Package 'HGNChelper'

March 5, 2019

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

Depends R (>= 3.5.0), methods, utils

Author Levi Waldron and Markus Riester

Version 0.7.1

Date 2019-03-04

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which have been converted to date format by Excel, withdrawn, or aliased.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

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

LazyData TRUE

RoxygenNote 6.1.1

Suggests testthat,
knitr,
rmarkdown

VignetteBuilder knitr

R topics documented:

Index   **10**

HGNChelper-package   *Handy functions for working with HGNC gene symbols and Affymetrix*
                            probeset identifiers.

Description

Contains functions for identifying and correcting HGNC gene symbols which have been converted
to date format by Excel, for reversibly converting between HGNC symbols and valid R names,
identifying invalid HGNC symbols and correcting synonyms and outdated symbols which can be
mapped to an official symbol.

Details

Package:	HGNChelper
Maintainer:	Levi Waldron <lwaldron@hsph.harvard.edu>
Depends:	R (>= 2.10)
Author:	Levi Waldron and Markus Riester
Version:	0.7.1
Date:	2019-03-04
License:	GPL (>2.0)
Title:	Handy functions for working with HGNC gene symbols and Affymetrix probeset identifiers.
URL:	https://bitbucket.org/lwaldron/hgnchelper
BugReports:	https://bitbucket.org/lwaldron/hgnchelper

Author(s)

Levi Waldron and Markus Riester

---

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checkGeneSymbols   *Title Identify outdated or Excel-mogrified gene symbols*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This function identifies gene symbols which are outdated or may have been mogrified by Excel or other spreadsheet programs.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, unmapped.as.na = TRUE, map = NULL,
  species = "human")
```

Arguments

x   Vector of gene symbols to check for mogrified or outdated values

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 col-names(map) 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.

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 = 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")
if (interactive()){
  ##Run checkGeneSymbols with a brand-new map downloaded from HGNC:
  source(system.file("hgncLookup.R", package = "HGNChelper"))
  ## 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, species=NULL, map=hgnc.table)
  checkGeneSymbols(human, species=NULL, map=mouse.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 input to the "map" argument of check-GeneSymbols(). Make sure to change the default species="human" argument to checkGeneSymbols() if you are doing this for mouse. getCurrentHumanMap() for HGNC human gene symbols from genenames.org getCurrentMouseMap() for MGI mouse gene symbols from www.informatics.jax.org/downloads/reports

Usage

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

Value

a 'data.frame' that can be used as the checkGeneSymbols "map" argument

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"
## so that human-like capitalization corrections aren't made
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 mogrifica-*
                                   tions.

Description

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

Usage

```
hgnc.table
```

Format

An object of class `data.frame` with 97729 rows and 2 columns.

Details

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

Source

Extracted from ftp://ftp.ebi.ac.uk/pub/databases/genenames/new/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 mogrifications.*

Description

A dataframe with the first column providing a gene symbol or known alias (including withdrawn symbols), second column providing the approved MOUSE symbol.

Usage

```
mouse.table
```

Format

An object of class `data.frame` with 134371 rows and 2 columns.

Details

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

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 "forward-slash".

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