Package 'DEET'

June 26, 2024

Title Differential Expression Enrichment Tool

Version 1.0.12

Maintainer Dustin Sokolowski <dustin.sokolowski@sickkids.ca>

Description Abstract of Manuscript. Differential gene expression analysis using RNA sequencing (RNA-seq) data is a standard approach for making biological discoveries. Ongoing large-scale efforts to process and normalize publicly available gene expression data enable rapid and systematic reanalysis. While several powerful tools systematically process RNA-seq data, enabling their reanalysis, few resources systematically recompute differentially expressed genes (DEGs) generated from individual studies. We developed a robust differential expression analysis pipeline to recompute 3162 human DEG lists from The Cancer Genome Atlas, Genotype-Tissue Expression Consortium, and 142 studies within the Sequence Read Archive. After measuring the accuracy of the recomputed DEG lists, we built the Differential Expression Enrichment Tool (DEET), which enables users to interact with the recomputed DEG lists. DEET, available through CRAN and RShiny, systematically queries which of the recomputed DEG lists share similar genes, pathways, and TF targets to their own gene lists. DEET identifies relevant studies based on shared results with the user's gene lists, aiding in hypothesis generation and data-driven literature review. Sokolowski, Dustin J., et al. ``Differential Expression Enrichment Tool (DEET): an interactive atlas of human differential gene expression." Nucleic Acids Research Genomics and Bioinformatics (2023).

Depends R (>= 3.5.0)

Imports ActivePathways, pbapply, dplyr, ggplot2, glmnet, utils, stats, ggrepel, downloader

License GPL-3

Encoding UTF-8

LazyData true

RoxygenNote 7.2.3

Suggests testthat, knitr, rmarkdown

VignetteBuilder knitr

NeedsCompilation no

Author Dustin Sokolowski [aut, cre], Jedid Ahn [aut],

Lauren Erdman [aut],
Kai Ellis [aut],
Huayun Hou [aut],
Anna Goldenberg [aut],
Michael Wilson [aut]

Repository CRAN

Date/Publication 2024-06-26 10:50:02 UTC

Contents

`adjust_DE_cutoffs` *adjust_DE_cutoffs*

Description

Utility function to adjust mean expression, FDR, and log2FC cutoffs of the database of DEGs inputted into DEET.

Usage

```
adjust_DE_cutoffs(
  DEET_combined,
  redo_pathways_instructions = FALSE,
  baseMean = 1,
  abslog2FoldChange = 0,
  padj = 0.05
)
```

Arguments

DEET_combined	The databank of the differential expression enrichment tool. Appropriate inputs here are "DEET_example_data" stored within DEET, the "DEET_combined.rda" file from the DEET stable repositoy found at X, and the DEET database developmental repository found at Y. The DEET_dataset is a named list where details of it's structure can be found ?DEET_example_data.
redo_pathways_instructions	Boolean value specifying whether to print the instructions required to update all pathway enrichments based on new DE cutoffs.
baseMean	Change the mean-expression cutoff.
abslog2FoldChange	Change the log2 Fold-change cutoff.
padj	Change the FDR-adjusted p-value cutoff.

Value

The DEET_combined object but with the user-inputted expression, log2FC, and FDR-adjusted p-value cutoffs. DEET_gmt_DE is also updated to the new cutoffs.

Author(s)

Dustin Sokolowski

Examples

```
data("DEET_example_data")
DEET_cutoff <- adjust_DE_cutoffs(DEET_example_data, abslog2FoldChange = 1, padj = 0.01)
```

DEET_data_download *DEET_data_download*

Description

Function to automatically download the files within the DEET database that are required for the DEET_enrich and DEET_feature_extractfunctions.

Usage

```
DEET_data_download(x = "enrich")
```

Arguments

x	categorical variable containing options "ALL", "enrich", "metadata" or "feature_matrix".

Value

Named list with the neccesary data required to input into DEET_feature_extract or DEET_enrich. The metadata within DEET can also be downloaded.

- feature_matrix - A gene by comparison matrix populated with the log2FC of gene expression for all genes, regardless of DE status.
- metadata - a comparison - by - explanatory piece of data dataframe providing important details to contextualize each study. For every pairwise comparison, the study name, source (SRA, TCGA, GTEx and SRA-manual), description from the DRA compendium, the number of samples (total, up-condition, and down-condition), samples (total ,up-condition, down-condition), tissue (including tumour from TCGA), number of DEs (total, up-condition, down-condition), age (mean +- sd), sex, top 15 DEGs - up, top 15 DEGs - down, top 5 enriched pathways, and top 5 enriched TFs. PMID are also available for studies selected from SRA. Lastly, each pairwise comparison was given an overall category based on those decided in Crow et al., 2019.
- DEET_enrich - A named list of seven objects containing the data frames summarizing the DEGs from comparisons within DEET, GMT objects of comparisons within DEET for enrichment through ActivePathways, GMT objects for basic pathway and TF enrichment, and a dataframe for the metadata of each study. For more detail on each element of the list, please consult the vignette or "?DEET_example_data", as it is a subset of this object

Author(s)

Dustin Sokolowski, Jedid Ahn

References

Engebretsen, S., & Bohlin, J. (2019). Statistical predictions with glmnet. Clinical epigenetics, 11(1), 1-3.

Examples

```
# Download the metadata. Downloading other
# files within DEET are larger and take
# a bit more time.
downloaded <- DEET_data_download(x = "metadata")

# extract metadata from the list
metadata <- downloaded[["metadata"]]
```

DEET_enrich *DEET_enrich*

Description

Core function of DEET where an input weighted human gene list will be queried to DEETs library of studies.

Usage

```
DEET_enrich(
  DEG_list,
  DEET_dataset,
  ordered = FALSE,
  background = NULL,
  abs_cor = FALSE
)
```

Arguments

DEG_list	Data frame or matrix of gene symbols with corresponding padj and log2FC values (3 columns in total). Can also be a character vector of gene symbols only. colnames of genes: c("gene_symbol", "padj", "coef") The rownames of the dataframe are also the gene symbols.
DEET_dataset	The databank of the differential expression enrichment tool. Appropriate inputs here are "DEET_example_data" stored within DEET, the "DEET_combined.rda" file from the DEET stable repositoy found at X, and the DEET database developmental repository found at Y. The DEET_dataset is a named list where details of it's structure can be found ?DEET_example_data.
ordered	Boolean value specifying whether DEG_list is a character vector of gene symbols that is ordered. Default value is FALSE.
background	Character vector of human gene symbols showing all possible genes. Default value is NULL.
abs_cor	Boolean value that forces log2FC's in DEET to be their absolute value. Use when the directionality of the coefficient is unknown (or includes both up- down-directions). Default value is FALSE.

Value

Named list where each element contains 6 objects. Each object will contain the results (enrichment or correlation) and corresponding metadata.

- AP_INPUT_BP_output - Enriched BPs of input gene list.
- AP_INPUT_TF_output - Enriched TFs of input gene list.
- AP_DEET_DE_output - Enrichment of input gene list on DEETs studies.

- AP_DEET_BP_output - Enrichment of BPs of input gene list on DEETs BPs of studies.
- AP_DEET_TF_output - Enrichment of TFs of input gene list on DEETs TFs of studies.
- DE_correlations - Correlation values of input gene list to DEETs studies (both Pearson and Spearman).

Author(s)

Dustin Sokolowski, Jedid Ahn

References

Paczkowska M, Barenboim J, Sintupisut N, et al. Integrative pathway enrichment analysis of multivariate omics data. Nat Commun. 2020;11(1):735. doi:10.1038/s41467-019-13983-9

Examples

```
data("example_DEET_enrich_input")
data("DEET_example_data")
DEET_out <- DEET_enrich(example_DEET_enrich_input, DEET_dataset = DEET_example_data)
```

`DEET_enrichment_plot` *DEET_enrichment_plot*

Description

Generate barplots or dotplots from the output of DEET

Usage

```
DEET_enrichment_plot(
  enrich_list,
  outname,
  width = 8,
  text_angle = 0,
  horizontal = FALSE,
  topn = 5,
  ol_size = 1,
  exclude_domain = "",
  cluster_order = NULL,
  dot = FALSE,
  colors = "Set2",
  split_domain = FALSE
)
```

Arguments

enrich_list	A list of enrichments from DEET, with each element post-processed with the barplot enrichment function.
outname	A character giving the title of the barplot or dotplot.
width	The number of inches in the barplot or dotplot.
text_angle	The angle of the enriched studies.
horizontal	Whether the output barplot is vertical or horizontal
topn	the top number of studies (by p-value) to be plotted.
ol_size	the minimum number of overlapping genes (or paths) in an enriched study.
exclude_domain	Exclude studies enriched based on DEGs, Paths, or TF if the user happened to aggregate the results into a single DF, generally unused.
cluster_order	Factor to group studies based on the researchers custom annotation.
dot	logical (T/F) of whether to produce a dotplot or a barplot
colors	Type of color pallete to input into 'scale_fill_brewer' of ggplot.
split_domain	logical (T/F) of whether to plot the "topn" studies for each "domain" (default is source) or to plot the topn pathwys regardless of domain. default is set to FALSE, meaning it plots the topn pathways regardless of domain.

Value

A ggplot2 object (barplot or dotplot) of enrichment identified within DEET.

Author(s)

Dustin Sokolowski, Hauyun Hou PhD

Examples

```
data("example_DEET_enrich_input")
data("DEET_example_data")
DEET_out <- DEET_enrich(example_DEET_enrich_input, DEET_dataset = DEET_example_data)

# converting output to format compatible with DEET_enrichment plot
DE_example <- DEET_out$AP_DEET_DE_output$results
DE_example$term.name <- DEET_out$AP_DEET_DE_output$metadata$DEET.Name
DE_example$domain <- "DE"
DE_example$overlap.size <- lengths(DE_example$overlap)
DE_example$p.value <- DE_example$adjusted_p_val

DE_example_plot <- DEET_enrichment_plot(list(DE_example = DE_example), "DE_example")
```

DEET_enrich_genesonly *DEET_enrich_genesonly*

Description

Altered version of the function of DEET where an input weighted human gene list will be queried to DEETs library of studies. This version does not include pathway enrichments.

Usage

```
DEET_enrich_genesonly(
  DEG_list,
  DEET_dataset,
  ordered = FALSE,
  background = NULL
)
```

Arguments

DEG_list	Data frame or matrix of gene symbols with corresponding padj and log2FC values (3 columns in total). Can also be a character vector of gene symbols only. colnames of genes: c("gene_symbol", "padj", "coef") The rownames of the dataframe are also the gene symbols.
DEET_dataset	The databank of the differential expression enrichment tool. Appropriate inputs here are "DEET_example_data" stored within DEET, the "DEET_combined.rda" file from the DEET stable repositoy found at X, and the DEET database developmental repository found at Y. The DEET_dataset is a named list where details of it's structure can be found ?DEET_example_data. Unlike in DEET_enrich, this dataset does not require the pathway-relevant elements of the DEET_dataset list, namely "gmt_BP", or "gmt_TF" "DEET_gmt_BP", "DEET_gmt_TF".
ordered	Boolean value specifying whether DEG_list is a character vector of gene symbols that is ordered. Default value is FALSE.
background	Character vector of human gene symbols showing all possible genes. Default value is NULL.

Value

Named list where each element contains 2 objects. Each object will contain the results (enrichment or correlation) and corresponding metadata.

- AP_DEET_DE_output - Enrichment of input gene list on DEETs studies.
- DE_correlations - Correlation values of input gene list to DEETs studies (both Pearson and Spearman).

Author(s)

Dustin Sokolowski, Jedid Ahn

References

Paczkowska M, Barenboim J, Sintupisut N, et al. Integrative pathway enrichment analysis of multivariate omics data. Nat Commun. 2020;11(1):735. doi:10.1038/s41467-019-13983-9

Examples

```
data("example_DEET_enrich_input")
data("DEET_example_data")
DEET_out <- DEET_enrich_genesonly(example_DEET_enrich_input, DEET_dataset = DEET_example_data)
```

`DEET_example_data` *DEET_example_data*

Description

Named list of gene-sets and representative metadata for studies associated with Alizada et al., 2021 This example data is the exact same as what is needed to run DEET enrich properly but subsetted to have 13 studies that are enriched by 'example_DEET_enrich_input'. This way, the example gives an output at all levels of enrichment and at the correlation level.

Usage

```
data(DEET_example_data)
```

Format

A named list of seven objects containing the data frames summarizing the DEGs from comparisons within DEET, GMT objects of comparisons within DEET for enrichment through ActivePathways, GMT objects for basic pathway and TF enrichment, and a dataframe for the metadata of each study.

#'

DEET_DE A list of data frames containing the significant DE genes, mean expression, log2fold-change, and padj from DESeq (padj < 0.05).

DEET_gmt_BP A list of class GMT, which is a list of studies where each study is populated by comparison id (internal DEET identifier), comparison name (interpretable comparison name), and a gene set. In this case the gene-set is the pathways that are enriched within that study.

DEET_gmt_TF A list of class GMT, which is a list of studies where each study is populated by comparison id (internal DEET identifier), comparison name (interpretable comparison name), and a gene set. In this case the gene-set is the TFs that are enriched within that study.

DEET_gmt_DE A list of class GMT, which is a list of studies where each study is populated by comparison id (internal DEET identifier), comparison name (interpretable comparison name), and a gene set. In this case the gene-set is the DEGs that are enriched within that study.

gmt_BP A list of class GMT, which is a list of gene ontology gene-sets acquired from the bader lab 'http://download.baderlab.org/EM_Genesets/'#'

gmt_TF A list of class GMT, which is a list of Transcription Factor gene-sets acquired from the bader lab 'http://download.baderlab.org/EM_Genesets/'

DEET_metadata For every pairwise comparison, the study name, source (SRA, TCGA, GTEx and SRA-manual), description from the DRA compendium, the number of samples (total, up-condition, and down-condition), samples (total ,up-condition, down-condition), tissue (including tumour from TCGA), number of DEs (total, up-condition, down-condition), age (mean +- sd), sex, top 15 DEGs - up, top 15 DEGs - down, top 5 enriched pathways, and top 5 enriched TFs. PMID are also available for studies selected from SRA. Lastly, each pairwise comparison was given an overall category based on those decided in Crow et al., 2019.

Examples

```
data(DEET_example_data)
```

`DEET_feature_extract` *DEET_feature_extract*

Description

Identify which genes are associated with pieces of metadata that a researcher queries.

Usage

```
DEET_feature_extract(mat, response, datatype, detection_cutoff = 0.7)
```

Arguments

`mat`	A gene-by-study matrix populated by the coefficients of that study. By default, the coefficient is the log2Fold-change of genes as long as they are differentially expressed (cutoff = padj < 0.05).
`response`	A vector (binomial, categorical, or continuous) that is used to associated the DEGs within the studies.
`datatype`	indication of whether the response variable is binomial, categorical, or continuous.
`detection_cutoff`	Proportion of studies where the gene is detected (not as DE but detected at all, designated with a FC != 0). Default value 0.7.

Value

Named list given the elastic net coefficients and the eleastic net regression between the response variable and the DEGs within DEET. It also outputs the correlation, ANOVA, and wilcoxon test of every gene against the response variable based on if it's continuous, categorical, or binomial in nature.

- elastic_net_coefficients - Association that a gene has with the response variable based on the elastic net regression.

- elastic_net - Output of the elastic net regression
- - basic_features gives the output of the correlation, ANOVA, and wilcoxon test of every gene against the response variable.

Author(s)

Dustin Sokolowski, Jedid Ahn

References

Engebretsen, S., & Bohlin, J. (2019). Statistical predictions with glmnet. Clinical epigenetics, 11(1), 1-3.

Examples

```
data(DEET_feature_extract_example_matrix)
data(DEET_feature_extract_example_response)
single1 <- DEET_feature_extract(DEET_feature_extract_example_matrix,
DEET_feature_extract_example_response,"categorical")
```

`DEET_feature_extract_example_matrix`

DEET_feature_extract_example_matrix

Description

An object of class data.frame where rows are genes and columns are comparisons. The matrix is populated by the log2Fold-change of each gene within each study. If the gene is not detected within that study, it is populated with 0 instead of the log2Fold-change. This object is inputted into the 'mat' input variable for the 'DEET_feature_extract' function. This example takes 1000 random genes and 200 random studies (seed = 1234s).

Usage

```
data(DEET_feature_extract_example_matrix)
```

Format

An object of class data.frame where rows are genes and columns are comparisons (1000 randomly selected genes and 200 randomly selected studies).

Examples

```
data(DEET_feature_extract_example_matrix)
```

DEET_feature_extract_example_response

DEET_feature_extract_example_response

Description

Character vector giving the source (TCGA SRA, GTEx, SRA-manual) of 200 comparisons within DEET. Used as the input for the ‘response‘ input of ‘DEET_feature_extract‘ in the example. For this response variable to work, the ‘datatype‘ input variable would also need to be set to "categorical".

Usage

```
data(DEET_feature_extract_example_response)
```

Format

Character vector giving the source (TCGA SRA, GTEx, SRA-manual) of 200 comparisons within DEET.

Examples

```
data(DEET_feature_extract_example_response)
```

DEET_Input_as_Reference

DEET_Input_as_Reference

Description

Alternative function to DEET enrich for when the inputted gene list is unordered. Here, we can increase the statistical rigour of enrichment by levaraging the p-values of the DEGs within DEET. Specifically, the inputted DE list is used as the reference and we test each DE list against your reference. Specifically. We convert your reference into a gmt file before inputting each pairwise DE list into ActivePathways. This function does not complete correlations or pathway-level analysis.

Usage

```
DEET_Input_as_Reference(genes, DEET_dataset, background = NULL)
```

Arguments

genes	A character vector of gene symbols within 'DEET_dataset'
DEET_dataset	The databank of the differential expression enrichment tool. Appropriate inputs here are "DEET_example_data" stored within DEET, the "DEET_combined.rda" file from the DEET stable repositoy found at X, and the DEET database developmental repository found at Y. The DEET_dataset is a named list where details of it's structure can be found ?DEET_example_data. Unlike in DEET_enrich, this dataset does not require the pathway-relevant elements of the DEET_dataset list, namely "gmt_BP", or "gmt_TF" "DEET_gmt_BP", "DEET_gmt_TF". It also does not need DEET_gmt_DE.
background	Character vector of human gene symbols showing all possible genes. Default value is NULL and the background is generated as all detected DEGs across any comparison.

Value

Named list containing the ActivePathways enrichment of each comparison on the user's inputted gene list, as well as the associated metadata of each enriched comparison.

Author(s)

Dustin Sokolowski

References

Paczkowska M, Barenboim J, Sintupisut N, et al. Integrative pathway enrichment analysis of multivariate omics data. Nat Commun. 2020;11(1):735. doi:10.1038/s41467-019-13983-9

Examples

```
data("example_DEET_enrich_input")
genes <- rownames(example_DEET_enrich_input)
data("DEET_example_data")
DEET_out_ref <- DEET_Input_as_Reference(genes, DEET_dataset = DEET_example_data)
```

DEET_plot_correlation *DEET_plot_correlation*

Description

Take significant correlation outputs and generate scatterplots of the genes DE in one or the other.

Usage

```
DEET_plot_correlation(correlation_input)
```

Arguments

correlation_input
: The "DE_correlations" element of the output of the DEET_enrich function. This function only works if there is at least one significantly correlated study.

Value

Named list of ggplot objects with the correlation between the input study and the study within DEET

Author(s)

Dustin Sokolowski, Jedid Ahn

Examples

```
data("example_DEET_enrich_input")
data("DEET_example_data")
DEET_out <- DEET_enrich(example_DEET_enrich_input, DEET_dataset = DEET_example_data)
correlation_input <- DEET_out$DE_correlations
correlation_plots <- DEET_plot_correlation(correlation_input)
```

---

example_DEET_enrich_input

example_DEET_enrich_input

Description

Exon-level DEGs of HAoEC after TNFa treatment for 45 mins from Alizada et al., 2021. Object is a data.frame with columns "gene_symbol" "padj" and "coef", which in this case is the log2Fold-change of differential expression.

Usage

```
data(example_DEET_enrich_input)
```

Format

A data frame with three columns. Rows are genes and it's populated by the gene symbol, padj of gene expression, and coef (log2Fold-change).

Examples

```
data(example_DEET_enrich_input)
```

proccess_and_plot_DEET_enrich
proccess_and_plot_DEET_enrich

Description

Generates barplots and dotplots based on the output of the DEET_enrich function.

Usage

```
proccess_and_plot_DEET_enrich(
  DEET_output,
  colour_barplot = "Source",
  width = 8,
  text_angle = 0,
  horizontal = F,
  topn = 5,
  ol_size = 1,
  exclude_domain = "",
  cluster_order = NULL,
  colors = "Set2"
)
```

Arguments

DEET_output	Direct output of the DEET_enrich function. A list with all of the same names as DEET_output.
colour_barplot	Pick dotplot or barplot colours. It can be NULL, in which all bars are the same or it can be a (case sensitive) column within the metadata. Defaults to "source".
width	The number of inches in the barplot or dotplot.
text_angle	The angle of the enriched studies.
horizontal	Whether the output barplot is vertical or horizontal
topn	the top number of studies (by p-value) to be plotted.
ol_size	the minimum number of overlapping genes (or paths) in an enriched study.
exclude_domain	Exclude studies enriched based on DEGs, Paths, or TF if the user happened to aggregate the results into a single DF, generally unused.
cluster_order	Factor to group studies based on the researchers custom annotation.
colors	Type of color pallete to input into 'scale_fill_brewer' of ggplot.

Value

Named list where each element is a ggplot object plotting the output of the enrichment tests within DEET. The final element is the output of ActivePathways (in DEET) that is directly compatible with the DEET_enrichment_barplot function.

- DEET_DotPlot - ggplot object of Dotplot of enrichment of enrichment of DEET studies based on DE, BP, and TF information. Only plotted if 2/3 levels contain at least one significant study.
- Pathway_barplot - ggplot object of Barplot of standard gene set enrichment based on gene ontology and TFs. Only plotted if there is at least one enriched significant pathway/TF.
- individual_barplot - ggplot object of Barplot of the top enriched pathways or studies (depending on the input list).Barplot is only generated if each list has at least one pathway (or study) is enriched.
- DEET_output_forplotting - output of Activepathways with "domain", "overlap.size", and "p.value" columns added to be compatible with the DEET_enrichment_barplot function.

Author(s)

Dustin Sokolowski, Hauyun Hou PhD

Examples

```
data("example_DEET_enrich_input")
data("DEET_example_data")
DEET_out <- DEET_enrich(example_DEET_enrich_input, DEET_dataset = DEET_example_data)
plotting_example <- proccess_and_plot_DEET_enrich(DEET_out, text_angle = 45,
horizontal = TRUE, topn=4)
```

Index