Package 'hgnc'

June 18, 2025

Type Package

Title Import Human Gene Nomenclature

Version 0.3.0

Description A set of routines to quickly download and import the HUGO Gene Nomenclature Committee (HGNC) data set on mapping of gene symbols to gene entries in other genomic databases or resources.

License MIT + file LICENSE

URL https://github.com/patterninstitute/hgnc, https://www.pattern.institute/hgnc/

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

Encoding UTF-8

RoxygenNote 7.3.2

Depends R (>= 4.2.0)

Imports cli, dplyr, httr2, memoise, prettyunits, purrr, readr, stringr, tibble

Suggests lubridate, spelling, tidyr

Language en-US

Config/Needs/website patterninstitute/chic, rmarkdown

NeedsCompilation no

Author Ramiro Magno [aut, cre] (ORCID: <https://orcid.org/0000-0001-5226-3441>), Isabel Duarte [aut] (ORCID: <https://orcid.org/0000-0003-0060-2936>), Jacob Munro [aut] (ORCID: <https://orcid.org/0000-0002-2751-0989>), Ana-Teresa Maia [ctb] (ORCID: <https://orcid.org/0000-0002-0454-9207>), Pattern Institute [cph, fnd] (ROR: <https://ror.org/04jrgd746>)

Maintainer Ramiro Magno <rmagno@pattern.institute>

Repository CRAN

Date/Publication 2025-06-18 02:10:02 UTC

Contents

`crosswalk` *Convert an HGNC value to another*

Description

`crosswalk()` will convert values found in one of the columns of an HGNC gene data set to values in another.

Usage

```
crosswalk(value, from, to = from, hgnc_dataset = import_hgnc_dataset())
```

Arguments

`value`	A character vector of values to be matched in the `from` column. These values must match once and only once in the `from` column values, otherwise `NA` is returned.
`from`	The name of the column in the HGNC gene data set (`hgnc_dataset`) where values passed in `value` are used as queries.
`to`	The name of the column whose values are to be returned, corresponding to matches in the `from` column.
`hgnc_dataset`	A data frame corresponding to a HGNC gene data set. Typically, you'd get hold of a HGNC gene data set with `import_hgnc_dataset()`. For testing purposes and an offline solution, you may use alternatively the function `hgnc_dataset_example()` providing a subset.

Examples

```
## Not run:
# Map a gene symbol to its HUGO identifier.
crosswalk(value = "A1BG", from = "symbol", to = "hgnc_id")

# If `from` and `to` refer to the same column, `crosswalk()` will filter
```

```
# out unmatched values by converting them to `NA`.
crosswalk(value = c("A1BG", "Not a gene"), from = "symbol", to = "symbol")

# This is the default behavior, so you can simply call:
crosswalk(value = c("A1BG", "Not a gene"), from = "symbol")

## End(Not run)
```

`download_hgnc_dataset` *Download the human gene nomenclature dataset*

Description

`download_hgnc_dataset()` gets the latest HGNC approved data set.

Usage

```
download_hgnc_dataset(
  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`	The directory path where the downloaded file is to be saved. By default, this value is `NULL` and the data is imported directly into memory without saving into disk.
`filename`	A character string with the name of the saved file. By default, this is inferred from the last part of the URL.

Value

The path to the saved file.

`filter_by_keyword` *Filter genes by keyword*

Description

Filter the HGNC data set by a keyword (or a regex) 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 dives into list-columns for matching and returns a gene entry 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)
```

---

hgnc_dataset_example *Example HGNC data set*

Description

hgnc_dataset_example() provides an example HGNC data set. This example contains only the first 10,000 gene entries of the HGNC data set dated of 2024-07-26 03:11:20.

This is mostly used in example code as it does not require internet connection.

Usage

```
hgnc_dataset_example()
```

Value

A tibble whose structure is documented in import_hgnc_dataset().

Examples

```
hgnc_dataset_example()
```

---

import_hgnc_dataset *Import HGNC data*

Description

import_hgnc_dataset() imports HGNC data into R. Specify a directory path in addition if you wish the save the data to disk.

Usage

```
import_hgnc_dataset(file = latest_archive_url())
```

Arguments

`file`	A file or URL of the complete HGNC data set (in TSV format). Use `list_archives()` to list previous versions of these data. Pass one of the URLs (column `url`) to `file` to import that specific version. By default the value of `file` is the URL corresponding to the latest version, i.e. the returned value of `latest_archive_url()`.

Value

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

- `hgnc_id`: A unique ID provided by 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, 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, including various types of RNA and other gene-related categories, such as pseudogenes and virus integration sites.
- `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 chromosome, e.g., `'19q13.43'`. In the absence of that information, it may be listed as `'not on reference assembly'`, `'unplaced'`, or `'reserved'`.
- `location_sortable`: A sortable version of the `location` column, allowing easier sorting by chromosomal location.
- `alias_symbol`: Alternative symbols that have been used to refer to the gene. Aliases may be from literature, other databases, or represent membership of a gene group.
- `alias_name`: Alternative names for the gene. Aliases may be from literature, other databases, or 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, or other databases.
- `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()
```

Value

A string with the latest HGNC archive URL.

Examples

```
try(latest_archive_url())
```

latest_monthly_url *Latest HGNC monthly URL*

Description

Latest HGNC monthly URL

Usage

```
latest_monthly_url()
```

Value

A string with the latest HGNC monthly archive URL.

Examples

```
try(latest_monthly_url())
```

latest_quarterly_url *Latest HGNC quarterly URL*

Description

Latest HGNC quarterly URL

Usage

```
latest_quarterly_url()
```

Value

A string with the latest HGNC monthly archive URL.

Examples

```
try(latest_quarterly_url())
```

list_archives *List monthly or quarterly archives*

Description

This function lists the monthly and quarterly archives currently available from the HGNC's Google Cloud Storage.

Usage

```
list_archives(release = c("monthly", "quarterly"))
```

Arguments

release	Series type: "monthly" or "quarterly".

Value

A tibble of available archives for download with the following columns:

- series: whether "monthly" or "quarterly".
- dataset: type of data set ("hgnc_complete_set", "symbol-changes-monthly", "withdrawn" or "symbol-changes-quarterly").
- file: file name.
- date: update date.

- `size`: file size.
- `last_modified`: date-time of file last modification on server.
- `md5sum`: MD5 checksum.
- `url`: direct address to the archive.

Examples

```
try(list_archives())
```

Index