Package 'ActivePathways'

July 9, 2020

Title Integrative Pathway Enrichment Analysis of Multivariate Omics Data

Version 1.0.2

Description Framework for analysing multiple omics datasets in the context of molecular pathways, biological processes and other types of gene sets. The package uses p-value merging to combine gene- or protein-level signals, followed by ranked hypergeometric tests to determine enriched pathways and processes. This approach allows researchers to interpret a series of omics datasets in the context of known biology and gene function, and discover associations that are only apparent when several datasets are combined. The package is part of the following publication: Integrative Pathway Enrichment Analysis of Multivariate Omics Data. Paczkowska M^, Barenboim J^, Sintupisut N, Fox NS, Zhu H, Abd-Rabbo D, Mee MW, Boutros PC, PCAWG Drivers and Functional Interpretation Working Group; Reimand J, PCAWG Consortium. Nature Communications (2020) <doi:10.1038/s41467-019-13983-9>.

Depends R (>= 3.6)

Imports data.table, ggplot2

License GPL-3

URL

BugReports https://github.com/reimandlab/ActivePathways/issues

Encoding UTF-8

LazyData true

RoxygenNote 7.1.0

Suggests testthat, knitr, rmarkdown, metap, EmpiricalBrownsMethod

VignetteBuilder knitr

NeedsCompilation no

Author Juri Reimand [aut, cre],
Jonathan Barenboim [aut]

Maintainer Juri Reimand <juri.reimand@utoronto.ca>

Repository CRAN

Date/Publication 2020-07-09 17:20:02 UTC

R topics documented:

`ActivePathways` *ActivePathways*

Description

ActivePathways

Usage

```
ActivePathways(
  scores,
  gmt,
  background = makeBackground(gmt),
  geneset.filter = c(5, 1000),
  cutoff = 0.1,
  significant = 0.05,
  merge.method = c("Brown", "Fisher"),
 correction.method = c("holm", "fdr", "hochberg", "hommel", "bonferroni", "BH", "BY",
    "none"),
  cytoscape.file.tag = NA
)
```

Arguments

`scores`	A numerical matrix of p-values where each row is a gene and each column represents an omics dataset (evidence). Rownames correspond to the genes and colnames to the datasets. All values must be 0<=p<=1. We recommend converting missing values to ones.
`gmt`	A GMT object to be used for enrichment analysis. If a filename, a GMT object will be read from the file.
`background`	A character vector of gene names to be used as a statistical background. By default, the background is all genes that appear in `gmt`.

`geneset.filter` A numeric vector of length two giving the lower and upper limits for the size of the annotated geneset to pathways in gmt. Pathways with a geneset shorter than `geneset.filter[1]` or longer than `geneset.filter[2]` will be removed. Set either value to NA to to not enforce a minimum or maximum value, or set `geneset.filter` to `NULL` to skip filtering.

`cutoff` A maximum merged p-value for a gene to be used for analysis. Any genes with merged, unadjusted p > `significant` will be discarded before testing.

`significant` Significance cutoff for selecting enriched pathways. Pathways with `adjusted.p.val` < `significant` will be selected as results.

`merge.method` Statistical method to merge p-values. See section on Merging P-Values

`correction.method` Statistical method to correct p-values. See `p.adjust` for details.

`cytoscape.file.tag` The directory and/or file prefix to which the output files for generating enrichment maps should be written. If NA, files will not be written.

Value

A data.table of terms (enriched pathways) containing the following columns:

term.id The database ID of the term

term.name The full name of the term

adjusted.p.val The associated p-value, adjusted for multiple testing

term.size The number of genes annotated to the term

overlap A character vector of the genes enriched in the term

evidence Columns of `scores` (i.e., omics datasets) that contributed individually to the enrichment of the term. Each input column is evaluated separately for enrichments and added to the evidence if the term is found.

Merging P-values

To obtain a single p-value for each gene across the multiple omics datasets considered, the p-values in `scores` #' are merged row-wise using a data fusion approach of p-value merging. The two available methods are:

Fisher Fisher's method assumes p-values are uniformly distributed and performs a chi-squared test on the statistic sum(-2 log(p)). This method is most appropriate when the columns in `scores` are independent.

Brown Brown's method extends Fisher's method by accounting for the covariance in the columns of `scores`. It is more appropriate when the tests of significance used to create the columns in `scores` are not necessarily independent. The Brown's method is therefore recommended for many omics integration approaches.

Cytoscape

To visualize and interpret enriched pathways, ActivePathways provides an option to further analyse results as enrichment maps in the Cytoscape software. If `!is.na(cytoscape.file.tag)`, four files will be written that can be used to build enrichment maps. This requires the EnrichmentMap and enhancedGraphics apps.

The four files written are:

pathways.txt A list of significant terms and the associated p-value. Only terms with `adjusted.p.val <= significant` are written to this file.

subgroups.txt A matrix indicating whether the significant terms (pathways) were also found to be significant when considering only one column from `scores`. A one indicates that that term was found to be significant when only p-values in that column were used to select genes.

pathways.gmt A Shortened version of the supplied GMT file, containing only the significantly enriched terms in pathways.txt.

legend.pdf A legend with colours matching contributions from columns in `scores`.

How to use: Create an enrichment map in Cytoscape with the file of terms (pathways.txt) and the shortened gmt file (pathways.gmt). Upload the subgroups file (subgroups.txt) as a table using the menu File > Import > Table from File. To paint nodes according to the type of supporting evidence, use the 'style' panel, set image/Chart1 to use the column 'instruct' and the passthrough mapping type. Make sure the app enhancedGraphics is installed. Lastly, use the file legend.pdf as a reference for colors in the enrichment map.

Examples

```
fname_scores <- system.file("extdata", "Adenocarcinoma_scores_subset.tsv",
     package = "ActivePathways")
fname_GMT = system.file("extdata", "hsapiens_REAC_subset.gmt",
     package = "ActivePathways")

dat <- as.matrix(read.table(fname_scores, header = TRUE, row.names = 'Gene'))
dat[is.na(dat)] <- 1

ActivePathways(dat, fname_GMT)
```

`brownsMethod` *Merge p-values using the Brown's method.*

Description

Merge p-values using the Brown's method.

Usage

```
brownsMethod(p.values, data.matrix = NULL, cov.matrix = NULL)
```

Arguments

p.values	A vector of m p-values.
data.matrix	An m x n matrix representing m tests and n samples. NA's are not allowed.
cov.matrix	A pre-calculated covariance matrix of data.matrix. This is more efficient when making many calls with the same data.matrix. Only one of data.matrix and cov.matrix must be given. If both are supplied, data.matrix is ignored.

Value

A single p-value representing the merged significance of multiple p-values.

---

columnSignificance *Determine which terms are found to be significant using each column individually.*

Description

Determine which terms are found to be significant using each column individually.

Usage

```
columnSignificance(
  scores,
  gmt,
  background,
  cutoff,
  significant,
  correction.method,
  pvals
)
```

Arguments

scores	A numerical matrix of p-values where each row is a gene and each column represents an omics dataset (evidence). Rownames correspond to the genes and colnames to the datasets. All values must be 0<=p<=1. We recommend converting missing values to ones.
gmt	A GMT object to be used for enrichment analysis. If a filename, a GMT object will be read from the file.
background	A character vector of gene names to be used as a statistical background. By default, the background is all genes that appear in gmt.
cutoff	A maximum merged p-value for a gene to be used for analysis. Any genes with merged, unadjusted p > significant will be discarded before testing.
significant	Significance cutoff for selecting enriched pathways. Pathways with adjusted.p.val < significant will be selected as results.

correction.method
: Statistical method to correct p-values. See p.adjust for details.

pvals
: p-value for the pathways calculated by ActivePathways

Value

a data.table with columns 'term.id' and a column for each column in scores, indicating whether each term (pathway) was found to be significant or not when considering only that column. For each term, either report the list of related genes if that term was significant, or NA if not.

---

export_as_CSV *Export the results from ActivePathways as a comma-separated values (CSV) file.*

Description

Export the results from ActivePathways as a comma-separated values (CSV) file.

Usage

```
export_as_CSV(res, file_name)
```

Arguments

res
: the data.table object with ActivePathways results.

file_name
: location and name of the CSV file to write to.

Examples

```
fname_scores <- system.file("extdata", "Adenocarcinoma_scores_subset.tsv",
     package = "ActivePathways")
fname_GMT = system.file("extdata", "hsapiens_REAC_subset.gmt",
     package = "ActivePathways")

dat <- as.matrix(read.table(fname_scores, header = TRUE, row.names = 'Gene'))
dat[is.na(dat)] <- 1

res <- ActivePathways(dat, fname_GMT)

export_as_CSV(res, "results_ActivePathways.csv")
```

GMT *Read and Write GMT files*

Description

Functions to read and write Gene Matrix Transposed (GMT) files and to test if an object inherits from GMT.

Usage

```
read.GMT(filename)

write.GMT(gmt, filename)

is.GMT(x)
```

Arguments

filename	Location of the gmt file.
gmt	A GMT object.
x	The object to test.

Format

A GMT object is a named list of terms, where each term is a list with the items:

id The term ID.

name The full name or description of the term.

genes A character vector of genes annotated to this term.

Details

A GMT file describes gene sets, such as biological terms and pathways. GMT files are tab delimited text files. Each row of a GMT file contains a single term with its database ID and a term name, followed all genes annotated to the term.

Value

`read.GMT` returns a GMT object.
`write.GMT` returns NULL.
`is.GMT` returns TRUE if `x` is a GMT object, else FALSE.

Examples

```
fname_GMT <- system.file("extdata", "hsapiens_REAC_subset.gmt", package = "ActivePathways")
gmt <- read.GMT(fname_GMT)
gmt[1:10]
gmt[[1]]
gmt[[1]]$id
gmt[[1]]$genes
gmt[[1]]$name
gmt$`REAC:1630316`
```

`hypergeometric` *Hypergemoetric test*

Description

Perform a hypergeometric test, also known as the Fisher's exact test, on a 2x2 contingency table with the alternative hypothesis 'greater' . In this application, the test finds the probability that counts[1, 1] or more genes would be found to be annotated to a term (pathway), assuming the null hypothesis of genes being distributed randomly to terms.

Usage

```
hypergeometric(counts)
```

Arguments

`counts`	A 2x2 numerical matrix representing a contingency table.

Value

a p-value of enrichment of genes in a term or pathway.

`makeBackground` *Make a background list of genes (i.e., the statistical universe) based on all the terms (gene sets, pathways) considered.*

Description

Returns A character vector of all genes in a GMT object.

Usage

```
makeBackground(gmt)
```

Arguments

`gmt`	A GMT object.

Value

A character vector containing all genes in GMT.

Examples

```
fname_GMT <- system.file("extdata", "hsapiens_REAC_subset.gmt", package = "ActivePathways")
gmt <- read.GMT(fname_GMT)
makeBackground(gmt)[1:10]
```

---

merge_p_values *Merge a list or matrix of p-values*

Description

Merge a list or matrix of p-values

Usage

```
merge_p_values(scores, method = "Fisher")
```

Arguments

scores	Either a list of p-values or a matrix where each column is a test.
method	Method to merge p-values. See 'methods' section below.

Value

If scores is a vector or list, returns a number. If scores is a matrix, returns a named list of p-values merged by row.

Methods

Two methods are available to merge a list of p-values:

Fisher Fisher's method (default) assumes that p-values are uniformly distributed and performs a chi-squared test on the statistic sum(-2 log(p)). This method is most appropriate when the columns in scores are independent.

Brown Brown's method extends Fisher's method by accounting for the covariance in the columns of scores. It is more appropriate when the tests of significance used to create the columns in scores are not necessarily independent. Note that the "Brown" method cannot be used with a single list of p-values. However, in this case Brown's method is identical to Fisher's method and should be used instead.

Examples

```
merge_p_values(c(0.05, 0.09, 0.01))
merge_p_values(list(a=0.01, b=1, c=0.0015, d=0.025), method='Fisher')
merge_p_values(matrix(data=c(0.03, 0.061, 0.48, 0.052), nrow = 2), method='Brown')
```

`orderedHypergeometric` *Ordered Hypergeometric Test*

Description

Perform a series of hypergeometric tests (a.k.a. Fisher's Exact tests), on a ranked list of genes ordered by significance against a list of annotation genes. The hypergeometric tests are executed with increasingly larger numbers of genes representing the top genes in order of decreasing scores. The lowest p-value of the series is returned as the optimal enriched intersection of the ranked list of genes and the biological term (pathway).

Usage

```
orderedHypergeometric(genelist, background, annotations)
```

Arguments

`genelist`	Character vector of gene names, assumed to be ordered by decreasing importance. For example, the genes could be ranked by decreasing significance of differential expression.
`background`	Character vector of gene names. List of all genes used as a statistical background (i.e., the universe)
`annotations`	Character vector of gene names. A gene set representing a functional term, process or biological pathway.

Value

a list with the items:

p.val The lowest obtained p-value

ind The index of `genelist` such that `genelist[1:ind]` gives the lowest p-value

Examples

```
orderedHypergeometric(c('HERC2', 'SP100'), c('PHC2', 'BLM', 'XPC', 'SMC3', 'HERC2', 'SP100'),
                      c('HERC2', 'PHC2', 'BLM'))
```

prepareCytoscape *Prepare files for building an enrichment map network visualization in Cytoscape*

Description

This function writes four text files that are used to build an network using Cytoscape and the EnrichmentMap app. The files are prefixed with `cytoscape.file.tag`. The four files written are:

pathways.txt A list of significant terms and the associated p-value. Only terms with `adjusted.p.val <= significant` are written to this file

subgroups.txt A matrix indicating whether the significant pathways are found to be significant when considering only one column (i.e., type of omics evidence) from `scores`. A 1 indicates that that term is significant using only that column to test for enrichment analysis

pathways.gmt A shortened version of the supplied GMT file, containing only the terms in pathways.txt.

legend.pdf A legend with colours matching contributions from columns in `scores`

Usage

```
prepareCytoscape(terms, gmt, cytoscape.file.tag, col.significance)
```

Arguments

`terms`	A data.table object with the columns 'term.id', 'term.name', 'adjusted.p.val'.
`gmt`	An abridged GMT object containing only the pathways that were found to be significant in the ActivePathways analysis.
`cytoscape.file.tag`	The user-defined file prefix and/or directory defining the location of the files.
`col.significance`	A data.table object with a column 'term.id' and a column for each type of omics evidence indicating whether a term was also found to be signficiant or not when considering only the genes and p-values in the corresponding column of the `scores` matrix. If term was not found, NA's are shown in columns, otherwise the relevant lists of genes are shown.

Value

None

Index