Package 'DAGItoolsr'

September 14, 2022

Type Package

Title Disease-Associated Gene Identification Tools

Version 0.0.1.0

Description Implements methods to identify and analyse disease-associated genes.
Several options are available: (1) gene search from databases, and (2) gene search from differential expression data.
It also provides options to predict protein-protein interaction (via STRINGdb) and perform functional enrichment (via clusterProfiler).

RoxygenNote 7.2.1

Depends R (>= 4.2.0)

biocViews

Imports pheatmap (>= 1.0.12),
ggplot2 (>= 3.3.6),
dplyr (>= 1.0.10),
rentrez (>= 1.2.3),
stringr (>= 1.4.1),
ggrepel (>= 0.9.1),
GEOquery (>= 2.64.2),
limma (>= 3.52.2),
AnnotationDbi (>= 1.58.0),
org.Hs.eg.db (>= 3.15.0),
clusterProfiler (>= 4.4.4),
enrichplot (>= 1.16.2),
pathview (>= 1.36.1),
DOSE (>= 3.22.1),
STRINGdb (>= 2.8.4),
igraph (>= 1.3.4),
knitr (>= 1.40),
rvest (>= 1.0.3),
tidyr (>= 1.2.0),
Biobase (>= 2.56.0),
magrittr (>= 2.0.3),
utils (>= 4.2.1)

Remotes github::ramiromagno/gwasrapidd,
github::cran/SPARQL,
bitbucket::ibi_group/disgenet2r

License GPL (>=3)

URL https://github.com/NurSyatila/DAGItoolsr

Encoding UTF-8

R topics documented:

---

analyse_deg   *Analyse Gene Expression Data (GEO Series)*

Description

Perform differential expression analysis and retrieve differentially expressed protein-coding genes from GEO Series

Usage

```
analyse_deg(queryTerms, GSEaccession, GSEplatform)
```

Arguments

queryTerms	A vector of disease name and/or keywords
GSEaccession	GSE Accession
GSEplatform	GSE Platform

Value

A dataframe of differentially expressed protein-coding genes

A CSV-formmated file containing results of differentially expressed protein-coding genes (FilteredDE.csv)

A text file containing Entrez identifiers of differentially expressed protein-coding genes (DEGenes.txt)

A PDF-formatted file containing visualization plots (Plots.pdf)

A text file containing top 20 genes in tabulated form (Top20DEGenes.txt)

Examples

```
queryTerms <- c("endometriosis");
GSEaccession <- "GSE23339";
GSEplatform <- "GPL6102";
results <- analyse_deg(queryTerms,GSEaccession,GSEplatform);
```

---

analyse_ppi_network   *Analyse PPI Network*

Description

A workflow to get predict protein-protein interaction, identify gene modules and perform functional enrichment for a given gene list

Usage

```
analyse_ppi_network(geneList)
```

Arguments

geneList   A vector of Entrez gene identifiers

Value

A directory named PPI

PPI: A summary of PPI analysis (PPISummary.txt)

PPI: A PDF file containing visualization plots from PPI analysis (STRING_PPI_Network.pdf)

PPI: A list of filtered genes from PPI where isolated nodes are discarded (FilteredGenes.txt)

PPI: A list of gene clusters with details on enriched KEGG Pathway (GeneCluster.txt)

A directory named PPI/FilteredGenes containing functional enrichment results for filtered genes

A directory named PPI/TopGenes containing KEGG enrichment results for top 10 or 20 genes

Examples

```
gene_file <- 'DatabaseGenes.txt';
geneList <- readLines(gene_file)
results <- analyse_ppi_network(geneList);
```

---

differentialexpression_single_workflow
Single Differential Expression Worflow

Description

Get differentially expressed genes from gene expression data associated to a disease, perform PPI and functional enrichment analyses for a single GEO Series

Usage

```
differentialexpression_single_workflow(queryTerms, GSEaccession, GSEplatform)
```

Arguments

queryTerms	A vector of disease name and/or keywords to be used for sample classification
GSEaccession	GEO Series (GSE) accession
GSEplatform	GEO Series (GSE) platform

Value

GeneExpression: A CSV-formmated file containing results of differentially expressed protein-coding genes (FilteredDE.csv) for individual GEO Series in a separate directory

GeneExpression: A text file containing Entrez identifiers of differentially expressed protein-coding genes (DEGenes.txt) for individual GEO Series in a separate directory

GeneExpression: A PDF-formatted file containing visualization plots (Plots.pdf) for individual GEO Series in a separate directory

GeneExpression: A text file containing top 20 genes in tabulated form (Top20DEGenes.txt) for individual GEO Series in a separate directory

PPI: A summary of PPI analysis (PPISummary.txt)

PPI: A list of filtered genes from PPI where isolated nodes are discarded (FilteredGenes.txt)

PPI: Top genes from PPI with highest interaction (Top_PPI_genes.txt)

PPI: A list of gene clusters with details on enriched KEGG Pathway (GeneCluster.txt)

PPI: A PDF file containing visualization plots from PPI analysis (STRING_PPI_Network.pdf)

Functional Enrichment: A separate directory of functional enrichment results for filtered genes (FilteredGenes/) and top genes (TopGenes) containing visualization plots (.pdf) and enrichment results (.csv)

Examples

```
queryTerms <- c("endometriosis");
GSEaccession <- "GSE23339";
GSEplatform <- "GPL6102";
results <- differentialexpression_single_workflow(queryTerms,GSEaccession,GSEplatform);
```

---

differentialexpression_workflow
Differential Expression Worflow

Description

Get differentially expressed genes from individual gene expression data (GEO series) associated to a disease, perform PPI and functional enrichment analyses

Usage

```
differentialexpression_workflow(queryTerms)
```

Arguments

queryTerms   A vector of disease name and/or keywords

Value

A directory named DifferentialExpression

GeneExpression: A CSV-formmated file containing GEO Datasets associated to a disease (GSESummary.csv)

GeneExpression: A CSV-formmated file containing GEO Datasets associated to a disease with 'diseased state' condition (DiseaseGSESummary.csv)

Individual directories for each GEO Series matching the query terms containing result files

GeneExpression: A CSV-formmated file containing results of differentially expressed protein-coding genes (FilteredDE.csv) for individual GEO Series in a separate directory

GeneExpression: A text file containing Entrez identifiers of differentially expressed protein-coding genes (DEGenes.txt) for individual GEO Series in a separate directory

GeneExpression: A PDF-formatted file containing visualization plots (Plots.pdf) for individual GEO Series in a separate directory

GeneExpression: A text file containing top 20 genes in tabulated form (Top20DEGenes.txt) for individual GEO Series in a separate directory

A directory named PPI

PPI: A summary of PPI analysis (PPISummary.txt)

PPI: A list of filtered genes from PPI where isolated nodes are discarded (FilteredGenes.txt)

PPI: Top genes from PPI with highest interaction (Top_PPI_genes.txt)

PPI: A list of gene clusters with details on enriched KEGG Pathway (GeneCluster.txt)

PPI: A PDF file containing visualization plots from PPI analysis (STRING_PPI_Network.pdf)

Functional Enrichment: A separate directory of functional enrichment results for filtered genes (FilteredGenes/) and top genes (TopGenes) containing visualization plots (.pdf) and enrichment results (.csv)

Examples

```
queryTerms <- c("endometriosis");
results <- differentialexpression_workflow(queryTerms);
```

---

genesearch_workflow *Database Gene Search Worflow*

Description

Get disease-associated genes (Entrez Gene identifiers) from available public resources

Usage

```
genesearch_workflow(
  queryTerms,
  disgenet = FALSE,
  email = NULL,
  password = NULL
)
```

Arguments

queryTerms	A vector of disease name and/or keywords
disgenet	An option to include results from DisGeNET. Note that This option will require email and password to DisGeNET account
email	Registered email to DisGeNET account
password	Password to DisGeNET account

Value

A directory named GeneSearch

Gene Search: CSV-formmated files containing results from individual database search

Gene Search: Text-formatted files containing a list of genes from individual database search ([Database]_Genes.txt)

Gene Search: Text-formatted files containing a list of protein-coding genes from individual database search ([Database]_PCGenes.txt)

Gene Search: A text-formatted file containing a list of protein-coding genes from all database searches (DatabaseGenes.txt)

PPI: A summary of PPI analysis (PPISummary.txt)

PPI: A list of filtered genes from PPI where isolated nodes are discarded (FilteredGenes.txt)

PPI: Top genes from PPI with highest interaction (Top_PPI_genes.txt)

PPI: A list of gene clusters with details on enriched KEGG Pathway (GeneCluster.txt)

PPI: A PDF file containing visualization plots from PPI analysis (STRING_PPI_Network.pdf)

Functional Enrichment: A separate directory of functional enrichment results

Examples

```
queryTerms <- c("endometriosis");
results <- genesearch_workflow(queryTerms,disgenet=FALSE);
results <- genesearch_workflow(queryTerms,disgenet=TRUE,"you@gmail.com","yourpassword");
```

---

get_clinvar *Gene search in CLinVar*

Description

Get disease-associated genes (Entrez Gene identifiers) from ClinVar (via rentrez)

Usage

```
get_clinvar(queryTerms)
```

Arguments

queryTerms   A vector of disease name and/or keywords

Value

A list of protein-coding genes

A CSV-formmated file containing results from ClinVar

A text-formatted file containing a list of genes (ClinVar_Genes.txt)

A text-formatted file containing a list of protein-coding genes (ClinVar_PCGenes.txt)

Examples

```
queryTerms <- c("endometriosis");
results <- get_clinvar(queryTerms);
```

---

get_diseasesdb *Gene search in DISEASES database*

Description

Get disease-associated genes (Entrez Gene identifiers) from DISEASES database

Usage

```
get_diseasesdb(queryTerms)
```

Arguments

queryTerms   A vector of disease name and/or keywords

Value

A list of protein-coding genes

A CSV-formmated file containing results from DISEASES database

A text-formatted file containing a list of genes (DISEASES_Genes.txt)

A text-formatted file containing a list of protein-coding genes (DISEASES_PCGenes.txt)

Examples

```
queryTerms <- c("endometriosis");
results <- get_diseasesdb(queryTerms);
```

---

get_disgenet   *Gene search in DisGeNET*

Description

Get disease-associated genes (Entrez Gene identifiers) from DisGeNET (via DisGeNET2r)

Usage

```
get_disgenet(queryTerms, userEmail, userPassword)
```

Arguments

queryTerms	A vector of disease name and/or keywords
userEmail	Registered email to DisGeNET account
userPassword	Password to DisGeNET account

Value

A list of protein-coding genes

A CSV-formmated file containing results from DisGeNET

A text-formatted file containing a list of genes (DisGeNET_Genes.txt)

A text-formatted file containing a list of protein-coding genes (DisGeNET_PCGenes.txt)

Examples

```
queryTerms <- c("endometriosis");
userEmail <- "you@gmail.com";
userPassword <- "yourpassword";
results <- get_disgenet(queryTerms,userEmail,userPassword);
```

---

get_enrichment_results
                        Get enrichment results

Description

Perform enrichment analysis for a gene list (over-representation analysis) or a dataframe consisting gene identifiers and log value (gene set enrichment analysis)

Usage

```
get_enrichment_results(geneList, ORAterm, enrichmenttype)
```

Arguments

geneList	A vector / dataframe of Entrez gene identifiers
ORAterm	Functional term, BP, MF, CC, KEGG or DO
enrichmenttype	ORA (over-representation analysis) or GSE (gene set enrichment analysis)

Value

A CSV-formatted file of enrichment results

A PDF file containing visualization plots from enrichment analysis

Examples

```
#gene_file <- 'DatabaseGenes.txt';
#geneList <- readLines(gene_file)
geneList <- c("1588","3586","5241","7124", "998", "54361", "55591", "283455", "3480","8390","83608","9687","4
results <- get_enrichment_results(geneList,"KEGG","ORA");
```

---

get_gene_modules   *Get gene modules*

Description

Get gene modules (clusters of genes according to interactions) and analyse individual gene clusters/modules (required by analyse_ppi_network())

Usage

```
get_gene_modules(geneNetwork, summaryFile)
```

Arguments

geneNetwork	An igraph object
summaryFile	A file name to store details of the clusters

Value

PPI: A list of gene clusters with details on enriched KEGG Pathway (SummaryFile)

Examples

```
geneList <- readLines(gene_file)
genelist <- as.data.frame(geneList)
mapped_genes <- suppressWarnings(string_db$map(genelist, "geneList", removeUnmappedRows = TRUE, quiet = TRUE )
string_db <- STRINGdb$new( version="11.5", species=9606, score_threshold=700, input_directory=".")
gene_network <- suppressMessages(string_db$get_subnetwork(mapped_genes$STRING_id))
gene_network2 <- igraph::delete.vertices(gene_network, degree(gene_network)==0)
gene_network2 <- igraph::simplify(gene_network2, remove.multiple=TRUE, remove.loops=TRUE)
results <- get_gene_modules(gene_network2)
```

---

get_gse_datasets   *Get GSE datasets associated to a disease*

Description

Retrieve GSE Datasets associated to a disease based on MeSH Terms

Usage

```
get_gse_datasets(queryTerms)
```

Arguments

queryTerms   A vector of disease name and/or keywords

Value

A dataframe of results (GEO Dataset with 'diseased state' condition) from GEO Dataset web search through Rentrez

A CSV-formmated file containing GEO Datasets associated to a disease (GSESummary.csv)

A CSV-formmated file containing GEO Datasets associated to a disease with 'diseased state' condition (DiseaseGSESummary.csv)

Examples

```
queryTerms <- c("endometriosis");
results <- get_gse_datasets(queryTerms);
```

---

get_gtr   *Gene search in GTR*

Description

Get disease-associated genes (Entrez Gene identifiers) from GTR (via rentrez)

Usage

```
get_gtr(queryTerms)
```

Arguments

queryTerms   A vector of disease name and/or keywords

Value

A list of protein-coding genes

A CSV-formmated file containing results from GTR database

A text-formatted file containing a list of genes (GTR_Genes.txt)

A text-formatted file containing a list of protein-coding genes (GTR_PCGenes.txt)

Examples

```
queryTerms <- c("endometriosis");
results <- get_gtr(queryTerms);
```

---

get_gwascatalog *Gene search in GWAS Catalog*

Description

Get disease-associated genes (Entrez Gene identifiers) from GWAS Catalog (via gwasrapidd)

Usage

```
get_gwascatalog(queryTerms)
```

Arguments

queryTerms   A vector of disease name and/or keywords

Value

A list of protein-coding genes

A CSV-formmated file containing results from GWAS Catalog database

A text-formatted file containing a list of genes (GWASCatalog_Genes.txt)

A text-formatted file containing a list of protein-coding genes (GWASCatalog_PCGenes.txt)

Examples

```
queryTerms <- c("endometriosis");
results <- get_gwascatalog(queryTerms);
```

---

get_medgen *Gene search in MedGen*

Description

Get disease-associated genes (Entrez Gene identifiers) from MedGen (via Rentrez)

Usage

```
get_medgen(queryTerms)
```

Arguments

queryTerms   A vector of disease name and/or keywords

Value

A list of protein-coding genes

A CSV-formmated file containing results from MedGen

A text-formatted file containing a list of genes (MedGen_Genes.txt)

A text-formatted file containing a list of protein-coding genes (MedGen_PCGenes.txt)

Examples

```
queryTerms <- c("endometriosis");
results <- get_medgen(queryTerms);
```

---

get_meshterms   *Get MeSH terms or identifiers*

Description

Get MeSH terms or identifiers associated to a disease (via rentrez)

Usage

```
get_meshterms(queryTerms, meshtype)
```

Arguments

queryTerms   A vector of disease name and/or keywords

meshtype   type of MeSH output, "mesh_id" or "mesh_term"

Value

A list of MeSH terms or identifers

Examples

```
queryTerms <- c("endometriosis");
results <- get_meshterms(queryTerms,"mesh_id");
results <- get_meshterms(queryTerms,"mesh_term");
```

---

get_omim *Gene search in OMIM*

Description

Get disease-associated genes (Entrez Gene identifiers) from OMIM (via rentrez)

Usage

```
get_omim(queryTerms)
```

Arguments

queryTerms   A vector of disease name and/or keywords

Value

A list of protein-coding genes

A CSV-formmated file containing results from OMIM

A text-formatted file containing a list of genes (OMIM_Genes.txt)

A text-formatted file containing a list of protein-coding genes (OMIM_PCGenes.txt)

Examples

```
queryTerms <- c("endometriosis");
results <- get_omim(queryTerms);
```

---

get_PCGenes *Get protein coding genes*

Description

Get protein coding genes from a list of genes

Usage

```
get_PCGenes(geneList, gene_file, pcgene_file, key_type)
```

Arguments

geneList   A vector of genes (identifiers/symbols)

gene_file   A text-formatted file name to store the list of genes in tabular form (ENTREZID, SYMBOL, GENETYPE)

pcgene_file   A text-formatted file name to store the list of protein-coding genes (Entrez identifiers)

key_type   A text-formatted file name to store the list of protein-coding genes (Entrez identifiers)

Value

Type of geneList, e.g. "ENSEMBL","ENTREZID","SYMBOL","UNIPROT" etc. Refer to key-types(org.Hs.eg.db)

Examples

```
geneList <- c("1588","3586","5241","100048912","10151");
gene_file <- "Genes.txt"
pcgene_file <- "PCGenes.txt"
results <- get_PCGenes(geneList,gene_file,pcgene_file,"ENTREZID");
```

---

get_phegeni *Gene search in PheGenI*

Description

Get disease-associated genes (Entrez Gene identifiers) from PheGenI

Usage

```
get_phegeni(queryTerms)
```

Arguments

queryTerms   A vector of disease name and/or keywords

Value

A list of protein-coding genes

A CSV-formmated file containing results from PheGenI

A text-formatted file containing a list of genes (PheGenI_Genes.txt)

A text-formatted file containing a list of protein-coding genes (PheGenI_PCGenes.txt)

Examples

```
queryTerms <- c("endometriosis");
results <- get_phegeni(queryTerms);
```

get_samplegroupings   *Assign sample groupings for a GEO Series)*

Description

Assign sample groupings for a GEO Series through automatic classification (required by analyse_deg function)

Usage

```
get_samplegroupings(queryTerms, GSEaccession, GSEplatform)
```

Arguments

queryTerms	A vector of disease name and/or keywords
GSEaccession	GSE Accession
GSEplatform	GSE Platform

Value

A vector of GEOquery results and sample groupings e.g. c(0,0,0,0,1,1,1,1)

Examples

```
queryTerms <- c("endometriosis");
GSEaccession <- "GSE23339";
GSEplatform <- "GPL6102";
results <- get_samplegroupings(queryTerms,GSEaccession,GSEplatform);
```

Index