Package 'ungeneanno'

September 21, 2016

Type Package

Title Collate Gene Annotation Data from Uniprot and NIH Gene Databases

Version 0.1.6

Description Taking a list of genes, the package collates together the summary information about those genes from the publicly available resources at Uniprot and NCBI. Additionally, the package is able to collate publication information from a search of the NCBI Pubmed database.

License GPL-3

Depends R (>= 2.10), httr, methods, XML

LazyData TRUE

RoxygenNote 5.0.1.9000

Suggests knitr, rmarkdown, testthat

VignetteBuilder knitr

NeedsCompilation no

Author I. Richard Thompson [aut, cre]

Maintainer I. Richard Thompson <ithompson@qf.org.qa>

Repository CRAN

Date/Publication 2016-09-21 12:35:11

R topics documented:

---

ungeneanno-package *Collate Gene Annotation Data from Uniprot and NIH Gene Databases*

Description

Taking groups of genes, the package collates together the summary information about those genes from the publicly available resources at Uniprot and NCBI. This is achieved through the `getUniqueGeneList`, `getGeneSummary`, `getGroupGeneList` and `produceOutputFiles` functions. Additionally, the package is able to collate publication information from a search of the NCBI Pubmed database via the `getPublicationList` function.

Details

A 2 column matrix, containing a column of 'group' identifers and a column of gene names, is required as input. Unique lists of both group identifiers and gene names are created and the data from the NCBI gene and Uniprot databases is downloaded. A vector of vectors is used to recreate the relationships between the group identifers and gene names, which is then used to create output files detailing the downloaded information for the genes in each group.

Author(s)

Richard Thompson <ithompson@qf.org.qa>

See Also

http://www.ncbi.nlm.nih.gov/books/NBK25500/

http://www.uniprot.org/help/uniprotkb_column_names

Examples

```
## Not run:
## Create geneanno object and set save directory
ga <- geneanno()
ga@fileroot <- "~/Desktop"

## Parse input gene names and group identifiers into unique lists
data("genematrix")
ga <- getUniqueGeneList(ga,genematrix)
```

```
## Query databases and Parse responses
gs <- getGeneSummary(ga)

## Create vector of vectors containing gene names for each group identifier
dgl <- getGroupGeneList(ga,genematrix)

## Collate data into out files
produceOutputFiles(ga, dgl, gs)

## Query PubMed database
query <- "Thompson IR HIV"
ReturnedPublications <- getPublicationList(query)

## List PubMedIDs from PubMed query
for (i in 1:length(ReturnedPublications)){
  print(ReturnedPublications[[i]]@Id)
  }

## End(Not run)
```

`gene-class` *An S4 class to represent a gene.*

Description

An S4 class to represent a gene.

Slots

`name` A character string stating the name of the gene as found in NCBI.

`nih_id` A numeric value stating the NCBI gene database identifier for the gene.

`nih_summary` A character string describing the function and interactions of the gene as found in the NCBI.

`uniprot_name` A character string stating the name of the gene as found in Uniprot.

`uniprot_protein_name` A character string stating the name of the protein gene product as found in Uniprot.

`uniprot_summary` A vector of character strings describing the function and interactions of the gene as found in the Uniprot dataset.

`alternatives` A vector of character strings stating commonly used alternative names for the gene (from NCBI).

`symbol` A character string stating the offical gene symbol (from NCBI).

`chromosome` A character string stating the chromosome upon which the gene resides (from NCBI).

`start` A character string stating the chromosomal position of the transcription start (from NCBI).

`stop` A character string stating the chromosomal position of the transcription stop (from NCBI).

`exon_count` A character stating the number of exons in the gene (from NCBI).

`chraccver` A character string stating the NIH accession number and version for the chromosome (from NCBI).

`map_location` A character string representing the location of the gene on the karyotype (from NCBI).

`direction` A character string [forward/reverse] identifying the direction of the gene on the chromosome (from NCBI).

`geneanno-class` *An S4 class to represent the base information used by the methods.*

Description

An S4 class to represent the base information used by the methods.

Slots

`db` A character string stating the NIH/NCBI database to be queried. (Default - "gene")

`nihbase` A character string stating the base URL for NIH/NCBI queries.

`ids` A Vector of character strings

`uniprotbase` A character string stating the base URL for Uniprot queries.

`uniprotquery` A character string stating additional information to optimize the query to the Uniport database.

`uniprotcolumns` A character string stating the database columns to be returned from the Uniport database.

`genelist` A vector of character strings representing a unique list of the genes queried.

`fileroot` A charcter string of the directory for file input and storage. (defaults to working directory)

`outputstem` A character string stating the sub-directory of 'fileroot' into which output files are saved. (Default - "gene_annotations")

`genefilestem` A character string stating the sub-directory of 'fileroot' into which gene objects are saved. (Default - "genes")

`groupnos` A numeric vector representing the identifiers of groups in the input file/list.

genematrix *Input Matrix of Drug/Gene Data*

Description

A matrix made up of a column of drug identifiers and a column of associated gene names, as used for input into `getUniqueGeneList()` and

Usage

```
genematrix
```

Format

A 2 column matrix where the first columns contains drug identifiers and the second column Gene names.

getGeneSummary *Create gene specific objects containing data from online resources*

Description

`getGeneSummary` populates and returns a vector of gene objects with information sourced from a series of html requests to the NIH and Uniport publicly available databases.

Usage

```
getGeneSummary(x)

## S4 method for signature 'geneanno'
getGeneSummary(x)
```

Arguments

x	object of class geneanno.

Details

Information returned from a database requests is parsed into a gene object, which are saved in a 'genes' subdirectory of the working directory. Each gene object is added to a vector of objects, which is then returned. Where gene information has previously been downloaded and objects saved (within the last seven days), gene objects are repopulated from the saved files so as to minimise server traffic. *N.B.* the function includes a random wait (of up to 5s) between each gene downloaded.

N.B. It is possible to define an alternative directory using `geneanno@fileroot <- "/path/to/directory"`

Value

vector of gene objects, each containing the collated data from the public resources.

Methods (by class)

- geneanno: Produces a vector of gene objects, each containing the collated data from the public resources.

Examples

```
## Not run:
geneanno@fileroot <- "~/Desktop"
genesummaries <- getGeneSummary(geneanno)

## End(Not run)
```

---

getGroupGeneList *Create unique list of group,gene combinations from input list*

Description

getGroupGeneList takes a mixed list of group identifiers (numeric) and gene names or a 2-column matrix of group identifiers and gene names, returning the seperated list of group identifiers and the related lists of genes.

Usage

```
getGroupGeneList(x, inputlist)

## S4 method for signature 'geneanno,character'
getGroupGeneList(x, inputlist)

## S4 method for signature 'geneanno,matrix'
getGroupGeneList(x, inputlist)
```

Arguments

x	object of class geneanno.
inputlist	vector of strings, being a mixed list of group numbers and gene names

Value

object of type geneanno; a copy of input object having the additional list of group numbers and a unique list of genes from s

Methods (by class)

- `x = geneanno,inputlist = character`: uses vector of characters of group ids and Gene names for inputlist.
- `x = geneanno,inputlist = matrix`: uses matrix of group ids and Gene names for inputlist.

Examples

```
## Not run:
data("genematrix")
dgl <- getGroupGeneList(geneanno(),genematrix)

## End(Not run)
```

---

getNihQuery *create initial NIH Query*

Description

`getNihQuery` contacts the NIH E-utilities and carries out an initial data query. Accessing the NCBI NIH databases is a two stage process; the `getNihQuery` function carries out the first stage, returning a short list of IDs and a unique query key. This Query key is then used in the second stage to fetch the results of the query.

Usage

```
getNihQuery(x, db, query)

## S4 method for signature 'query'
getNihQuery(x, db, query)
```

Arguments

`x`	`query` object.
`db`	The name of the NCBI database to be searched for the query, e.g. "gene" or "pubmed"
`query`	The query string for the database

Value

`query` object a copy of the input object 'x', having the query specifiers from NIH added.

Methods (by class)

- `query`: `query` object a copy of the input object 'x', having the query specifiers from NIH added.

See Also

http://www.ncbi.nlm.nih.gov/books/NBK25500/

Examples

```
## Not run:
f <- query()
f@gene <- "BRAF"
db <- "gene"
f <- getNihQuery(f,db,f@gene)
gene <- getNihSummary(gene(),f)

## End(Not run)
```

`getNihSummary` *Gather NIH Gene Data*

Description

`getNihSummary` collates data returned from a html request to the NCBI publicly available gene database to populate the gene object. Accessing the NCBI NIH databases is a two stage process; the `getNihSummary` function carries out the second stage in relation to the NCBI "gene" database, populating NIH specific slots in a gene object.

Usage

```
getNihSummary(x, y)

## S4 method for signature 'gene'
getNihSummary(x, y)
```

Arguments

x	gene object
y	query object (see `getNihQuery()`)

Value

gene object a copy of the input object 'x', having the uniprot data added.

Methods (by class)

- gene: gene object a copy of the input object 'x', having the uniprot data added.

Examples

```
## Not run:
f <- query()
f@gene <- "BRAF"
db <- "gene"
f <- getNihQuery(f,db,f@gene)
gene <- getNihSummary(gene(),f)

## End(Not run)
```

`getPublicationList` *Gather NIH Pubmed Data*

Description

`getPublicationList` collates Pubmed data returned from a html request for input string.

Usage

```
getPublicationList(query)

## S4 method for signature 'character'
getPublicationList(query)
```

Arguments

`query`	A query string to be submitted to Pubmed

Value

A vector of `pubmed` objects each detailing a publication relating to the entered query combination requested.

Methods (by class)

- `character`: as above.

Examples

```
## Not run:
query <- "Thompson IR HIV"
ReturnedPublications <- getPublicationList(query)

## End(Not run)
```

getUniprotSummary *Gather Uniprot Data*

Description

getUniprotSummary collates data returned from a html request to the Uniport publicly available databases to populate the gene object. The query currently returns id, entry name, reviewed, protein names, genes, organism, length and "comment(FUNCTION)".

Usage

```
getUniprotSummary(x, y)

## S4 method for signature 'gene'
getUniprotSummary(x, y)
```

Arguments

x	gene object
y	query object (see getNihQuery())

Value

gene object - a copy of the input object 'y', having the uniprot data added.

Methods (by class)

- gene: gene object - a copy of the input object 'y', having the uniprot data added.

See Also

http://www.uniprot.org/help/uniprotkb_column_names

Examples

```
## Not run:
f <- query()
f@gene <- "BRAF"
gene <- gene()
gene@name <- f@gene
gene <- getUniprotSummary(gene,f)

## End(Not run)
```

getUniqueGeneList *Create unique list of genes from input list*

Description

getUniqueGeneList takes a 2-column matrix of group identifiers and gene names, returning the seperated list of unique group identifiers and a unique list of genes.

Usage

```
getUniqueGeneList(x, inputlist)

## S4 method for signature 'geneanno,character'
getUniqueGeneList(x, inputlist)

## S4 method for signature 'geneanno,matrix'
getUniqueGeneList(x, inputlist)
```

Arguments

x	object of class geneanno.
inputlist	vector of strings, being a mixed list of group numbers and gene names

Value

object of type geneanno; a copy of input object having the additional list of group numbers and a list of genes from uniquelist

Methods (by class)

- x = geneanno,inputlist = character: object of type geneanno; a copy of input object having the additional list of group numbers and a unique list of genes from s
- x = geneanno,inputlist = matrix: object of type geneanno; a copy of input object having the additional list of group numbers and a unique list of genes from s

Examples

```
## Not run:
data("genematrix")
ga <- getUniqueGeneList(ga,genematrix)

## End(Not run)
```

inputlist *Input Vector of Drug/Gene Data*

Description

A mixed list of drug identifiers and gene names, as output by `parseInputFile()` and used for input into `getUniqueGeneList()` and

Usage

```
inputlist
```

Format

A vector of 10 drug identification numbers, intersperced with 90 Gene names.

parseInputFile *Parse files of gene names and group identifiers to create unique lists of each*

Description

`parseInputFile` takes a mixed file containing group identifiers (numeric) and gene names, returning the list of group identifiers and genes with the remaining columns removed. The package was originally written to work from a file laid out thus: group_id1 gene_name1 gene_name2 group_id2 gene_name1 gene_name3 The methods assume that both group identifiers and gene names are alphanumeric; the group identifiers, where present, begining with a number and gene names starting with a character. please note, this populates the vector with only the alphanumric strings begining each line of the input file. Also, RNA genes (begining ENSG000) are excluded.

Usage

```
parseInputFile(x, file)

## S4 method for signature 'geneanno'
parseInputFile(x, file)
```

Arguments

`x`	object of class geneanno.
`file`	character string providing the name of the input file

Value

vector of character strings, as exemplified by the inputlist data object.

Methods (by class)

- geneanno: vector of character strings

---

produceOutputFiles *Save Gene Information to group Specific Output Files*

Description

produceOutputFiles saves the downloaded gene summary information into group specific files.

Usage

```
produceOutputFiles(x, dgl, gs, pub)

## S4 method for signature 'geneanno,vector,vector,missing'
produceOutputFiles(x, dgl, gs)

## S4 method for signature 'geneanno,vector,vector,vector'
produceOutputFiles(x, dgl, gs, pub)
```

Arguments

x	object of class geneanno.
dgl	a Vector of group specific vectors relating the list of genes to the groups, as in the input file
gs	Vector of Gene Summary information downloaded from NIH and Uniprot databases, typically output by getGeneSummary
pub	Matrix containing journal article information from the PubMed resource, as retruned by the searchPublications function

Methods (by class)

- x = geneanno,dgl = vector,gs = vector,pub = missing: Save object data to text files.
- x = geneanno,dgl = vector,gs = vector,pub = vector: Save object data, including journal articles, to text files.

pubmed-class *An S4 class to represent a query to the Pubmed database, inherits from* geneanno *class.*

Description

An S4 class to represent a query to the Pubmed database, inherits from geneanno class.

Slots

`Id` The pubmed ID for the publication

`Authors` List of article authors

`Date` The publication date of the article

`Title` Publication Title

`Journal` The publication journal of the article

`Volume` The journal volume of the article

`Issue` The journal volume issue number of the article

`Pages` The journal pages of the article

`DOI` The unique article Digital Object Identifier (DOI, see www.doi.org)

query-class *An S4 class to represent an initial query to the NCBI e-utilities, inherits from* geneanno *class.*

Description

An S4 class to represent an initial query to the NCBI e-utilities, inherits from geneanno class.

Slots

`gene` A character vector of the gene name.

`querykey` a character string identifier from the NCBI Eutils.

`webenv` A character string uniquely identifying the query from the NCBI Eutils.

searchPublications *Carry out Pubmed search for a mtrix*

Description

searchPublications carries out a series of Pubmed searches for each row of information in the given matrix. The function uses the same matrix used for input to the getUniqueGeneList function, however given the nature of the search it is best to ensure the groups have meaningful names and not arbitrary numbers.

Usage

```
searchPublications(query)

## S4 method for signature 'matrix'
searchPublications(query)
```

Arguments

query	A matrix, as input to getUniqueGenes, string to be submitted to Pubmed

Value

A 2 column matrix containg the query string and a list of pubmed objects each detailing a publication relating to the respective query.

Methods (by class)

- matrix: as above.

Examples

```
## Not run:
query   f <- matrix(c("Axitinib","BRAF","Imatinib","BRAF"),ncol=2,byrow=TRUE)
ReturnedPublications <- searchPublications(query)

## End(Not run)
```

Index