Package 'forensicpopdata'

May 2, 2025

Type Package

Title Allele Frequency Data for Human Genetic Markers

Version 1.0.4

Description Provides allele frequency data for Short Tandem Repeat human
genetic markers commonly used in forensic genetics for human
identification and kinship analysis. Includes published population
frequency data from the US National Institute of Standards and
Technology, Federal Bureau of Investigation and the UK government.

License GPL (>= 3)

Encoding UTF-8

LazyData true

RoxygenNote 7.3.2

Depends R (>= 3.5)

Imports xml2

NeedsCompilation no

Author Maarten Kruijver [aut, cre] (ORCID:
<https://orcid.org/0000-0002-6890-7632>)

Maintainer Maarten Kruijver <maarten.kruijver@esr.cri.nz>

Repository CRAN

Date/Publication 2025-05-02 08:40:02 UTC

Contents

---

allele_counts_to_freqs

Convert allele counts data frame to list of frequencies by locus

---

Description

Convert allele counts data frame to list of frequencies by locus

Usage

```
allele_counts_to_freqs(x, remove_zeroes = TRUE)
```

Arguments

x   A data fram with columns: `locus`, `allele`, and `count`.

remove_zeroes   Logical. Should zero-count alleles be removed? Default is `TRUE`.

Value

Named list with frequencies per locus. Each element is a named numeric vector of allele frequencies. An attribute `N` gives the number of allele observations per locus.

Examples

```
x <- data.frame(
  locus = "D3S1358",
  allele = c("12", "13", "14", "15", "15.2", "16", "17", "18", "19"),
  count = c(3, 2, 62, 211, 1, 218, 145, 39, 3)
)
freqs <- allele_counts_to_freqs(x)
freqs
attr(freqs, "N")
```

FBI2015freqs   *FBI 2015 Population Data for the expanded CODIS core STR loci*

Description

A data set containing allele frequencies for 23 autosomal STR loci from the FBI 2015 population data set. Frequencies are provided for are determined with both the GlobalFiler and Fusion kits in African Americans, Caucasians, Southeastern Hispanics, Southwestern Hispanics, Bahamians, Jamaicans, Trinidadians, Apaches, Navajos, Chamorros and Filipinos.

Usage

```
FBI2015freqs
```

Format

A named list of length 12.

Each element is itself a named list of 23 STR loci, with named numeric vectors of allele frequencies.

Details

Each population group is a named list of 23 elements, where each element corresponds to a specific STR locus (e.g., D3S1358, vWA, FGA, etc.). Each locus is represented as a named numeric vector:

- **Names**: allele values (as character strings, e.g., "12", "14.2")
- **Values**: allele frequencies for that population group

An attribute "N" is attached to each population list, specifying the sample size (number of alleles) for each locus.

Source

Raw data (public domain) on which the data set is based is available online on https://ucr.fbi.gov/lab/biometric-analysis/codis/expanded-fbi-str-2015-final-6-16-15.pdf

References

Moretti, T.R., et al. (2016) *Population data on the expanded CODIS core STR loci for eleven populations of significance for forensic DNA analyses in the United States*. Forensic Sci. Int. Genet. 25:p175–181. doi:10.1016/j.fsigen.2016.07.022

Examples

```
# Access allele frequencies for D3S1358 in African American population
FBI2015freqs$`African American`$D3S1358

# Frequency of allele "15" at D3S1358 in Caucasian population
FBI2015freqs$Caucasian$D3S1358["15"]
```

NIST1036freqs   *NIST 1036 Allele Frequency Data for 29 STR Loci*

Description

A dataset containing allele frequencies for 29 autosomal STR loci from the NIST 1036 U.S. Population dataset. Frequencies are provided for four population groups: African American (AfAm), Asian (Asian), Caucasian (Cauc), and Hispanic (Hisp).

Usage

```
NIST1036freqs
```

Format

A named list of length 4:

AfAm African American allele frequencies

Asian Asian allele frequencies

Cauc Caucasian allele frequencies

Hisp Hispanic allele frequencies

Each element is itself a named list of 29 STR loci, with named numeric vectors of allele frequencies.

Details

This dataset is based on the revised genotypes from 2017. The 2017 revision incorporates some changes to the dataset from Hill et al. (2013). Details are provided in the referenced NIST presentation explaining revisions (2017) and Steffen et al. (2017).

Each population group is a named list of 29 elements, where each element corresponds to a specific STR locus (e.g., D3S1358, vWA, FGA, etc.). Each locus is represented as a named numeric vector:

- **Names**: allele values (as character strings, e.g., "12", "14.2")
- **Values**: allele frequencies for that population group

An attribute "N" is attached to each population list, specifying the sample size (number of alleles) for each locus.

Source

Raw data (public domain) on which the data set is based is listed as U.S. Population Dataset 1036 (NIST) on https://strbase.nist.gov

References

Hill, C. R., Duewer, D. L., Kline, M. C., et al. (2013). *U.S. population data for 29 autosomal STR loci*. Forensic Sci. Int. Genet. 7:e82–e83. doi:10.1016/j.fsigen.2012.12.004

Steffen, C. R., Coble, M. D., Gettings, K. B., et al. (2017). *Corrigendum to "U.S. Population Data for 29 Autosomal STR Loci" [Forensic Sci. Int. Genet. 7 (2013) e82–e83]*. Forensic Sci. Int. Genet. 31:e36–e40. doi:10.1016/j.fsigen.2017.08.011

NIST presentation explaining revisions (2017): https://strbase.nist.gov/NIST_Resources/Population_Data/Vallone-Error-Management-July-25-2017.pdf

Examples

```
# Access allele frequencies for D3S1358 in African American population
NIST1036freqs$AfAm$D3S1358

# Frequency of allele "15" at D3S1358 in Caucasian population
NIST1036freqs$Cauc$D3S1358["15"]
```

---

read_allele_freqs   *Read allele frequencies in FSIgen format (.csv)*

Description

Read allele frequencies in FSIgen format (.csv)

Usage

```
read_allele_freqs(filename, remove_zeroes = TRUE, normalise = TRUE)
```

Arguments

filename	Path to csv file.
remove_zeroes	Logical. Should frequencies of 0 be removed from the return value? Default is TRUE.
normalise	Logical. Should frequencies be normalised to sum to 1? Default is TRUE.

Details

Reads allele frequencies from a .csv file. The file should be in FSIgen format, i.e. comma separated with the first column specifying the allele labels and one column per locus. The last row should be the number of observations. No error checking is done since the file format is only loosely defined, e.g. we do not restrict the first column name or the last row name.

Value

Named list with frequencies by locus. The frequencies at a locus are returned as a named numeric vector with names corresponding to alleles.

Examples

```
# below we read an allele freqs file that comes with the package
filename <- system.file("extdata","FBI_extended_Cauc_022024.csv",package = "forensicpopdata")
freqs <- read_allele_freqs(filename)
freqs # the output is a list with an attribute named \code{N} giving the sample size.
```

---

read_STRidER_xml   *Parse allele frequencies from STRidER database*

Description

Parse allele frequencies from STRidER database

Usage

```
read_STRidER_xml(xml_file = "https://strider.online/frequencies/xml")
```

Arguments

xml_file   Path to XML file. Default is `"https://strider.online/frequencies/xml"`

Value

A named list by population. Each population is a list of loci with named numeric vectors of allele frequencies. Each vector has an attribute N for sample size (number of alleles observed).

References

Bodner M. et al. (2016), 'Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER).', Forensic Sci. Int. Genet. 24, 97-102. doi:10.1016/j.fsigen.2016.06.008

@importFrom xml2 read_xml xml_find_all xml_text xml_find_first xml_attr @importFrom stats setNames

@examplesIf interactive() # Import STRidER database freqs <- read_STRidER_xml()

Origins names(freqs)

Access frequencies at the TH01 locus for the NORWAY origin freqs$NORWAY$TH01

UKDNA17freqs   *UK DNA-17 Allele Frequency Data for 16 STR Loci*

Description

A dataset containing allele frequencies for 16 autosomal STR loci from the UK Population dataset. Frequencies are provided for four population groups: `"White_-_EA1_&_EA2"`, `"Black_African_&_Caribbean_-_EA3"`, `"Indian_-_EA4"` and `"Chinese_-_EA5"`.

Usage

```
UKDNA17freqs
```

Format

A named list of length 4.

Each element is itself a named list of 16 STR loci, with named numeric vectors of allele frequencies.

Details

Each population group is a named list of 16 elements, where each element corresponds to a specific STR locus (e.g., D3S1358, vWA, FGA, etc.). Each locus is represented as a named numeric vector:

- **Names**: allele values (as character strings, e.g., ″12″, ″14.2″)
- **Values**: allele frequencies for that population group

An attribute ″N″ is attached to each population list, specifying the sample size (number of alleles) for each locus.

Source

Raw data on which the data set is based is available from https://www.gov.uk/government/statistics/dna-population-data-to-support-the-implementation-of-national-dna-database-dna-17-profi under the Open Government licence https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/

Examples

```
# Access allele frequencies for D3S1358 in the Indian_-_EA4 population
UKDNA17freqs$`Indian_-_EA4`$D3S1358

# Frequency of allele ″15″ at D3S1358 in the Indian_-_EA4 population
UKDNA17freqs$`Indian_-_EA4`$D3S1358[″15″]
```

Index