Package 'hgnc'

August 29, 2023

Type Package

Title Download and Import the HUGO Gene Nomenclature Committee ('HGNC') Data Set into R

Version 0.1.4

Description A set of routines to quickly download and import the 'HGNC' data set on mapping of gene symbols to gene entries in other popular databases or resources.

License MIT + file LICENSE

URL https://github.com/ramiromagno/hgnc, https://rmagno.eu/hgnc/

BugReports https://github.com/ramiromagno/hgnc/issues

Encoding UTF-8

RoxygenNote 7.2.3

Imports dplyr, hms, httr, jsonlite, lubridate, magrittr, purrr, readr, rlang, rvest, stringr, tibble

Suggests spelling

Language en-US

NeedsCompilation no

Author Ramiro Magno [aut, cre] (<https://orcid.org/0000-0001-5226-3441>), Ana-Teresa Maia [aut] (<https://orcid.org/0000-0002-0454-9207>), CINTESIS [cph, fnd], Pattern Institute [cph, fnd]

Maintainer Ramiro Magno <rmagno@pattern.institute>

Repository CRAN

Date/Publication 2023-08-29 07:20:02 UTC

R topics documented:

`download_archive` *Download HGNC dataset*

Description

Download the latest HGNC approved data set.

Usage

```
download_archive(
  url = latest_archive_url(),
  path = getwd(),
  filename = basename(url),
  ...
)
```

Arguments

`url`	A character string naming the URL of the HGNC dataset. It defaults to the latest available archive.
`path`	A character string with the directory path where the downloaded file is to be saved. By default, this is the current working directory.
`filename`	A character string with the name of the saved file. By default, this is inferred from the last part of the URL.
`...`	Additional arguments passed on to `download.file()`.

`filter_by_keyword` *Filter genes by keyword*

Description

Filter the HGNC data set by a keyword to be looked up in the columns containing gene names or symbols. By default, it will look up in `symbol`, `name`, `alias_symbol`, `alias_name`, `prev_symbol` and `prev_name`. Note that this function in dive into list-columns and match return a hit result if at least one of the strings matches the `keyword`.

Usage

```
filter_by_keyword(
  tbl,
  keyword,
 cols = c("symbol", "name", "alias_symbol", "alias_name", "prev_symbol", "prev_name")
)
```

Arguments

tbl	A tibble containing the HGNC data set, typically obtained with `import_hgnc_dataset()`.
keyword	A keyword or a regular expression to be used as search criterion.
cols	Columns to be looked up.

Value

A tibble of the HGNC data set filtered by observations matching the `keyword`.

Examples

```
## Not run:
# Start by retrieving the HGNC data set
hgnc_tbl <- import_hgnc_dataset()

# Search for entries containing "TP53" in the HGNC data set
hgnc_tbl %>%
  filter_by_keyword('TP53') %>%
  dplyr::select(1:4)

# The same as above but restrict the search to the `symbol` column
hgnc_tbl %>%
  filter_by_keyword('TP53', cols = 'symbol') %>%
  dplyr::select(1:4)

# Match "TP53" exactly in the `symbol` column
hgnc_tbl %>%
  filter_by_keyword('^TP53$', cols = 'symbol') %>%
  dplyr::select(1:4)

# `filter_by_keyword()` is vectorised over `keyword`
hgnc_tbl %>%
  filter_by_keyword(c('^TP53$', '^PIK3CA$'), cols = 'symbol') %>%
  dplyr::select(1:4)

## End(Not run)
```

`import_hgnc_dataset` *Import HGNC data*

Description

This function imports into memory, as a tibble, the complete HGNC data set from a TSV file.

Usage

```
import_hgnc_dataset(file = latest_archive_url(), ...)
```

Arguments

`file`	A file or URL of the complete HGNC data set (in TSV format).
`...`	Additional arguments to be passed on to `readr::read_tsv()`.

Value

A tibble of the HGNC data set consisting of 55 columns:

`hgnc_id` A unique ID provided by the HGNC for each gene with an approved symbol. IDs are of the format HGNC:n, where n is a unique number. HGNC IDs remain stable even if a name or symbol changes.

`hgnc_id2` A stripped down version of `hgnc_id` where the prefix `"HGNC:"` has been removed (this column is added by the package `{hgnc}`).

`symbol` The official gene symbol approved by the HGNC, which is typically a short form of the gene name. Symbols are approved in accordance with the Guidelines for Human Gene Nomenclature.

`name` The full gene name approved by the HGNC; corresponds to the approved symbol above.

`locus_group` A group name for a set of related locus types as defined by the HGNC. One of: `"protein-coding gene"`, `"non-coding RNA"`, `"pseudogene"` or `"other"`.

`locus_type` Specifies the genetic class of each gene entry:

- `"gene with protein product"` Protein-coding genes (the protein may be predicted and of unknown function), SO:0001217.
- `"RNA, cluster"` Region containing a cluster of small non-coding RNA genes.
- `"RNA, long non-coding"` Non-protein coding genes that encode long non-coding RNAs (lncRNAs) SO:0001877; these are at least 200 nt in length. Subtypes include intergenic SO:0001463, intronic SO:0001903 and antisense SO:0001904.
- `"RNA, micro"` Non-protein coding genes that encode microRNAs (miRNAs), SO:0001265.
- `"RNA, ribosomal"` Non-protein coding genes that encode ribosomal RNAs (rRNAs), SO:0001637.
- `"RNA, small nuclear"` Non-protein coding genes that encode small nuclear RNAs (snRNAs), SO:0001268.
- `"RNA, small nucleolar"` Non-protein coding genes that encode small nucleolar RNAs (snoRNAs) containing C/D or H/ACA box domains, SO:0001267.

"RNA, small cytoplasmic" Non-protein coding genes that encode small cytoplasmic RNAs (scRNAs), SO:0001266.

"RNA, transfer" Non-protein coding genes that encode transfer RNAs (tRNAs), SO:0001272.

"RNA, small misc" Non-protein coding genes that encode miscellaneous types of small ncRNAs, such as vault (SO:0000404) and Y (SO:0000405) RNA genes.

"phenotype only" Mapped phenotypes where the causative gene has not been identified, SO:0001500.

"pseudogene" Genomic DNA sequences that are similar to protein-coding genes but do not encode a functional protein, SO:0000336.

"complex locus constituent" Transcriptional unit that is part of a named complex locus.

"endogenous retrovirus" Integrated retroviral elements that are transmitted through the germline, SO:0000100.

"fragile site" A heritable locus on a chromosome that is prone to DNA breakage.

"immunoglobulin gene" Gene segments that undergo somatic recombination to form heavy or light chain immunoglobulin genes (SO:0000460). Also includes immunoglobulin gene segments with open reading frames that either cannot undergo somatic recombination, or encode a peptide that is not predicted to fold correctly; these are identified by inclusion of the term "non-functional" in the gene name.

"immunoglobulin pseudogene" Immunoglobulin gene segments that are inactivated due to frameshift mutations and/or stop codons in the open reading frame.

"protocadherin" Gene segments that constitute the three clustered protocadherins (alpha, beta and gamma)

"readthrough" A naturally occurring transcript containing coding sequence from two or more genes that can also be transcribed individually.

"region" Extents of genomic sequence that contain one or more genes, also applied to non-gene areas that do not fall into other types.

"T cell receptor gene" Gene segments that undergo somatic recombination to form either alpha, beta, gamma or delta chain T cell receptor genes (SO:0000460). Also includes T cell receptor gene segments with open reading frames that either cannot undergo somatic recombination, or encode a peptide that is not predicted to fold correctly; these are identified by inclusion of the term "non-functional" in the gene name.

"T cell receptor pseudogene" T cell receptor gene segments that are inactivated due to frameshift mutations and/or stop codons in the open reading frame.

"transposable element" A segment of repetitive DNA that can move, or retrotranspose, to new sites within the genome (SO:0000101).

"unknown" Entries where the locus type is currently unknown.

"virus integration site" Target sequence for the integration of viral DNA into the genome.

status Status of the symbol report, which can be either "Approved" or "Entry Withdrawn".

location Chromosomal location. Indicates the cytogenetic location of the gene or region on the chromsome, e.g. "19q13.43". In the absence of that information one of the following may be listed:

"not on reference assembly" Named gene is not annotated on the current version of the Genome Reference Consortium human reference assembly; may have been annotated on previous assembly versions or on a non-reference human assembly.

"unplaced" Named gene is annotated on an unplaced/unlocalized scaffold of the human reference assembly.

"reserved" Named gene has never been annotated on any human assembly.

location_sortable A sortable version of the location column (see above).

alias_symbol Alternative symbols that have been used to refer to the gene. Aliases may be from literature, from other databases or may be added to represent membership of a gene group.

alias_name Alternative names for the gene. Aliases may be from literature, from other databases or may be added to represent membership of a gene group.

prev_symbol This field displays any symbols that were previously HGNC-approved nomenclature.

prev_name This field displays any names that were previously HGNC-approved nomenclature.

gene_group A gene group. Each gene has been assigned to one or more groups, according to either sequence similarity or information from publications, specialist advisors for that group or other databases. Groups may be either structural or functional.

gene_group_id Gene group identifier, an integer number. This column contains the gene group identifiers, see gene_group for the gene group name.

date_approved_reserved The date the entry was first approved.

date_symbol_changed The date the gene symbol was last changed.

date_name_changed The date the gene name was last changed.

date_modified Date the entry was last modified.

entrez_id Entrez gene identifier.

ensembl_gene_id Ensembl gene identifier.

vega_id VEGA gene identifier.

ucsc_id UCSC gene identifier.

ena International Nucleotide Sequence Database Collaboration (GenBank, ENA and DDBJ) accession number(s).

refseq_accession The Reference Sequence (RefSeq) identifier for that entry, provided by the NCBI.

ccds_id Consensus CDS identifier.

uniprot_ids UniProt protein accession.

pubmed_id Pubmed and Europe Pubmed Central PMIDs.

mgd_id Mouse genome informatics database identifier.

rgd_id Rat genome database gene identifier.

lsdb The name of the Locus Specific Mutation Database and URL for the gene.

cosmic Symbol used within the Catalogue of somatic mutations in cancer for the gene.

omim_id Online Mendelian Inheritance in Man (OMIM) identifier.

mirbase miRBase identifier.

homeodb Homeobox Database identifier.

snornabase snoRNABase identifier.

bioparadigms_slc Symbol used to link to the SLC tables database at bioparadigms.org for the gene.

`orphanet` Orphanet identifier.

`pseudogene.org` Pseudogene.org identifier.

`horde_id` Symbol used within HORDE for the gene.

`merops` Identifier used to link to the MEROPS peptidase database.

`imgt` Symbol used within international ImMunoGeneTics information system.

`iuphar` The objectId used to link to the IUPHAR/BPS Guide to PHARMACOLOGY database.

`kznf_gene_catalog` Lawrence Livermore National Laboratory Human KZNF Gene Catalog (LLNL) identifier.

`mamit-trnadb` Identifier to link to the Mamit-tRNA database.

`cd` Symbol used within the Human Cell Differentiation Molecule database for the gene.

`lncrnadb` lncRNA Database identifier.

`enzyme_id` ENZYME EC accession number.

`intermediate_filament_db` Identifier used to link to the Human Intermediate Filament Database.

`rna_central_ids` Identifier in the RNAcentral, The non-coding RNA sequence database.

`lncipedia` The LNCipedia identifier to which the gene belongs. This will only appear if the gene is a long non-coding RNA.

`gtrnadb` The GtRNAdb identifier to which the gene belongs. This will only appear if the gene is a tRNA.

`agr` The Alliance of Genomic Resources HGNC ID for the Human gene page within the resource.

`mane_select` MANE Select nucleotide accession with version (i.e. NCBI RefSeq or Ensembl transcript ID and version).

`gencc` Gene Curation Coalition (GenCC) Database identifier.

Examples

```
## Not run: import_hgnc_dataset()
```

---

last_update *Last update of HGNC data set*

Description

This function returns the date of the most recent update of the HGNC data set.

Usage

```
last_update()
```

Value

A POSIXct date-time object.

Examples

```
try(last_update())
```

latest_archive_url *Latest HGNC archive URL*

Description

Latest HGNC archive URL

Usage

```
latest_archive_url(type = c("tsv", "json"))
```

Arguments

type	The format of the archive: "tsv" or "json".

Value

A string with the latest HGNC archive URL.

Examples

```
latest_archive_url()
```

list_archives *List monthly and quarterly archives*

Description

This function lists the monthly and quarterly archives currently available.

Usage

```
list_archives()
```

Value

A tibble of available archives for download.

Index