Package ‘fst4pg’

October 13, 2022

Title Genetic Distance Segmentation for Population Genetics

Version 1.0.0

Date 2022-06-15

Description Provides efficient methods to compute local and genome wide genetic distances (corresponding to the so called Hudson Fst parameters) through moment method, perform chromosome segmentation into homogeneous Fst genomic regions, and selection sweep detection for multi-population comparison. When multiple profile segmentation is required, the procedure can be parallelized using the future package.

License GPL (>= 2)

Depends R (>= 2.10)

Imports dplyr, fpopw, furrr, future, ggplot2, gplots, graphics, grDevices, purrr, rlang, scales, stats, stringr, tibble, tidyr, utils

LazyData true

RoxygenNote 7.2.0

NeedsCompilation no

Author Tristan Mary-Huard [aut, cre] (<`https://orcid.org/0000-0002-3839-9067`>), Guillem Rigaill [aut] (<`https://orcid.org/0000-0002-7176-7511`>)

Maintainer Tristan Mary-Huard <`tristan.mary-huard@agroparistech.fr`>

Repository CRAN

Date/Publication 2022-07-03 18:20:02 UTC

R topics documented:

`BuildFreqNbG` *Convert the Freq and NbGamete tables into a list.*

Description

The function builds a list where each element corresponds to a population present in both Freq and NbGametes (all other populations being discarded). Each element consists of a data.frame with 2 columns, Freq and NbGamete.

Usage

```
BuildFreqNbG(Freq, NbGamete)
```

Arguments

`Freq`	A data.frame or matrix of frequencies where each row corresponds to a marker, each column corresponds to a population,
`NbGamete`	A data.frame or matrix of number of gametes where each row corresponds to a marker, and each column corresponds to a population

Value

a list of data.frames, each corresponding to a population.

Examples

```
## Load the HGDP data
data(Freq);data(NbGamete)
FreqNbG <- BuildFreqNbG(Freq,NbGamete)
```

`Compute_Denominator` *Computation of the numerator of the moment estimator*

Description

Computation of the numerator of the moment estimator

Usage

```
Compute_Denominator(p1, p2)
```

Arguments

`p1`	numeric, frequencies in population 1
`p2`	numeric, frequencies in population 2

Value

a vector with the denominators of the Fst moment estimator

`Compute_Nominator` *Computation of the numerator of the moment estimator*

Description

Computation of the numerator of the moment estimator

Usage

```
Compute_Nominator(p1, p2, n1, n2)
```

Arguments

`p1`	numeric, frequencies in population 1
`p2`	numeric, frequencies in population 2
`n1`	numeric, number of gametes in population 1
`n2`	numeric, number of gametes in population 2

Value

a vector with the numerators of the Fst moment estimator

`ContrastGraphSummary` *ContrastGraphSummary*

Description

Display mean ratio and/or number of selection graphs

Usage

```
ContrastGraphSummary(
  CS,
  Info,
  Ratio.thres,
  Coef = 1,
  CutNbSel = NULL,
  CutMeanRatio = NULL
)
```

Arguments

`CS`	a contrast summary, as provided by function `ContrastSummary`
`Info`	a data.frame providing information about markers
`Ratio.thres`	a numeric value, regions exhibiting Fst levels whose ratio with the reference level is higher than Ratio.thres will be highlighted.
`Coef`	a scalar, controling font sizes for the graph, optional
`CutNbSel`	a scalar, providing a y-value for an horizontal line on the NbSel graph
`CutMeanRatio`	a scalar, providing a y-value for an horizontal line on the MeanRatio graph

Value

two ggplots objects, called NbSel and MeanRatio, respectively.

`ContrastSummary` *ContrastSummary*

Description

Summarize multiple Fst profiles

Usage

```
ContrastSummary(PS, RefLevel, Ratio.thres = 3, NbSnp.min = 1)
```

Arguments

`PS` a list of profile summaries, as provided by the `ProfilingSummary` function

`RefLevel` a list of reference (i.e. baseline) Fst levels

`Ratio.thres` a numeric value, regions exhibiting Fst levels whose ratio with the reference level is higher than Ratio.thres will be highlighted.

`NbSnp.min` an integer. The minimum number of markers required to highlight a region

Value

a tibble

Examples

```
## The full example execution takes a few seconds.
# data(Freq);data(NbGamete)
# FreqNbG <- BuildFreqNbG(Freq,NbGamete)
# HFst.m <- HudsonFst.m(FreqNbG)

## Two sets of populations to contrast
# Contrast <- list(America=c("Colombian","Maya"),Europe=c("Tuscan","Italian"))
# Profiles <- HudsonFst.prof(HFst.m,Contrast=Contrast)
# PS <- ProfilingSummary(Profiles,Info)

# RefLevel <- rapply(Profiles,median,classes = "numeric",how='list')
# Ratio.thres <- 3
# NbSnp.min <- 1
# CS <- ContrastSummary(PS, RefLevel,
#                       Ratio.thres=Ratio.thres,
#                       NbSnp.min=NbSnp.min)
```

`ContrastTopRegions` *ContrastTopRegions*

Description

ContrastTopRegions

Usage

```
ContrastTopRegions(CS, Crit, Info, Thres, Simplify = FALSE)
```

Arguments

`CS` a list of contrast summaries as obtained from function `ContrastSummary`

`Crit` a string providing the name of the variable to use to select regions

`Info` a data.frame providing information about markers

`Thres` the threshold to be used on the Crit variable

`Simplify` a boolean specifying whether the results should be displayed as a list (by-default option) or as a single data.frame

Value

a data.frame or a list of data.frames

Examples

```
## The full example execution takes a few seconds.
# data(Freq);data(NbGamete)
# FreqNbG <- BuildFreqNbG(Freq,NbGamete)
# HFst.m <- HudsonFst.m(FreqNbG)

## Two sets of populations to contrast
# Contrast <- list(America=c("Colombian","Maya"),Europe=c("Tuscan","Italian"))
# Profiles <- HudsonFst.prof(HFst.m,Contrast=Contrast)
# PS <- ProfilingSummary(Profiles,Info)

# RefLevel <- rapply(Profiles,median,classes = "numeric",how='list')
# Ratio.thres <- 3
# NbSnp.min <- 1
# CS <- ContrastSummary(PS, RefLevel,
#                       Ratio.thres=Ratio.thres,
#                       NbSnp.min=NbSnp.min)
# NbSel.thres <- 2
# TopRegions <- ContrastTopRegions(CS = CS,Crit = 'NbSel',Info = Info,
#                                  Thres = NbSel.thres, Simplify=TRUE)
```

DF4Plot1Prof *DF4Plot1Prof*

Description

Shape the data for Fst profile representation

Usage

```
DF4Plot1Prof(Info, HF, FstProf, Coord = NULL, Threshold = NULL)
```

Arguments

`Info`	a data.frame providing information about markers
`HF`	a data.frame with 2 columns Fst and Weight, as obtained from the `HudsonFst.m` function
`FstProf`	an Fst profile, as obtained from the `HudsonFst.prof` function
`Coord`	a vector with the minimum and maximum coordinates (i.e. positions along the genome), providing the range of the genomic region that will be plotted, optional.
`Threshold`	a numeric value. Markers belonging to regions whose Fst profile is higher than threshold will be highlighted. Optional.

Value

a data.frame that can be used as an input for function `Plot1Prof`

`Freq` *Frequencies of the American and European HGDP populations*

Description

`Freq` is a data.frame containing the frequencies of 49,636 markers located on chromosome 1, for the 13 American and European populations described in the Stanford HGDP dataset.

Usage

```
Freq
```

Format

A data.frame

`Freq.filt` *Filtering markers based on allelic frequencies*

Description

Filtering markers based on allelic frequencies

Usage

```
Freq.filt(FreqNbG, Maf = 0)
```

Arguments

`FreqNbG`	a list of data.frames (one per population) with 2 columns: Freq and NbGamete
`Maf`	a numerci value for the thresholding of minor allelic frequencies

Value

a vector of positions to be removed

HeatMap *HeatMap*

Description

Make a frequency heatmap

Usage

```
HeatMap(
  Min,
  Max,
  chr = NULL,
  Info,
  FreqNbG,
  Dir = NULL,
  Weights = NULL,
  Weight.thres = 0.05,
  NbAdjM = 0,
  Subsets = NULL
)
```

Arguments

`Min`	the starting position value of the region
`Max`	the end position value of the region
`chr`	a string providing the chromosome name, optional
`Info`	a data.frame providing information about markers
`FreqNbG`	a list of data.frames (one per population) with two columns: Freq and NbGamete
`Dir`	a string providing the name of the directory where the graph should be saved, optional
`Weights`	a vector of weights associated with each marker, optional
`Weight.thres`	a numeric value. Markers with weights lower than this threshold will be discarded from the graphical representation. Optional.
`NbAdjM`	an integer providing the number of markers before and after the highlighted regions that should be added to the graphical representation, optional.
`Subsets`	a list of character vectors with the population names, optional.

Value

A heatmap where rows correspond to markers, and columns to populations.

Examples

```
## The full example execution takes a few seconds.
# data(Freq);data(NbGamete)
# FreqNbG <- BuildFreqNbG(Freq,NbGamete)
# HFst.m <- HudsonFst.m(FreqNbG)

## Two sets of populations to contrast
# Contrast <- list(America=c("Colombian","Maya"),Europe=c("Tuscan","Italian"))
# Profiles <- HudsonFst.prof(HFst.m,Contrast=Contrast)
# PS <- ProfilingSummary(Profiles,Info)

# RefLevel <- rapply(Profiles,median,classes = "numeric",how='list')
# Ratio.thres <- 3
# NbSnp.min <- 1
# CS <- ContrastSummary(PS, RefLevel,
#                       Ratio.thres=Ratio.thres,
#                       NbSnp.min=NbSnp.min)
# NbSel.thres <- 2
# TopRegions <- ContrastTopRegions(CS = CS,Crit = 'NbSel',Info = Info,
#                                  Thres = NbSel.thres, Simplify=TRUE)
# HeatMap(Min = TopRegions[1,]$Start,
#         Max = TopRegions[1,]$End,
#         chr = TopRegions[1,]$Chromosome,
#         Info = Info,
#         FreqNbG = FreqNbG,
#         Subsets = Contrast)
```

`HudsonFst.gw` *HudsonFst.gw*

Description

Compute genome-wide Hudson Fst moment estimator

Usage

```
HudsonFst.gw(HFst.m, Mat = TRUE)
```

Arguments

`HFst.m`	a list of data.frames as obtained with function `HudsonFst.m`
`Mat`	boolean, should the result be output as a matrix.

Value

By default a matrix of Hudson Fst coefficients, a vector otherwise.

Examples

```
data(Freq);data(NbGamete)
FreqNbG <- BuildFreqNbG(Freq,NbGamete)
HFst.m <- HudsonFst.m(FreqNbG)
HFst.chr <- HudsonFst.gw(HFst.m)
```

`HudsonFst.m` *HudsonFst.m*

Description

Compute Hudson Fst moment estimator at marker level

Usage

```
HudsonFst.m(FreqNbG)
```

Arguments

`FreqNbG`	a list of data.frames (one per population) with 2 columns: Freq and NbGamete

Value

a list of data.frames with 2 columns: Fst and Weight.

Examples

```
## Load the FreqNbG object build from the HGDP data
data(Freq);data(NbGamete)
FreqNbG <- BuildFreqNbG(Freq,NbGamete)
HFst.m <- HudsonFst.m(FreqNbG)
```

`HudsonFst.plot` *Plot Fst values along chromosomes*

Description

Plot Fst values along chromosomes

Usage

```
HudsonFst.plot(
  Info,
  HFst.m,
  HFst.prof = NULL,
  Coord = NULL,
  Ref = NULL,
  Threshold = NULL
)
```

Arguments

`Info`	a data.frame providing information about markers
`HFst.m`	a data.frame with 2 columns, Fst and Weight, as provided by the `HudsonFst.m` function
`HFst.prof`	a data.frame corresponding to one item of the output of the `HudsonFst.prof` function
`Coord`	a vector with the minimum and maximum coordinates (i.e. positions along the genome) providing the range of the genomic region that will be plotted.
`Ref`	a value to plot a reference line
`Threshold`	a value to plot a threshold line

Value

a ggplot object

Examples

```
## The full example execution takes a few seconds.
data(Freq);data(NbGamete)
FreqNbG <- BuildFreqNbG(Freq,NbGamete)
HFst.m <- HudsonFst.m(FreqNbG)
TwoPops <- list(First="Colombian",Second="Tuscan")
HFst.prof <- HudsonFst.prof(HFst.m,Contrast=TwoPops)

## Plot the raw data

HudsonFst.plot(Info,HFst.m$Colombian_Tuscan)

## Plot the raw data and the segmentation

HudsonFst.plot(Info,HFst.m$Colombian_Tuscan,HFst.prof$Colombian_Tuscan)

## Add a background/reference level

RefLevel <- median(HFst.prof$Colombian_Tuscan)
HudsonFst.plot(Info,HFst.m$Colombian_Tuscan,HFst.prof$Colombian_Tuscan,
               Ref=RefLevel)

## Add a threshold

Threshold <- 3*RefLevel
HudsonFst.plot(Info,HFst.m$Colombian_Tuscan,HFst.prof$Colombian_Tuscan,
               Ref=RefLevel,Threshold = Threshold)
```

HudsonFst.prof *HudsonFst.prof*

Description

Perform FST profiling between pairs of pops, as requested by Contrast. If no contrast is provided, all pairs are considered

Usage

```
HudsonFst.prof(
  HFst.m,
  Contrast = NULL,
  Kmax = 100,
  NbSegCrit = "biggest.S3IB",
  parallel = TRUE
)
```

Arguments

HFst.m	A list of data.frame with two columns each, Fst and Weight, as provided by the HudsonFst.m function
Contrast	a list of two vectors with the names of the populations to be contrasted
Kmax	maximum number of breakpoints to be considered
NbSegCrit	the criterion used for the choice of the number of segments
parallel	a boolean, should the profiling be parallelized (using future) or not

Value

a smoothed profile

Examples

```
data(Freq);data(NbGamete)
FreqNbG <- BuildFreqNbG(Freq,NbGamete)
HFst.m <- HudsonFst.m(FreqNbG)

## Two population analysis
TwoPops <- list(First="Colombian",Second="Tuscan")
HFst.prof <- HudsonFst.prof(HFst.m,Contrast=TwoPops)

## The full example execution takes a few seconds.
## Two sets of populations to contrast

Contrast <- list(America=c("Colombian","Maya"),Europe=c("Tuscan","Italian"))
Profiles <- HudsonFst.prof(HFst.m,Contrast=Contrast)
```

```
## For larger lists and/or larger marker sets,
## use the future package for parallel computation:

future::plan("multisession",workers=4)
Profiles <- HudsonFst.prof(HFst.m,Contrast=Contrast)
future::plan("default")
```

Info *Marker information*

Description

Info is a data.frame describing the markers located on chromosome 1 in the Stanford HGDP dataset. Each row corresponds to a marker, and the 5 columns provide information about the marker name, its chromosome membership, its position, and its reference and alternative alleles.

Usage

```
Info
```

Format

A data.frame

MakeProfile.op *MakeProfile.op*

Description

Perform segmentation on a given dataset and returns the segmented profile

Usage

```
MakeProfile.op(DF, coef.pen.value = 1, sd.y = NULL)
```

Arguments

DF	a data.frame with two columns, Fst and Weights, as provided by the HudsonFst.m function
coef.pen.value	coef to use for penaly 2*coef.pen.value*log(n)
sd.y	a numeric value corresponding to the (estimated) standard deviation of the signal. If NULL (default) the value is automatically estimated.

Value

a smoothed profile

MakeProfile.sn_nomemory *MakeProfile.sn_nomemory*

Description

Perform segmentation on a given dataset and returns the segmented profile

Usage

```
MakeProfile.sn_nomemory(DF, Kmax, method = "biggest.S3IB", sd.y = NULL)
```

Arguments

`DF`	a data.frame with two columns, Fst and Weights, as provided by the `HudsonFst.m` function
`Kmax`	max number of changes used for model selection (check with crops)
`method`	a string, the name of the criterion used for model selection: (1) "givenVariance" = using the penalty of Lebarbier 2005 given a estimator of the variance, (2) "biggest.S3IB" = biggest=TRUE in saut taken from S3IB, (3) "notbiggest.S3IB" To be chosen amongs3ib.jump, s3ib.nojump, pre, ddse (capushe) or jump (capushe)
`sd.y`	a numeric value corresponding to the (estimated) standard deviation of the signal. If NULL (default) the value is automatically estimated.

Value

a smoothed profile

MergeRegion *MergeRegion*

Description

Merge adjacent top regions

Usage

```
MergeRegion(DT, Crit)
```

Arguments

`DT`	a data.frame
`Crit`	a string corresponding to the name of the criterion used for selecting the top regions

Value

a simplified data.frame (tibble)

`NbGamete` *Number of gametes of the American and European HGDP populations*

Description

`NbGamete` is a data.frame containing the number of gametes collected for the 13 American and European populations described in the Stanford HGDP dataset.

Usage

```
NbGamete
```

Format

A data.frame

`Plot1Prof` *Plot1Prof*

Description

Display the graphical representation of an Fst profile

Usage

```
Plot1Prof(DF, Title = "", Range = NULL)
```

Arguments

`DF` a data.frame, as provided by function `DF4Plot1Prof`

`Title` a string providing a title for the graph, optional.

`Range` a vector with the minimum and maximum values for the y-axis

Value

a ggplot object

ProfilingSummary *ProfilingSummary*

Description

Summary of Fst profiles

Usage

```
ProfilingSummary(FstProfiles, SnpInfo)
```

Arguments

`FstProfiles`	a list of Fst profiles as obtained from function `HudsonFst.prof`
`SnpInfo`	a data.frame providing information about markers

Value

a list of data.frame. Each data.frame summarizes a Fst profile, in terms of number of segments, Start and End positions, length (i.e. number of markers) and Fst level of each segment.

Examples

```
data(Freq);data(NbGamete);data(Info)
FreqNbG <- BuildFreqNbG(Freq,NbGamete)
HFst.m <- HudsonFst.m(FreqNbG)
TwoPops <- list(First="Colombian",Second="Tuscan")
HFst.prof <- HudsonFst.prof(HFst.m,Contrast=TwoPops)
PS <- ProfilingSummary(HFst.prof,Info)
```

Ratio_Average *Computation of the Fst moment estimator*

Description

Computation of the Fst moment estimator

Usage

```
Ratio_Average(Nominator, Denominator)
```

Arguments

`Nominator`	numeric, numerators of the Fst moment estimator
`Denominator`	numeric, denominators of the Fst moment estimator

Value

a vector with the global Fst estimator

`RawPlot` *RawPlot*

Description

Plot the Fst estimates along a (portion of) chromosome

Usage

```
RawPlot(Info, HF, Coord = NULL, Title = "")
```

Arguments

`Info` a data.frame providing information about markers

`HF` a data.frame with 2 columns, Fst and Weight, as provided by the `HudsonFst.m` function

`Coord` a vector with the minimum and maximum coordinates (i.e. positions along the genome), providing the range of the genomic region that will be plotted.

`Title` a string providing a title for the graph.

Value

a ggplot object.

`Summarise1Profile` *Summarise1Profile*

Description

Summarise1Profile

Usage

```
Summarise1Profile(profile, snpinfo)
```

Arguments

`profile` a vector, corresponding to the Fst profile of a pair of populations

`snpinfo` a data.frame, providing information about markers

Value

a data.frame that combines both objects

Index