Package ‘diffEnrich’

June 27, 2022

Title  Given a List of Gene Symbols, Performs Differential Enrichment Analysis

Version  0.1.2

Description  Compare functional enrichment between two experimentally-derived groups of genes or proteins (Peterson, DR., et al.(2018)) <doi:10.1371/journal.pone.0198139>. Given a list of gene symbols, ‘diffEnrich’ will perform differential enrichment analysis using the Kyoto Encyclopedia of Genes and Genomes (KEGG) REST API. This package provides a number of functions that are intended to be used in a pipeline. Briefly, the user provides a KEGG formatted species id for either human, mouse or rat, and the package will download and clean species specific ENTREZ gene IDs and map them to their respective KEGG pathways by accessing KEGG's REST API. KEGG's API is used to guarantee the most up-to-date pathway data from KEGG. Next, the user will identify significantly enriched pathways from two gene sets, and finally, the user will identify pathways that are differentially enriched between the two gene sets. In addition to the analysis pipeline, this package also provides a plotting function.

License  GPL-2

Encoding  UTF-8

LazyData  true

RoxygenNote  7.0.2

URL  https://github.com/SabaLab/diffEnrich

BugReports  https://github.com/SabaLab/diffEnrich/issues

Suggests  knitr, rmarkdown, kableExtra, diagram

Depends  dplyr, ggplot2, R (>= 2.10)

Imports  here, stats, rlang, stringr, reshape2, ggnewscale

VignetteBuilder  knitr

NeedsCompilation  no

Author  Harry Smith [aut, cre],
          Laura Saba [aut]

Maintainer  Harry Smith <harry.smith@ucdenver.edu>
Description

This function connects to the KEGG API, downloads, and cleans ncbi gene ID data, KEGG pathway descriptions, and species specific data. Currently, this function supports Human, Mouse, and Rat. Files will be written to the working directory unless otherwise specified by the user.

Usage

.api_pull(species, path = path)

Arguments

species character. The species to use in kegg data pull
path character. A character string describing the path to write out KEGG API data sets. If not provided, defaults to current working directory.

Value

kegg_out: A named list of the data pulled from kegg api when the function was run. This may be different if the function is run at different times. For reproducible results, use text files generated by function that include the date they were pulled.
**Description**

This is a helper function for `diffEnrich`. This function takes the objects generated from `pathEnrich`. If performing a differential enrichment analysis, the user will have 2 objects. There will be one for list1 and one for list2 (see example for `pathEnrich`). This function then merges the two data frames using the following columns that should be present in both objects (by = c("KEGG_PATHWAY_ID", "KEGG_PATHWAY_description", "KEGG_PATHWAY_cnt", "KEGG_DATABASE_cnt")). This merged data frame will be used as the input for the differential enrichment function. Any pathways that do not contain any genes from either gene list will be removed.

**Usage**

```
.combineEnrich(list1_pe, list2_pe)
```

**Arguments**

- `list1_pe` object of class `pathEnrich` generated from `pathEnrich`. See example for `pathEnrich`.
- `list2_pe` object of class `pathEnrich` generated from `pathEnrich`. See example for `pathEnrich`.

**Value**

`combined_enrich`: An object of class data.frame that is the result of merging `list1_pe` and `list2_pe`, using the default joining implemented in the base `merge` function.

**Description**

This function reads in the text files generated from a previous `get_kegg` call and saves them as a names list formatted for down stream analysis.

**Usage**

```
.data_read(path = path, date = date, release = release)
```

**Arguments**

- `path` character. A character string describing the path to write out KEGG API data sets. If not provided, defaults to current working directory.
- `date` character. A character string describing the date that was used to time stamp files from previous call. Must be formatted like YYYY-MM-DD.
- `release` character. A character string describing the KEGG release that was used to time stamp files from previous call (e.g. "90" or "92")
**Value**

kegg_out: A named list of the data pulled from kegg api when the function was run. This may be different if the function is run at different times. For reproducible results, use text files generated by function that include the date they were pulled.

**Description**

This function takes the objects generated from `pathEnrich`. If performing a differential enrichment analysis, the user will have 2 objects. There will be one for the genes of interest in gene list 1 and one for the genes of interest in gene list 2 (see example for `pathEnrich`). This function then uses a Fisher's Exact test to identify differentially enriched pathways between the terms enriched in the gene-of-interest lists. `diffEnrich` will remove KEGG pathways that do not contain any genes from either gene list as these cannot be tested, and will print a warning message telling the user how many pathways were removed. `diffEnrich` returns a dataframe containing differentially enriched pathways with their associated estimated odds ratio, unadjusted p-value, and fdr adjusted p-value. S3 generic functions for `print` and `summary` are provided. The `print` function prints the results table as a tibble, and the `summary` function returns the number of pathways that reached statistical significance as well as their descriptions, the number of genes used from the KEGG data base, the KEGG species, the number of pathways that were shared (and therefore tested) between the gene lists and the method used for multiple testing correction.

**Usage**

```r
diffEnrich(list1_pe, list2_pe, method = "BH", cutoff = 0.05)
```

```r
## S3 method for class 'diffEnrich'
print(x, ...)
```

```r
## S3 method for class 'diffEnrich'
summary(object, ...)
```

**Arguments**

- `list1_pe` object of class `pathEnrich` generated from `pathEnrich`. See example for `pathEnrich`.
- `list2_pe` object of class `pathEnrich` generated from `pathEnrich`. See example for `pathEnrich`.
- `method` character. Character string telling `diffEnrich` which method to use for multiple testing correction. Available methods are those provided by `p.adjust`, and the default is "BH", or False Discovery Rate (FDR).
- `cutoff` Numeric. The p-value threshold to be used as the cutoff when determining statistical significance, and used to filter list of significant pathways.
- `x` object of class `diffEnrich`
- `...` Unused
- `object` object of class `diffEnrich`
**Value**

A list object of class `diffEnrich` that contains 5 items:

- **species**  The species used in enrichment
- **padj**  The method used to correct for multiple testing for the differential enrichment
- **sig_paths**  The KEGG pathways the reached statistical significance after multiple testing correction.
- **path_intersect**  the number of pathways that were shared (and therefore tested) between the gene lists.
- **de_table**  A data frame that summarizes the results of the differential enrichment analysis and contains the following variables:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>KEGG_PATHWAY_ID</strong></td>
<td>KEGG Pathway Identifier</td>
</tr>
<tr>
<td><strong>KEGG_PATHWAY_description</strong></td>
<td>Description of KEGG Pathway (provided by KEGG)</td>
</tr>
<tr>
<td><strong>KEGG_PATHWAY_cnt</strong></td>
<td>Number of Genes in KEGG Pathway</td>
</tr>
<tr>
<td><strong>KEGG_DATABASE_cnt</strong></td>
<td>Number of Genes in KEGG Database</td>
</tr>
<tr>
<td><strong>KEGG_PATHWAY_in_list1</strong></td>
<td>Number of Genes from gene list 1 in KEGG Pathway</td>
</tr>
<tr>
<td><strong>KEGG_DATABASE_in_list1</strong></td>
<td>Number of Genes from gene list 1 in KEGG Database</td>
</tr>
<tr>
<td><strong>expected_list1</strong></td>
<td>Expected number of genes from list 1 to be in KEGG pathway by chance (i.e., not enriched)</td>
</tr>
<tr>
<td><strong>enrich_p_list1</strong></td>
<td>P-value for enrichment of list 1 genes related to KEGG pathway</td>
</tr>
<tr>
<td><strong>p_adj_list1</strong></td>
<td>Multiple testing adjustment of enrich_p_list1 (default = False Discovery Rate (Benjamini and Hochberg))</td>
</tr>
<tr>
<td><strong>fold_enrichment_list1</strong></td>
<td>KEGG_PATHWAY_in_list1/expected_list1</td>
</tr>
<tr>
<td><strong>KEGG_PATHWAY_in_list2</strong></td>
<td>Number of Genes from gene list 2 in KEGG Pathway</td>
</tr>
<tr>
<td><strong>KEGG_DATABASE_in_list2</strong></td>
<td>Number of Genes from gene list 2 in KEGG Database</td>
</tr>
<tr>
<td><strong>expected_list2</strong></td>
<td>Expected number of genes from list 2 to be in KEGG pathway by chance (i.e., not enriched)</td>
</tr>
<tr>
<td><strong>enrich_p_list2</strong></td>
<td>P-value for enrichment of list 2 genes related to KEGG pathway</td>
</tr>
<tr>
<td><strong>p_adj_list2</strong></td>
<td>Multiple testing adjustment of enrich_p_list2 (default = False Discovery Rate (Benjamini and Hochberg))</td>
</tr>
<tr>
<td><strong>fold_enrichment_list2</strong></td>
<td>KEGG_PATHWAY_in_list2/expected_list2</td>
</tr>
<tr>
<td><strong>odd_ratio</strong></td>
<td>Odds of a gene from list 2 being from this KEGG pathway / Odds of a gene from list 1 being from this KEGG pathway</td>
</tr>
<tr>
<td><strong>diff_enrich_p</strong></td>
<td>P-value for differential enrichment of this KEGG pathway between list 1 and list 2</td>
</tr>
<tr>
<td><strong>diff_enrich_adjusted</strong></td>
<td>Multiple testing adjustment of diff_enrich_p (default = False Discovery Rate (Benjamini and Hochberg))</td>
</tr>
</tbody>
</table>
Examples

```r
## Generate individual enrichment results
list1_pe <- pathEnrich(gk_obj = kegg, gene_list = geneLists$list1)
list2_pe <- pathEnrich(gk_obj = kegg, gene_list = geneLists$list2)

## Perform differential enrichment
dif_enrich <- diffEnrich(list1_pe = list1_pe, list2_pe = list2_pe, method = 'none', cutoff = 0.05)
```

Description

This is a list object that contains the list background genes and significant genes used in pathway enrichment. This object is mostly meant for running examples and vignettes. The data provided is for the rat, and is loaded from org.Rn.eg.db version 3.7.0.

Usage

geneLists

Format

A list with two names items which are:

- **background** List of ENTREZ gene IDs that will considered background
- **sigGenes** List of ENTREZ gene IDs that were significant

Description

This function calls an internal helper function that connects to the KEGG API, downloads, and stores ncbi gene ID data, KEGG pathway descriptions, and species specific data. Currently, this function supports Human, Mouse, and Rat. Files will be written to the working directory unless otherwise specified by the user.

Usage

get_kegg(species, read = FALSE, path = NULL, date, release)
get_kegg

Arguments

- **species** character. The species to use in kegg data pull.
- **read** logical. Should get_kegg read in files from previous call. If TRUE, all 3 files generated by get_kegg must be in the same directory and the user must provide a file path that points to that directory.
- **path** character. A character string describing the path to write out KEGG API data sets. If not provided, defaults to current working directory.
- **date** character. A character string describing the date that was used to time stamp files from previous call. Must be formatted like YYYY-MM-DD.
- **release** character. A character string describing the KEGG release that was used to time stamp files from previous call (e.g. "90" or "92")

Details

The get_kegg function is used to connect to the KEGG REST API and download the data sets required to perform downstream analysis. Currently, this function supports three species, and recognizes the KEGG code for Homo sapiens ("hsa"), Mus musculus ("mmu"), and Rattus norvegicus ("rno"). For a given species, three data sets are generated: 1) Because the user must provide their own gene lists in downstream analysis using ENTREZ gene IDs, the data set maps NCBI/ENTREZ gene IDs to KEGG gene IDs, 2) a data set that maps KEGG gene IDs to their respective KEGG pathway IDs, and 3) a data set that maps KEGG pathway IDs to their respective pathway descriptions. This function allows the user save versioned (based on KEGG release) and time-stamped text files of the three data sets described above. In addition to these flat files, get_kegg() will also create a named list with the three relevant KEGG data sets. The names of this list will describe the data set.

Table 1. Description of get_kegg list object

<table>
<thead>
<tr>
<th><strong>get_kegg_list_object</strong></th>
<th><strong>Object_description</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>ncbi_to_kegg</td>
<td>ncbi gene ID &lt;- mapped to -&gt; KEGG gene ID</td>
</tr>
<tr>
<td>kegg_to_pathway</td>
<td>KEGG gene ID &lt;- mapped to -&gt; KEGG pathway ID</td>
</tr>
<tr>
<td>pathway_to_species</td>
<td>KEGG pathway ID &lt;- mapped to -&gt; KEGG pathway species description</td>
</tr>
</tbody>
</table>

Value

- **kegg_out**: A named list of the data pulled from kegg api when the function was run. This may be different if the function is run at different times. For reproducible results, use text files generated by function that include the date they were pulled.

Examples

```r
## Not run:
kegg <- get_kegg(species = "rno")
```
```r
## End(Not run)
## Not run:
kegg <- get_kegg(species = "mmu", path = "usr/data/out/")
kegg <- get_kegg(species = "mmu", path = "usr/data/out/",
read = TRUE,
date = "2019-09-30",
release = "92")

## End(Not run)

kegg

### Description
This is a list object that contains the output generated from the get_kegg function. This object is mostly meant for running examples and vignettes.

### Usage
kegg

### Format
A list with three names items which are:

- **kegg_to_pathway** List of kegg IDs mapped to pathway IDs
- **ncbi_to_kegg** List of ENTREZ gene IDs that map to kegg IDs
- **pathway_to_species** List of pathways IDs that map to rat pathways

### Description
This function takes the list generated in get_kegg as well as a vector of NCBI (ENTREZ) geneIDs, and identifies significantly enriched KEGG pathways using a Fisher’s Exact Test. Unadjusted p-values as well as FDR corrected p-values are calculated.

### Usage
pathEnrich(gk_obj, gene_list, method = "BH", cutoff = 0.05, N = 2)

```

```
Arguments

gk_obj  list. Object generated from `get_kegg`, or a list containing the output generated from a past `get_kegg` call. Names of the list must match those defined in `get_kegg`. If the user wishes to use an older version of data generated by `get_kegg`, they must first load that data and put it in a named list that matches the names given in the list generated by `get_kegg`.

gene_list  Vector. Vector of NCBI (ENTREZ) geneIDs.

method  Character. Character string telling `diffEnrich` which method to use for multiple testing correction. Available methods are those provided by `p.adjust`, and the default is "BH", or False Discovery Rate (FDR).

cutoff  Numeric. The p-value threshold to be used as the cutoff when determining statistical significance, and used to filter list of significant pathways.

N  Numeric. The number of genes from the gene list that must be present in a KEGG pathway in order for that pathway to be retained and tested.

x  object of class `pathEnrich`

...  Unused

object  object of class `pathEnrich`

Details

This function may not always use the complete list of genes provided by the user. Specifically, it will only use the genes from the list provided that are also in the most current species list pulled from the KEGG REST API, or from the older data KEGG loaded by the user. The ‘cutoff’ only filters the list of pathways provided in the ‘sig_paths’ list item. It is not used to filter the ‘enrich_table’ list object. S3 generic functions for `print` and `summary` are provided. The `print` function prints the results table as a `tibble`, and the `summary` function returns the number of pathways that reached statistical significance, as well as their descriptions, the number of genes used from the KEGG data base, the KEGG species, and the method used for multiple testing correction, and the p-value cutoff required for reaching statistical significance.

Value

A list object of class `pathEnrich` that contains 6 items:

species  The species used in enrichment

padj  The method used to correct for multiple testing

sig_paths  The KEGG pathways the reached statistical significance after multiple testing correction.

cutoff  The p-value threshold to be used as the cutoff when determining statistical significance, and used to filter final results data set.

N  The number of genes from the gene list that must be present in a KEGG pathway in order for that pathway to be retained and tested.

enrich_table  A data frame that summarizes the results of the pathway analysis and contains the following variables:
**plotFoldEnrichment**

<table>
<thead>
<tr>
<th>KEGG_PATHWAY_ID</th>
<th>KEGG Pathway Identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEGG_PATHWAY_description</td>
<td>Description of KEGG Pathway (provided by KEGG)</td>
</tr>
<tr>
<td>KEGG_PATHWAY_cnt</td>
<td>Number of Genes in KEGG Pathway</td>
</tr>
<tr>
<td>KEGG_PATHWAY_in_list</td>
<td>Number of Genes from gene list in KEGG Pathway</td>
</tr>
<tr>
<td>KEGG_DATABASE_cnt</td>
<td>Number of Genes in KEGG Database</td>
</tr>
<tr>
<td>KEGG_DATABASE_in_list</td>
<td>Number of Genes from gene list in KEGG Database</td>
</tr>
<tr>
<td><strong>expected</strong></td>
<td>Expected number of genes from list to be in KEGG pathway by chance (i.e., not enriched)</td>
</tr>
<tr>
<td><strong>enrich_p</strong></td>
<td>P-value for enrichment of list genes related to KEGG pathway</td>
</tr>
<tr>
<td><strong>p_adj</strong></td>
<td>False Discovery Rate (Benjamini and Hochberg) to account for multiple testing across KEGG pathways</td>
</tr>
<tr>
<td><strong>fold_enrichment</strong></td>
<td>KEGG_PATHWAY_in_list/expected</td>
</tr>
</tbody>
</table>

**Examples**

```r
list1_pe <- pathEnrich(gk_obj = kegg, gene_list = geneLists$list1)
## Not run:
list2_pe <- pathEnrich(gk_obj = kegg, gene_list = geneLists$list2, method = 'none', N = 4)
## End(Not run)
```

**Description**

This function uses the results generated using `diffEnrich` to generate a bar plot describing the fold enrichment of a set of given KEGG pathways stratified by their enrichment in list 1 or list 2. Users can plot all pathways based on the adjusted p-value threshold used in `diffEnrich` and the top N pathways sorted by the adjusted p-value threshold used in `diffEnrich`. `plotFoldEnrichment` returns a ggplot2 object so users can add additional customizations.

**Usage**

```r
plotFoldEnrichment(de_res, pval, N)
```

**Arguments**

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>de_res</td>
<td>Dataframe. Generated using <code>diffEnrich</code></td>
</tr>
<tr>
<td>pval</td>
<td>Numeric. Threshold for filtering pathways based on adjusted pvalue in de_res</td>
</tr>
<tr>
<td>N</td>
<td>Numeric. Number of top pathways to plot after filtering based on pval</td>
</tr>
</tbody>
</table>
Details

This function generates a grouped bar plot using ggplot2 and the ggnewscale package. KEGG pathways are plotted on the y-axis and fold enrichment is plotted on the x-axis. Each KEGG pathway has a bar plotting its fold enrichment in list 1 (red) and its fold enrichment in list 2 (blue). The transparency of the bars correspond to the adjusted p-value for the pathway’s enrichment in the given list. The p-value presented as text to the right of each pair of bars is the adjusted p-value associated with the differential enrichment of the pathway between the two lists, and the pathways are ordered from top to bottom by this p-value (i.e. smallest p-value on top of plot, and largest p-value on bottom of plot).

Value

ggplot object. If the user calls plotFoldEnrich and assigns it to an object (see example) then no plot will print in viewer, but if plotFoldEnrich is called without being assigned to an object the plot will print to the viewer. Users can edit the ggplot object as they would any other ggplot object (e.g. add title, theme, etc.).

Examples

## Not run:
plot <- plotFoldEnrichment(de_res = diff_enrich, pval = 0.05, N = 5)

## End(Not run)
Index

* datasets
  geneLists, 6
  kegg, 8
  .api_pull, 2
  .combineEnrich, 3
  .data_read, 3
  diffEnrich, 4, 10
  geneLists, 6
  get_kegg, 6, 8
  kegg, 8
  merge, 3
  p.adjust, 4, 9
  pathEnrich, 3, 4, 8
  plotFoldEnrichment, 10
  print.diffEnrich (diffEnrich), 4
  print.pathEnrich (pathEnrich), 8
  summary.diffEnrich (diffEnrich), 4
  summary.pathEnrich (pathEnrich), 8