Package 'VarfromPDB'

October 12, 2022

Type Package

Title Disease-Gene-Variant Relations Mining from the Public Databases and Literature

Version 2.2.10

Depends R (>= 3.4.0), XML, XML2R, curl, stringr

Imports stringi, RISmed, utils

Suggests tools,knitr,rmarkdown

LazyData true

biocViews Software

Date 2018-9-7

Author Zongfu Cao <caozongfu@gmail.com>; Lei Wang <isan.wong@gmail.com>

Maintainer Zongfu Cao <caozongfu@gmail.com>

Description Captures and compiles the genes and variants related to a disease, a phenotype or a clinical feature from the public databases including HPO (Human Phenotype Ontology, <http://human-phenotype-ontology.github.io/about.html>), Orphanet <http://www.orpha.net/consor/cgi-bin/index.php>, OMIM (Online Mendelian Inheritance in Man, <http://www.omim.org>), ClinVar <http://www.ncbi.nlm.nih.gov/clinvar>, and UniProt (Universal Protein Resource, <http://www.uniprot.org>) and PubMed abstracts. HPO provides a standardized vocabulary of phenotypic abnormalities encountered in human disease. HPO currently contains approximately 11,000 terms and over 115,000 annotations to hereditary diseases. Orphanet is the reference portal for information on rare diseases and orphan drugs, whose aim is to help improve the diagnosis, care and treatment of patients with rare diseases. OMIM is a continuously updated catalog of human genes and genetic disorders and traits, with particular focus on the molecular relationship between genetic variation and phenotypic expression. ClinVar is a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence. UniProt focuses on amino acid altering variants imported from Ensembl Variation databases. For Homo sapiens, the variants including human polymorphisms and disease mutations in the UniProt are manually curated from UniProtKB/Swiss-Prot. Additionally, PubMed provides the primary and latest source of the information. Text mining was employed to capture the information from PubMed abstracts.

License GPL-2

NeedsCompilation no

VignetteBuilder knitr

Repository CRAN

Date/Publication 2018-09-13 07:30:03 UTC

R topics documented:

`aa` *a table of Codon-Amino Acid Abbreviations*

Description

the table list the relation of the codon and amino acid, in cluding full Name, 3-letter Abbreviation and 1-letter Abbreviation.

`extract_clinvar` *Extract the genes and variants related to a genetic disorder from ClinVar*

Description

extract_clinvar extracts the genes and variants associated to a known genetic disorder or a clinical feature from NCBI ClinVar database. It annotates the phenotypes from GeneReview, MedGen, and OMIM. The alias of a disease/phenotye are considered in HPO database. Furtherly, the variants on a use-defined gene list can be captured at the same time.

Usage

```
extract_clinvar(keyword, localPDB.path = paste(getwd(),"localPDB",sep="/"),
           type = "both", HPO.disease = NULL, genelist = NULL, OMIM = NULL)
```

Arguments

keyword	character string: keyword, to describe a disease, clinical feature, or phenotype.
localPDB.path	the path of localized public data bases. The default value is set in the working directory.
type	the type of the information to extract, must be one of "gene", "variant", "both"(default).
HPO.disease	MIM number of the disease. The default value is NULL, which means that all the MIM number of the disease in HPO are added.
genelist	the gene(s) associated to the disease, or the genes you are interested.
OMIM	whether use the information from OMIM database. The default value is NULL. It can be set 'yes' when you make sue you have a OMIM API key.

Details

The function extracts the genes and variants associated to a disease, clinical feature or phenotype from ClinVar database. The keyword is searched not only in ClinVar, but also in HPO to considered the different alias of a diesease. You can prepare the files from OMIM, ClinVar, Orhanet, Uniprot, HPO, MedGen, and GeneReview using *localPDB()* before you start the job, which maybe more efficient. More details ablout ClinVar can be seen from http://www.ncbi.nlm.nih.gov/clinvar/.

Value

A list containing two components:

gene2dis	subset of the file gene_condition_source_id, which include all the information about genes and phenotypes in ClinVar.
variants	subset of the file variant_summary.txt, but added sevetal colomns which describe the phenotype from GeneReview, MedGen, and OMIM databases.

Author(s)

Zongfu Cao (caozongfu@nrifp.org.cn)

References

1.Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014 Jan 1;42(1):D980-5. doi: 10.1093/nar/gkt1113. PubMed PMID: 24234437

2.Melissa Landrum, PhD, Jennifer Lee, PhD, George Riley, PhD, Wonhee Jang, PhD, Wendy Rubinstein, MD, PhD, Deanna Church, PhD, and Donna Maglott, PhD. ClinVar. http://www.ncbi.nlm.nih.gov/books/NBK17458

3.Sebastian K?hler, Sandra C Doelken, Christopher J. Mungall, Sebastian Bauer, Helen V. Firth, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data Nucl. Acids Res. (1 January 2014) 42 (D1): D966-D974 doi:10.1093/nar/gkt1026

See Also

pheno_extract_HPO, extract_omim, extract_uniprot, extract_genes_orphanet

Examples

```
#    library(VarfromPDB)
#    clinvar.phenotype = extract_clinvar(keyword="retinoblastoma")
#    genes.clinvar = clinvar.phenotype[[1]]
#    print(dim(genes.clinvar))
#    variants.clinvar = clinvar.phenotype[[2]]
#    print(dim(variants.clinvar))
```

`extract_genes_orphanet` *Extract the genes related to a genetic disorder from Orphanet*

Description

extract_genes_orphanet extracts the genes associated to a known genetic disorder or a clinical feature from Orphanet database. The alias of a disease/phenotye are considered based on HPO database and then capture the information in Orphanet.

Usage

```
extract_genes_orphanet(keyword,
          localPDB = paste(getwd(),"localPDB",sep="/"), HPO.disease = NULL)
```

Arguments

`keyword`	character string: keyword, to search the disease, clinical feature, or phenotype.
`localPDB`	the path of localized public data bases. The default value is set in the working directory.
`HPO.disease`	Orpha Number of the disease. The default value is NULL, which means that all the Orpha Numbers of the disease in HPO are added.

Details

The function extracts the genes associated to a genetic disease especial rare disease, or a clinical feature or phenotype from Orphanet database. The keyword is searched not only in Ophanet, but also in HPO considering the alias of the diesease. More details about Ophanet can be seen in http://www.orpha.net/consor/cgi-bin/index.php.

Value

a matrix will be returned including

1.OrphaNumber

2.Phenotype

3.GeneSymbol

4.GeneName

5.GeneType

6.AssociationType

7.AssociationStatus

Author(s)

Zongfu Cao (caozongfu@nrifp.org.cn)

References

1.Orphanet: an online rare disease and orphan drug data base. INSERM 1997. Available on http://www.orpha.net. Accessed [date accessed].

2.Orphadata: Free access data from Orphanet. INSERM 1997. Available on http://www.orphadata.org. Data version [XML]

See Also

pheno_extract_HPO, extract_omim, extract_uniprot, extract_clinvar

Examples

```
    ## extract genes from Orphanet
#    orphanet.phenotype = extract_genes_orphanet(keyword = "retinoblastoma")
```

`extract_omim` *Extract the genes and variants related to a genetic disorder from OMIM*

Description

extract_omim extracts the genes and variants related to a known genetic disorder or a clinical feature from NCBI OMIM database.The alias of a disease/phenotype are caputured from HPO database and searched in OMIM. Furtherly, the variants on a use-defined gene list can be captured meanwhile.

Usage

```
extract_omim(keyword, omim.apiKey,
             localPDB.path = paste(getwd(),"localPDB",sep="/"),
             type = "both", HPO.disease = NULL, genelist = NULL)
```

Arguments

keyword	character string: keyword, to search the disease, clinical feature, or phenotype.
omim.apiKey	the API key of OMIM.
localPDB.path	the path of localized public data bases. The default value is set in the working directory.
type	the type of the information to extract, must be one of "gene", "variant", "both"(default).
HPO.disease	MIM number of the disease. The default value is NULL, which means that all the MIM number of the disease in HPO are added.
genelist	the gene(s) related to the disease, or the genes you are interested.

Details

extract_omim extracts the genes from OMIM first, and then translate to appoved gene symbol by HGNC. Then the variants are captured for each gene from OMIM API. However, you should apply for an account and an API key from OMIM.

We recommend to make the files ready locally before a job, in order to avoid a possible failure by the bad network environment.

Value

A list containing two components:

morbidmap	the subset of the file *morbidmap*, which include all the information about genes and phenotypes in OMIM.
mutations	all the mutations in the genes in OMIM.

Author(s)

Zongfu Cao (caozongfu@nrifp.org.cn)

References

1.OMIM:http://www.omim.org/

2.Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015 Jan;43(Database issue):D789-98. doi: 10.1093/nar/gku1205. Epub 2014 Nov 26. PubMed PMID: 25428349; PubMed Central PMCID: PMC4383985.

See Also

pheno_extract_HPO, extract_uniprot, extract_genes_orphanet, extract_clinvar

`extract_pubmed` *Extract the genes and variants related to a genetic disorder from PubMed*

Description

extract_pubmed extracts the genes and variants related to a known genetic disorder or a clinical feature from NCBI PubMed.

Usage

```
extract_pubmed(query, keyword, localPDB.path = paste(getwd(), "localPDB",
              sep = "/"))
```

Arguments

`query`	searching strategy in PubMed, such as "pubmed AND gene AND mutation AND chinese NOT meta analysis".
`keyword`	character string: keyword, to search the disease, clinical feature, or phenotype.
`localPDB.path`	the path of localized public data bases. The default value is set in the working directory.

Details

extract_pubmed extracts the phenotypes, genes and mutations from PubMed abstracts, and check the gene names to appoved symbol by HGNC. We recommend to check the searching strategy and the results carefully.

Value

A list containing two components:

`pubmed_captures`	the relationships among phenotypes, genes, and mutations captured from PubMed
`abstracts`	all the abstracts captured from PubMed.

Author(s)

Zongfu Cao (caozongfu@nrifp.org.cn)

Examples

```
## capture the genes and mutations related to cataract in Chinese populations
## from PubMed
#cataract_pubmed <- extract_pubmed(
#    query = "cataract AND gene AND mutation AND chinese NOT meta analysis",
#               keyword="cataract",
#               localPDB="/public/home/czf/project/rare.disease/localPDB")
```

`extract_uniprot` *Extract the genes and variants related to a genetic disorder from UniprotKB*

Description

extract_uniprot extracts the genes and variants associated to a known genetic disorder or a clinical feature from the UniProt Knowledgebase (UniprotKB). The alias of a disease/phenotye are caputured from HPO database. Furtherly, the gene mutations on a gene list can also be captured at the same time.

Usage

```
extract_uniprot(keyword, localPDB.path = paste(getwd(),"localPDB",sep="/"),
                HPO.disease = NULL, genelist = NULL)
```

Arguments

`keyword`	character string: keyword, to search the disease, clinical feature, or phenotype.
`localPDB.path`	the path of localized public data bases. The default value is set in the working directory.
`HPO.disease`	MIM number of the disease. The default value is NULL, which means that all the MIM number of the disease in HPO are added.
`genelist`	the gene(s) associated to the disease, or the genes you are interested.

Details

extract_uniprot extracts the genes and variants from Uniprot, which focus on amino acid altering variants, and manually curated Human polymorphisms and disease mutations from UniProtKB/Swiss-Prot.

The Uniprot file *humsavar* can be downloaded automatically. However, the speed may depend on the network environment. So, we recommend to make the file ready locally before the jobs using *localPDB()*.

Value

A list containing two components:

`genes.extr`	genes captured from Uniprot.
`dat.extr`	variants captured from Uniprot.

Author(s)

Zongfu Cao (caozongfu@nrifp.org.cn)

References

1.The UniProt Consortium UniProt: a hub for protein information. Nucleic Acids Res. 43: D204-D212 (2015).

See Also

pheno_extract_HPO, extract_omim, extract_genes_orphanet, extract_clinvar

Examples

```
## extract the genes and variants associated to a known mendelian
## disorder from uniprot
#    uniprot.phenotype = extract_uniprot(keyword="retinoblastoma")
```

---

genes_add_pubmed *Compile the disease-related genes from PubMed abstracts into the gene set from the public databases*

Description

To compile the genes related to a disease especially for a rare disease from PubMed abstracts into the gene set from the public databases, including HPO, orphanet, omim, clinvar and uniprot.

Usage

```
genes_add_pubmed(keyword,genepdb, pubmed,
             localPDB.path = paste(getwd(), "localPDB", sep = "/"))
```

Arguments

keyword	character string: keyword, to search the disease, clinical feature, or phenotype.
genepdb	the object from function *genes_compiled*.
pubmed	the object from function *extract_pubmed*. The object need to be checked manually.
localPDB.path	the path of localized public databases.

Details

The relationships between genes and a phenotype are compared with those from public databases, then the additional relationships can be merged together. For the object from function *extract_pubmed* maybe have noise, we strongly recommend that the additional relationships between genes and phenotypes should be pay more attention and checked mannually.

Value

A matrix containing the following information

GeneSymbol	gene symbols from HGNC.
chr	chorosomes of the genes.
strand	strands of the genes.
start	start positions (hg19) of the genes.
end	end positions (hg19) of the genes.
EntrezGeneID	Entrez GeneID
ApprovedName	Approved gene name from HGNC.
Synonyms	gene Synonyms.
HPO	the phenotypes from HPO.
Orphanet	the phenotypes from orphanet.
OMIM	the phenotypes from OMIM.
ClinVar	the phenotypes from ClinVar.
Uniprot	the phenotypes from Uniprot.
pubmed	the phenotypes from PubMed.

Author(s)

Zongfu Cao (caozongfu@nrifp.org.cn)

See Also

extract_pubmed, genes_compile

Examples

```
## add the relationships between genes and phenotypes into those from
## the public databases.
#geneAll <- genes_add_pubmed(genepdb= genesPDB, pubmed=genes.pubmed)
```

---

genes_compile *Compile the disease-related genes from multiple public databases*

Description

To compile a gene set related to a disease especially for a rare disease from multiple databases, including HPO, orphanet, omim, clinvar and uniprot.

Usage

```
genes_compile(HPO, orphanet, omim, clinvar, uniprot,
       localPDB.path = paste(getwd(),"localPDB",sep="/"))
```

Arguments

HPO	the object from *pheno.extract.HPO* function.
orphanet	the object from *extract.genes.orphanet* function.
omim	the object from *extract.omim* function.The default value is NULL.
clinvar	the object from *extract.clinvar* function.
uniprot	the object from *extract.uniprot* function.
localPDB.path	the path of localized public databases.

Details

The relationships between genes and a phenotype in different databases can be intergrated automatically.

Value

A matrix containing the following information

GeneSymbol	gene symbols from HGNC.
chr	chorosomes of the genes.
strand	strands of the genes.
start	start positions (hg19) of the genes.
end	end positions (hg19) of the genes.
EntrezGeneID	Entrez GeneID
ApprovedName	Approved gene name from HGNC.
Synonyms	gene Synonyms.
HPO	the phenotypes from HPO.
Orphanet	the phenotypes from orphanet.
OMIM	the phenotypes from OMIM.
ClinVar	the phenotypes from ClinVar.
Uniprot	the phenotypes from Uniprot.

Author(s)

Zongfu Cao (caozongfu@nrifp.org.cn)

See Also

pheno_extract_HPO, extract_omim, extract_genes_orphanet, extract_clinvar, extract_uniprot

Examples

```
## compile the gene-disease relationship from multiple databases
#genesPDB <- genes_compile(HPO = HPO.Joubert, orphanet = orphanet.joubert,
#                          omim = genes.omim,
#                          clinvar = genes.clinvar,
#                          uniprot = genes.uniprot)
```

grep_split *Extention for grep function*

Description

grep a string whether in another string or vector, the string are split by space.

Usage

```
grep_split(keyword, x)
```

Arguments

keyword	a character string, separator "|" is permitted.
x	a character vector where matches are sought.

Details

Extention for grep functin.

Value

The function return the numbers vector which contain the keyword.

Author(s)

Zongfu Cao (caozongfu@nrifp.org.cn)

See Also

grep

Examples

```
x <- c("you and he and I", "you", "Tom", "I", "you and I", "he and I")
grep_split("you and I | Tom", x)
#[1] 1 5 3
```

`localPDB` *Localize the public databases including HPO, MedGen, GeneReview, HGNC, Orphanet, ClinVar and Uniprot.*

Description

localPDB downloads the necessary files from the public databases including HPO, MedGen, GeneReview, HGNC, Orphanet, ClinVar and Uniprot.

Usage

```
localPDB(localPDB.path = paste(getwd(), "localPDB", sep = "/"), PDB = "all",
         omim.url = NULL, download.method = "curl_fetch_disk")
```

Arguments

`localPDB.path`	the path to localize the public databases.
`PDB`	which database to localize. The value must be one of "all"(default), "HPO", "MedGen", "GeneReview", "HGNC", "Orphanet", "ClinVar" or "Uniprot".
`omim.url`	the FTP URL of OMIM.
`download.method`	the method for downloading files, including "curl_fetch_disk", "curl_download", "download.file".

Details

The function gets the necessary files from the public databases including HPO, MedGen, GeneReview, HGNC, OMIM, Orphanet, ClinVar and Uniprot.

For the *omim.url*, you shoul apply for an OMIM account from http://omim.org/downloads and get the FTP URL.

Author(s)

Zongfu Cao (caozongfu@nrifp.org.cn)

References

1.Sebastian Kohler, Sandra C Doelken, Christopher J. Mungall, Sebastian Bauer, Helen V. Firth, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data Nucl. Acids Res. (1 January 2014) 42 (D1): D966-D974 doi:10.1093/nar/gkt1026

2.Orphanet: an online rare disease and orphan drug data base. INSERM 1997. Available on http://www.orpha.net. Accessed [date accessed].

3.Orphadata: Free access data from Orphanet. INSERM 1997. Available on http://www.orphadata.org. Data version [XML]

4.Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014 Jan 1;42(1):D980-5. doi: 10.1093/nar/gkt1113. PubMed PMID: 24234437

5.The UniProt Consortium UniProt: a hub for protein information. Nucleic Acids Res. 43: D204-D212 (2015).

6.Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015 Jan;43(Database issue):D789-98. doi:10.1093/nar/gku1205. Epub 2014 Nov 26. PubMed PMID: 25428349; PubMed Central PMCID: PMC4383985.

7.GeneReviews: http://www.ncbi.nlm.nih.gov/books/NBK1116/

8.MedGen: http://www.ncbi.nlm.nih.gov/medgen

9.OMIM:http://www.omim.org/

`pheno_extract_HPO` *Extract the genes related to a disease or disease alias from HPO database.*

Description

Extract the genes associated to a disease or disease alias from the Human Phenotype Ontology (HPO) database. The keyword can also be a clinical feature. All the genes and alias of a disease here can be considered in other databases, including Ophanet, OMIM, ClinVar and Uniprot.

Usage

```
pheno_extract_HPO(keyword, localPDB.path = paste(getwd(),"localPDB",sep="/"))
```

Arguments

`keyword`	character string: keyword, to search a disease, a clinical feature, or a phenotype.
`localPDB.path`	the path of localized public data bases. The default value is set in the working directory.

Details

Many genetic diseases have multiple aliases, and for a clinical feature, there are many different disease names too. All he information can be gotten from HPO. More details about HPO, please see `http://www.human-phenotype-ontology.org/`.

The HPO files include phenotype_annotation.tab and diseases_to_genes, which can be downloaded automatically. However, the speed may depend on the network environment. So, we recommend to make the files ready locally before the jobs using *localPDB()*.

Value

A list contains two complements

HPO	subset of HPO
diseases_to_genes	extract the genes and alias for a disease(phenotype), or a clinical feature.

Author(s)

Zongfu Cao (caozongfu@nrifp.org.cn)

References

1.Sebastian K?hler, Sandra C Doelken, Christopher J. Mungall, Sebastian Bauer, Helen V. Firth, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data Nucl. Acids Res. (1 January 2014) 42 (D1): D966-D974 doi:10.1093/nar/gkt1026

See Also

extract_omim, extract_uniprot, extract_genes_orphanet, extract_clinvar

Examples

```
#extract the phenotypes and genes from HPO
#    HPO.phenotype = pheno_extract_HPO("retinoblastoma")
```

---

variants_compile *Compile the disease-related variants from multiple public databases*

Description

To get a variant set related to a disease especially for a rare disease from multiple database, including omim, clinvar and uniprot.

Usage

```
variants_compile(omim, clinvar, uniprot,
     localPDB.path = paste(getwd(), "localPDB", sep = "/"))
```

Arguments

omim	the object from *extract.omim* function.
clinvar	the object from *extract.clinvar* function.
uniprot	the object from *extract.uniprot* function.
localPDB.path	the path of localized public data bases.

Details

A variant maybe have different names in different databases. All the variants in OMIM and Uniprot are compared with ClinVar, then the additional variants are appended the clinVar-like summary variant set.

Value

A matrix containing all the variant-gene-phenotypes relationships.

Author(s)

Zongfu Cao (caozongfu@nrifp.org.cn)

See Also

pheno_extract_HPO, extract_omim, extract_genes_orphanet, extract_clinvar, extract_uniprot

Examples

```
#variantsPDB <- variants_compile(omim = variants.omim,
#                                 clinvar = variants.clinvar,
#                                 uniprot = variants.uniprot)
```

Index