

Package ‘csdR’

April 12, 2022

Title Differential gene co-expression

Version 1.0.0

Date 09-09-2021

Description This package contains functionality to run differential gene co-expression across two different conditions.

The algorithm is inspired by Voigt et al. 2017 and finds Conserved, Specific and Differentiated genes (hence the name CSD).

This package include efficient and variance calculation by bootstrapping and Welford's algorithm.

Depends R (>= 4.1.0)

Imports WGCNA, glue, RhpcBLASctl, matrixStats, Rcpp

License GPL-3

Encoding UTF-8

NeedsCompilation yes

BugReports <https://github.com/AlmaasLab/csdR/issues>

URL <https://almaaslab.github.io/csdR>,

<https://github.com/AlmaasLab/csdR>

Roxygen list(markdown = TRUE)

RoxygenNote 7.1.1

LinkingTo Rcpp

biocViews DifferentialExpression, GraphAndNetwork, GeneExpression,
Network

Suggests rmarkdown, knitr, testthat (>= 3.0.0), BiocStyle, magrittr,
igraph, dplyr

Config/testthat/edition 3

VignetteBuilder knitr

git_url <https://git.bioconductor.org/packages/csdR>

git_branch RELEASE_3_14

git_last_commit 4d1e1ad

git_last_commit_date 2021-10-26

Date/Publication 2022-04-12

Author Jakob Peder Pettersen [aut, cre]
(<https://orcid.org/0000-0002-3485-1634>)

Maintainer Jakob Peder Pettersen <jakobpeder.pettersen@gmail.com>

R topics documented:

data-expression	2
partial_argsort	3
run_cor_bootstrap	3
run_csd	4
Index	7

data-expression	<i>Sample expression matrices for CSD</i>
-----------------	---

Description

Sample expression matrices of thyroid gland tissue for thyroid cancer patients and healthy individuals. These datasets were pre-processed by Gulla et al. (2019). Due to size requirements, only 1000 randomly selected genes are provided in the dataset. Number of samples are 399 and 504 in the healthy controls and the sick samples, respectively.

Usage

```
data(normal_expression)
data(sick_expression)
```

Format

Numeric matrices of normalized gene expression. Genes are in columns, whereas samples are in rows.

Source

For the expression matrix for healthy individuals, GenotypeTissue Expression (GTEx) V7. For the thyroid cancer patients, the data are obtained for the Thyroid Cancer project (THCA) from The Cancer Genome Atlas (TCGA).

References

Gulla, Almaas, Eivind, & Voigt, André (2019). An integrated systems biology approach to investigate transcriptomic data of thyroid carcinoma. NTNU. <http://hdl.handle.net/11250/2621725>

partial_argsort	<i>Extract indecies corresponding to the largest elements</i>
-----------------	---

Description

Extracts the indecies of the n largest elements of the input This procedure is equivalent to `order(x, decreasing = TRUE)[1:n_elements]`, but is much faster and avoids the overhead of sorting discarded elements. This function is useful for extracting the rows in a data frame having the largest values in one of the columns.

Usage

```
partial_argsort(x, n_elements)
```

Arguments

<code>x</code>	Numeric vector, the vector containing the numbers to sort.
<code>n_elements</code>	Integer scalar, the number of indecies to return.

Value

Numeric vector, the indecies of the largest elements (in sorted order) in `x`.

Examples

```
x <- c(10L,5L,-2L,12L,15L)
max_indecies <- partial_argsort(x,3L)
max_indecies
x[max_indecies]
order(x)[1:3]
mtcars[partial_argsort(mtcars$hp,5L),]
```

run_cor_bootstrap	<i>Run bootstrapping of Spearman correlations within a dataset</i>
-------------------	--

Description

This function provides the more low-level functionality of bootstrapping the Spearman correlations of the columns within a dataset. Only use this function if you want a low-level interface, else [run_csd](#) provides a more streamlined approach if you want to do a CSD analysis.

Usage

```
run_cor_bootstrap(
  x,
  n_it = 20L,
  nThreads = 1L,
  verbose = TRUE,
  iterations_gap = 1L
)
```

Arguments

<code>x</code>	Numeric matrix, the gene expression matrix to analyse. Genes are in columns, samples are in rows.
<code>n_it</code>	Integer, number of bootstrap iterations
<code>nThreads</code>	Integer, number of threads to use for computations
<code>verbose</code>	Logical, should progress be printed?
<code>iterations_gap</code>	If output is verbose - number of iterations after issue a status message (Default=1 - displayed only if verbose=TRUE)

Value

A list with two fields

rho Numeric matrix constaining the bootstrapped mean of the Spearman correlation between each column

var Numeric matrix constaining the bootstrapped variance of the Spearman correlation between each column

Examples

```
data("normal_expression")
cor_res <- run_cor_bootstrap(
  x = normal_expression,
  n_it = 100, nThreads = 2L
)
```

run_csd

Run CSD analysis

Description

This function implements the a CSD based on the one presented by Voigt et al. 2017. All pairs of genes are first compared within each condition by the Spearman correlation and the correlation and its variance are estimated by bootstrapping. Finally, the results for the two conditions are compared and C-, S- and D-values are computed and returned.

Usage

```
run_csd(
  x_1,
  x_2,
  n_it = 20L,
  nThreads = 1L,
  verbose = TRUE,
  iterations_gap = 1L
)
```

Arguments

<code>x_1</code>	Numeric matrix, the gene expression matrix for the first condition. Genes are in columns, samples are in rows. The columns must be named with the name of the genes.
<code>x_2</code>	Numeric matrix, the gene expression matrix for the second condition.
<code>n_it</code>	Integer, number of bootstrap iterations
<code>nThreads</code>	Integer, number of threads to use for computations
<code>verbose</code>	Logical, should progress be printed?
<code>iterations_gap</code>	If output is verbose - number of iterations after issue a status message (Default=1 - displayed only if verbose=TRUE)

Details

The gene names in `x_1` and `x_2` do not need to be in the same order, but must be in the same namespace. Only genes present in both datasets will be considered for the analysis.

Value

A `data.frame` with the additional class attribute `csd_res` with the results of the CSD analysis. This frame has a row for each pair of genes and has the following columns:

- Gene1** Character, the name of the first gene
- Gene2** Character, the name of the second gene
- rho1** Mean correlation of the two genes in the first condition
- rho2** Mean correlation of the two genes in the second condition
- var1** The estimated variance of rho1 determined by bootstrapping
- var2** The estimated variance of rho2 determined by bootstrapping
- cVal** Numeric, the conserved score. A high value indicates that the co-expression of the two genes have the same sign in both conditions
- sVal** Numeric, the specific score. A high value indicates that the co-expression of the two genes have a high degree of co-expression in one condition, but not the other.
- dVal** Numeric, the differentiated score. A high value indicates that the co-expression of the two genes have a high degree of co-expression in both condition, but the sign of co-expression is different.

References

Voigt A, Nowick K and Almaas E 'A composite network of conserved and tissue specific gene interactions reveals possible genetic interactions in glioma' In: *PLOS Computational Biology* 13(9): e1005739. (doi: <https://doi.org/10.1371/journal.pcbi.1005739>)

Examples

```
data("sick_expression")
data("normal_expression")
cor_res <- run_csd(
  x_1 = sick_expression, x_2 = normal_expression,
  n_it = 100, nThreads = 2L
)
c_max <- max(cor_res$cVal)
```

Index

* datasets

data-expression, 2

data-expression, 2

normal_expression (data-expression), 2

partial_argsort, 3

run_cor_bootstrap, 3

run_csd, 3, 4

sick_expression (data-expression), 2