Package ‘HGNChelper’

January 20, 2025

Maintainer Levi Waldron <lwaldron.research@gmail.com>

Depends R (>= 4.1.0), methods, utils

Version 0.8.15

Date 2024-11-16

License GPL (>= 2.0)

Title Identify and Correct Invalid HGNC Human Gene Symbols and MGI
Mouse Gene Symbols

Description Contains functions for
identifying and correcting HGNC human gene symbols and MGI mouse gene symbols
which have been converted to date format by Excel, withdrawn, or aliased.
Also contains functions for reversibly converting between HGNC
symbols and valid R names.

URL https://github.com/waldronlab/HGNChelper,
https://waldronlab.io/HGNChelper/

BugReports https://github.com/waldronlab/HGNChelper/issues

LazyData TRUE

RoxygenNote 7.3.2

Encoding UTF-8

Imports splitstackshape

Suggests testthat, knitr, rmarkdown, markdown

VignetteBuilder knitr

NeedsCompilation no

Author Sehyun Oh [aut],
Ayush Aggarwal [aut],
Markus Riester [aut],
Levi Waldron [aut, cre]

Repository CRAN

Date/Publication 2024-11-16 17:10:11 UTC

Contents

`affyToR` *Title function to convert Affymetrix probeset identifiers to valid R names*

Description

This function simply prepends "affy." to the probeset IDs to create valid R names. Reverse operation is done by the rToAffy function.

Usage

```
affyToR(x)
```

Arguments

`x` vector of Affymetrix probeset identifiers, or any identifier which may with a digit.

Value

a character vector that is simply x with "affy." prepended to each value.

`checkGeneSymbols` *Identify outdated or Excel-mogrified gene symbols*

Description

This function identifies gene symbols which are outdated or may have been mogrified by Excel or other spreadsheet programs. If output is assigned to a variable, it returns a data.frame of the same number of rows as the input, with a second column indicating whether the symbols are valid and a third column with a corrected gene list.

Usage

```
checkGeneSymbols(
  x,
  chromosome = NULL,
  unmapped.as.na = TRUE,
  map = NULL,
  species = "human",
  expand.ambiguous = FALSE
)
```

Arguments

`x` A character vector of gene symbols to check for modified or outdated values

`chromosome` An optional integer vector containing the chromosome number of each gene provided through the argument `x`. It should be the same length as the input for `x`. Currently, this argument is implemented only for human gene cases.

`unmapped.as.na` If `TRUE` (default), unmapped symbols will appear as NA in the `Suggested.Symbol` column. If `FALSE`, the original unmapped symbol will be kept.

`map` Specify if you do not want to use the default maps provided by setting species equal to "mouse" or "human". Map can be any other data.frame with colnames identical to `c("Symbol", "Approved.Symbol")`. The default maps can be updated by running the interactive example below.

`species` A character vector of length 1, either "human" (default) or "mouse". If `NULL`, or anything other than "human" or "mouse", then the map argument must be provided.

`expand.ambiguous`

If `FALSE` (default), genes with multiple mapping will only map to its approved symbol as the correct one. If `TRUE`, genes with multiple/ambiguous mapping will map to all the symbols linked to it.

Value

The function will return a data.frame of the same number of rows as the input, with corrections possible from map.

See Also

`mouse.table` for the mouse lookup table, `hgnc.table` for the human lookup table

Examples

```
library(HGNChelper)

## Human
human <- c("FN1", "TP53", "UNKNOWNGENE","7-Sep", "9/7", "1-Mar", "Oct4", "4-Oct",
      "OCT4-PG4", "C19ORF71", "C19orf71")
checkGeneSymbols(human)
```

```
## Mouse
mouse <- c("1-Feb", "Pzp", "A2m")
checkGeneSymbols(mouse, species="mouse")

## expand.ambiguous

## Human
human <- "AAVS1"
checkGeneSymbols(human, expand.ambiguous=FALSE)
checkGeneSymbols(human, expand.ambiguous=TRUE)

## Mouse
mouse <- c("Cpamd8", "Mug2")
checkGeneSymbols(mouse, species = "mouse", expand.ambiguous = FALSE)
checkGeneSymbols(mouse, species = "mouse", expand.ambiguous = TRUE)

## Updating the map
if (interactive()) {
    currentHumanMap <- getCurrentHumanMap()
    checkGeneSymbols(human, map=currentHumanMap)

    # You should save this if you are going to use it multiple times,
    # then load it from file rather than burdening HGNC's servers.
    save(hgnc.table, file="hgnc.table.rda", compress="bzip2")
    load("hgnc.table.rda")
    checkGeneSymbols(human, map=hgnc.table)
}
```

findExcelGeneSymbols *Title function to identify Excel-mogrified gene symbols*

Description

This function identifies gene symbols which may have been mogrified by Excel or other spreadsheet programs. If output is assigned to a variable, it returns a vector of the same length where symbols which could be mapped have been mapped.

Usage

```
findExcelGeneSymbols(
  x,
 mog.map = read.csv(system.file("extdata/mog_map.csv", package = "HGNChelper"), as.is =
    TRUE),
  regex = "impossibletomatch^"
)
```

Arguments

`x` Vector of gene symbols to check for mogrified values

`mog.map` Map of known mogrifications. This should be a dataframe with two columns: original and mogrified, containing the correct and incorrect symbols, respectively.

`regex` Regular expression, recognized by the base::grep function which is called with ignore.case=TRUE, to identify mogrified symbols. The default regex will not match anything. The regex in the examples is an attempt to match all Excel-mogrified HGNC human gene symbols. It is not necessary for all matches to have a corresponding entry in mog.map$mogrified; values in x which are matched by this regex but are not found in mog.map$mogrified simply will not be corrected.

Value

if the return value of the function is assigned to a variable, the function will return a vector of the same length as the input, with corrections possible from mog.map made.

Examples

```
## Available maps from this package:
human <- read.csv(system.file("extdata/mog_map.csv",
                              package = "HGNChelper"), as.is=TRUE)
mouse <- read.csv(system.file("extdata/HGNChelper_mog_map_MGI_AMC_2016_03_30.csv",
                              package = "HGNChelper"), as.is=TRUE)
## This regex is based that provided by Zeeberg et al.,
##  Mistaken Identifiers: Gene name errors can be introduced
## inadvertently when using Excel in bioinformatics.  BMC
##  Bioinformatics 2004, 5:80.
re <- "[0-9]\\-(JAN|FEB|MAR|APR|MAY|JUN|JUL|AUG|SEP|OCT|NOV|DEC)|[0-9]\\.[0-9][0-9]E\\+[[0-9][0-9]"
findExcelGeneSymbols(c("2-Apr", "APR2"), mog.map=human, regex=re)
findExcelGeneSymbols(c("1-Feb", "Feb1"), mog.map=mouse)
```

`getCurrentMaps` *Get the current maps for correcting gene symbols*

Description

Valid human and mouse gene symbols can be updated frequently. Use these functions to get the most current lists of valid symbols, which you can then use as an input to the `map` argument of `checkGeneSymbols`. Make sure to change the default `species="human"` argument to `checkGeneSymbols` if you are doing this for mouse. Use `getCurrentHumanMap` for HGNC human gene symbols from `https://www.genenames.org/` and `getCurrentMouseMap` for MGI mouse gene symbols from `https://www.informatics.jax.org/downloads/reports/MGI_EntrezGene.rpt`.

Usage

```
getCurrentHumanMap()
getCurrentMouseMap()
```

Value

A `data.frame` that can be used for `map` argument of `checkGeneSymbols` function

Examples

```
## Not run:
## human
new.hgnc.table <- getCurrentHumanMap()
checkGeneSymbols(c("3-Oct", "10-3", "tp53"), map=new.hgnc.table)

## mouse
new.mouse.table <- getCurrentMouseMap()
## Set species to NULL or "mouse"
checkGeneSymbols(c("Gm46568", "1-Feb"), map=new.mouse.table, species="mouse")

## End(Not run)
```

`hgnc.table` *All current and withdrawn HGNC gene symbols and Excel-mogrified symbols*

Description

A `data.frame` with the first column providing a gene symbol or known alias (including withdrawn symbols), second column providing the approved HGNC human gene symbol.

- `Symbol`: All valid, Excel-mogrified, and withdrawn symbols
- `Approved.Symbol`: Approved symbols

Usage

```
hgnc.table
```

Format

An object of class `data.table` (inherits from `data.frame`) with 103939 rows and 3 columns.

Source

Extracted from `https://storage.googleapis.com/public-download-files/hgnc/tsv/tsv/hgnc_complete_set.txt` and system.file("extdata/mog_map.csv", package="HGNChelper")

Examples

```
data("hgnc.table", package="HGNChelper")
head(hgnc.table)
```

---

mouse.table *All current and withdrawn MGI mouse symbols and Excel-mogrified symbols*

Description

A `data.frame` with the first column providing a gene symbol or known alias (including withdrawn symbols), second column providing the approved MGI mouse gene symbol.

- `Symbol`: All valid, Excel-mogrified, and withdrawn symbols
- `Approved.Symbol`: Approved symbols

Usage

```
mouse.table
```

Format

An object of class `data.frame` with 790110 rows and 2 columns.

Source

Extracted from `http://www.informatics.jax.org/downloads/reports/MGI_EntrezGene.rpt` and system.file("extdata/HGNChelper_mog_map_MGI_AMC_2016_03_30.csv", package="HGNChelper")

Examples

```
data("mouse.table", package="HGNChelper")
head(mouse.table)
```

rToAffy — *Title function to convert the output of affyToR back to the original Affymetrix probeset identifiers.*

Description

This function simply strips the "affy." added by the affyToR function.

Usage

```
rToAffy(x)
```

Arguments

`x`	the character vector returned by the affyToR function.

Value

a character vector of Affymetrix probeset identifiers.

rToSymbol — *Title function to reverse the conversion made by symbolToR*

Description

This function reverses the actions of the symbolToR function.

Usage

```
rToSymbol(x)
```

Arguments

`x`	the character vector returned by the symbolToR function.

Value

a character vector of HGNC gene symbols, which are not in general valid R names.

See Also

`symbolToR`

`symbolToR`	*Title function to *reversibly* convert HGNC gene symbols to valid R names.*

Description

This function reversibly converts HGNC gene symbols to valid R names by prepending "symbol.", and making the following substitutions: "-" to "hyphen", "@" to "ampersand", and "/" to "forwardslash".

Usage

```
symbolToR(x)
```

Arguments

`x`	vector of HGNC symbols

Value

a vector of valid R names, of the same length as x, which can be converted to the same HGNC symbols using the rToSymbol function.

See Also

`rToSymbol`

Index