a data.frame with the information. The data.frame will have the same structure as the input summary file, i.e., genes in rows and samples in columns.

Usage

```
readGeneSummaries(useToyData, summaryFile)
```

Arguments

useToyData	Boolean argument to indicate whether sample data sets included in the TIN package should be used in the analysis.
summaryFile	Tab separated file with expression values for each combination of gene (row) and sample (column). Gene expression summary files can be obtained by using for instance APT (Affymetrix Power Tools) or Expression Console. The file is expected to have one initial header line with sample names, and one initial column with Affymetrix transcript cluster id's as row names.

Value

A data.frame with gene summary values. It consists of one column for each sample, and one row for each gene.

Examples

```
# Read pre-processed gene summary values from file
# To use sample test data included in the TIN package as input:
    gs <- readGeneSummaries(useToyData=TRUE)
# To use your own data, provide path to txt file with expression values:
#   gs <- readGeneSummaries("/tmp/path/to/GeneLevelExpressionValues.txt")
```

---

sampleSetFirmaScores

sampleSetFirmaScores

Description

A data.frame containing preprocessed log2 expression values from 16 samples in 10,000 randomly selected Affymetrix probesets.

Usage

```
data(sampleSetFirmaScores)
```

Format

data.frame that contains preprocessed log2 expression values from 16 samples in 10,000 randomly selected Affymetrix probesets.

Value

A data.frame with preprocessed log2 expression values in 10,000 randomly selected Affymetrix probesets for a test data set of 16 samples.

---

sampleSetGeneSummaries

sampleSetGeneSummaries

Description

A data.frame containing gene expression summary values from 16 samples in 22,011 Affymetrix transcript clusters.

Usage

```
data(sampleSetGeneSummaries)
```

Format

A data.frame containing gene expression summary values from 16 samples in 22,011 Affymetrix transcript clusters.

Value

A data.frame with expression values at the gene level for a test data set of 16 samples.

`scatterPlot` *Scatterplot showing relative amounts of aberrant exon usage per sample*

Description

Scatterplot showing sample-wise relative amounts (blue dots) of aberrant exon inclusion (horizontal axis) and exon skipping (vertical axis). Random sample-wise amounts calculated from permuted FIRMA scores can also be included in the plot (yellow dots).

Usage

```
scatterPlot(fileName, permutations, percentileHits,
permPercentileHits)
```

Arguments

`fileName`	Output filename. File format is optional, but must be one of png, jpg, eps or pdf.
`permutations`	Boolean argument to indicate whether permuted data should be included. 'TRUE' adds permutation data to the plot.
`percentileHits`	List with two items containing sample-wise numbers of exons denoted with aberrant exon skipping and inclusion, i.e., having FIRMA scores in the indicated lower and upper percentiles, respectively. This list is returned from the function 'aberrantExonUsage'.
`permPercentileHits`	List with two items containing random sample-wise numbers of exons denoted with aberrant exon skipping and inclusion, i.e., having FIRMA scores in the indicated lower and upper percentiles, respecively (calculated from FIRMA score permutations). This list is returned from the function 'probesetPermutations'.

Value

scatterPlot is used for the side-effect of producing a scatter plot showing relative amounts of aberrant exon usage per sample.

Examples

```
    data(splicingFactors)
    fs <- firmaAnalysis(useToyData=TRUE)
    gs <- readGeneSummaries()
    tra <- aberrantExonUsage(1.0, fs)
#   The aberrantExonUsage function also creates the 'quantiles' object with
#   upper and lower threshold values for accepting aberrant exon usage, and
#   the list object 'aberrantExons' with the sample-wise number of exons
#   outside the threshold values.
    aberrantExonsPerms <- probesetPermutations(fs, quantiles)

# Create scatter plot with the samples
    scatterPlot("scatter.png", TRUE, aberrantExons, aberrantExonsPerms)
```

`splicingFactors` *A list of 280 splicing factor genes*

Description

The gene set is a comprehensive collection of 280 genes involved in pre-mRNA splicing events (Sveen et al., Genome Medicine, 2011, 3:32).

Usage

```
data(splicingFactors)
```

Format

data.frame with two columns; Affymetrix transcript cluster id and gene symbol.

Value

A data.frame with 280 genes involved in pre-mRNA splicing events.

Index