Package 'MMD'

January 26, 2021

Type Package

Title Minimal Multilocus Distance (MMD) for Source Attribution and Loci Selection

Imports e1071, plyr, bigmemory

Version 1.0.0

Maintainer Francisco Perez-Reche <fperez-reche@abdn.ac.uk>

Description The aim of the package is two-fold: (i) To implement the MMD method for attribution of individuals to sources using the Hamming distance between multilocus genotypes. (ii) To select informative genetic markers based on information theory concepts (entropy, mutual information and redundancy). The package implements the functions introduced by Perez-Reche, F. J., Rotariu, O., Lopes, B. S., Forbes, K. J. and Strachan, N. J. C. Mining whole genome sequence data to efficiently attribute individuals to source populations. Scientific Reports 10, 12124 (2020) <doi:10.1038/s41598-020-68740-6>. See more details and examples in the README file.

License GPL-3

Encoding UTF-8

LazyData true

Suggests knitr, rmarkdown

VignetteBuilder knitr

RoxygenNote 7.1.0

NeedsCompilation no

Author Francisco Perez-Reche [aut, cre]

Repository CRAN

Date/Publication 2021-01-26 13:40:08 UTC

R topics documented:

`MMD_attr` *Attribution of individuals to sources using the MMD method*

Description

Attribution of individuals to sources using the MMD method

Usage

```
MMD_attr(
  datafile,
  popfile,
  NL,
  sourcenames,
  ToAttribute,
  SelfA = "no",
  fSelfA = 0.5,
  randomSelfA = "yes",
  quantile = 0.01,
  optq = "no",
  pqmin = 0,
  pqmax = 0.5,
  np = 20,
  Nbootstrap = 10000,
  verbose = FALSE
)
```

Arguments

`datafile`	character; Name of the file *.csv (with full path in the file system) containing the genotypes (features) of individuals.
`popfile`	character; Name of the file *.pop (with full path in the file system) containing the genotypes (features) of individuals.
`NL`	integer; number of loci. If larger than the number of available loci in the data set, NL is reduced to the maximum available number of loci.
`sourcenames`	a character vector listing the names of the sources.
`ToAttribute`	character giving the name of the individuals of aknown origin (i.e. those that will be attributed to source).
`SelfA`	character; if "no" attribution of individuals to sources is made; if "yes", self-attribution of selected individuals from sources is made. (Default "no")
`fSelfA`	real number in the interval (0,1). When SelfA="yes", fSelfA specifies the fraction of individuals from the source specified by ToAttribute that will be assumed to be of unknown origin. (Default 0.1)

randomSelfA	character only relevant if SelfA="yes". If "yes", individuals to be considered as unknown are randomly selected from the source specified by ToAttribute; if "no" a list of names for individuals is read from filepoplist. (Default "yes")
quantile	real number with values in (0,1) giving the q-quantile for the MMD method. Only used if the quantile is not obatined through optimisation of the probability of correct self-attribution. (Default 0.01)
optq	character; if "no", the specified quantile value is used; if "yes", the q-quantile is optimised (only meaningful for self-attribution so optq="no" automatically if SelfA="no"). (Default "no")
pqmin	real number with values in (0,1); minimum value of q-quantile when optq="yes". (Default 0)
pqmax	real number with values in (0,1); maximum value of q-quantile when optq="yes". (Default 0.5)
np	integer giving the number of values of q-quantile in the interval (pqmin,pqmax) when optq="yes". (Default 20)
Nbootstrap	integer giving the number of samples used for bootstrapping to estimate the uncertainty of the attribution probability p_s bootstrap. (Default 10000)
verbose	boolean (TRUE/FALSE) for the display of a progress bar (Default FALSE)

Value

If optq="yes", the output is a list with seven elements:

1. Number of individuals from unknown origin.
2. Number of sources.
3. Statistics of the attribution probability to sources, p_s.
4. Probability of attribution of each unknown individual to each source p_u,s
5. Runtime of the calculation.
6. Number of loci.
7. Parameter q used to calculate the q-quantile of the Hamming distance in the MMD method.
8. Data frame giving the probability of correct attribution vs. q-quantile.

If optq="no", the output list contains all the items in the list above except the last one.

Author(s)

Francisco J. Perez-Reche (Univeristy of Aberdeen)

Examples

```
## This example uses a small dataset stored in the MMD package
datafile <- system.file("extdata", "Campylobacter_10SNP_HlW.csv", package = "MMD")
popfile <- system.file("extdata", "Campylobacter_10SNP_HlW.pop", package = "MMD")

NL <- 100
sourcenames <- c("Cattle","Chicken","Pig","Sheep","WB")
```

```
##----- Source attribution
ToAttribute <- "Human"
SelfA="no"
attribution <- MMD_attr(datafile,popfile,NL,sourcenames,ToAttribute)

## See more detailed examples in the vignette.
```

`MMD_Entropy` *Loci entropies to measure allele diversity*

Description

Loci entropies to measure allele diversity

Usage

```
MMD_Entropy(datafile, popfile, NL, sourcenames, verbose = FALSE)
```

Arguments

`datafile`	character; Name of the file *.csv (with full path in the file system) containing the genotypes (features) of individuals.
`popfile`	character; Name of the file *.pop (with full path in the file system) containing the genotypes (features) of individuals.
`NL`	integer; number of loci. If larger than the number of available loci in the data set, NL is reduced to the maximum available number of loci.
`sourcenames`	a character vector listing the names of the sources.
`verbose`	boolean (TRUE/FALSE) for the display of a progress bar (Default FALSE)

Value

A list with

1. Number of loci.
2. Number of individuals in sources.
3. Table with proportional weight of each population, qs.
4. Number of alleles in the dataset.
5. Value of the alleles in the dataset.
6. Data frame with three columns for entropies: HlT, HlW, HlB
7. Runtime.

Author(s)

Francisco J. Perez-Reche (Univeristy of Aberdeen)

Examples

```
## This example uses a small dataset stored in the MMD package
datafile <- system.file("extdata", "Campylobacter_10SNP_HlW.csv", package = "MMD")
popfile <- system.file("extdata", "Campylobacter_10SNP_HlW.pop", package = "MMD")

NL <- 100
sourcenames <- c("Cattle","Chicken","Pig","Sheep","WB")

EntropyLoci <- MMD_Entropy(datafile,popfile,NL,sourcenames)

## See more detailed examples in the vignette.
```

MMD_Rn *Loci redundancy in sequences*

Description

Loci redundancy in sequences

Usage

```
MMD_Rn(datafile, popfile, NL, sourcenames, verbose = FALSE)
```

Arguments

datafile	character; Name of the file *.csv (with full path in the file system) containing the genotypes (features) of individuals.
popfile	character; Name of the file *.pop (with full path in the file system) containing the genotypes (features) of individuals.
NL	integer; number of loci. If larger than the number of available loci in the data set, NL is reduced to the maximum available number of loci.
sourcenames	a character vector listing the names of the sources.
verbose	boolean (TRUE/FALSE) for the display of a progress bar (Default FALSE)

Value

A list with

1. Number of loci.
2. Number of individuals in sources.
3. Data frame with proportional weight of each population, qs.
4. Number of alleles in the dataset.
5. Value of the alleles in the dataset.
6. Dataframe with two columns: (1) Index of locus. (2) Rn for loci in the original dataset.

7. numerical; index of loci with increasing Rn.
8. Dataframe with two columns: (1) Index of loci sorted by increasing Rn. (2) Value of Rn in increasing order.
9. Runtime.

Author(s)

Francisco J. Perez-Reche (Univeristy of Aberdeen)

Examples

```
## This example uses a small dataset stored in the MMD package
datafile <- system.file("extdata", "Campylobacter_10SNP_HlW.csv", package = "MMD")
popfile <- system.file("extdata", "Campylobacter_10SNP_HlW.pop", package = "MMD")

NL <- 100
sourcenames <- c("Cattle","Chicken","Pig","Sheep","WB")

RedundancyLoci <- MMD_Rn(datafile,popfile,NL,sourcenames)

## See more detailed examples in the vignette.
```

Index