Package: twigstats (via r-universe)

October 28, 2024

Type Package

Title twigstats

Version 1.0.1

Encoding UTF-8

Date 2022-02-02

Maintainer Leo Speidel <leo.speidel@outlook.com>

Description This package takes Relate genealogies as input to compute time-stratified f-statistics.

License MIT

SystemRequirements zlib

Imports Rcpp (>= 1.0.7), RcppArmadillo

LinkingTo testthat, Rcpp, RcppArmadillo

URL https://github.com/leospeidel/twigstats

BugReports https://github.com/leospeidel/twigstats/issues

Suggests xml2, testthat (>= 3.0.0), knitr, rmarkdown, tidyr, lsei

Config/testthat/edition 3

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

Depends R (>= 3.6.0), dplyr

VignetteBuilder knitr

Repository https://leospeidel.r-universe.dev

RemoteUrl https://github.com/leospeidel/twigstats

RemoteRef HEAD

RemoteSha aeb8552450ae2dd7ab7163c5dffcf95e99c43c41

Contents

`f2_blocks_from_Relate` *Function to calculate f2 statistics from Relate trees for pairs of populations specified in poplabels.*

Description

This function will calculate f2 statistics in blocks of prespecified size for all pairs of populations specified in the poplabels file. Please refer to the Relate documentation for input file formats (https://myersgroup.github.io/relate/). The output is in a format that is directly accepted by the admixtools R package to calculate f3, f4, f4ratio, D statistics and more (https://uqrmaie1.github.io/admixtools/).

Usage

```
f2_blocks_from_Relate(
  file_anc,
  file_mut,
  poplabels,
  file_map = NULL,
  chrs = NULL,
  blgsize = NULL,
  mu = NULL,
  tmin = NULL,
  t = NULL,
  transitions = NULL,
  use_muts = NULL,
  minMAF = NULL,
  dump_blockpos = NULL,
  apply_corr = NULL
)
```

Arguments

`file_anc` Filename of anc file. If chrs is specified, this should only be the prefix, resulting in filenames of ${file_anc}_chr${chr}.anc(.gz).

`file_mut` Filename of mut file. If chrs is specified, this should only be the prefix, resulting in filenames of ${file_anc}_chr${chr}.anc(.gz).

`poplabels` Filename of poplabels file

`file_map` (Optional) File prefix of recombination map. Not needed if blgsize is given in base-pairs, i.e. blgsize > 100

`chrs` (Optional) Vector of chromosome IDs

`blgsize` (Optional) SNP block size in Morgan. Default is 0.05 (5 cM). If blgsize is 100 or greater, if will be interpreted as base pair distance rather than centimorgan distance. If blgsize is negative, every tree is its own block.

`mu` (Optional) Per base per generation mutation rate to scale f2 values. Default: 1.25e-8

`tmin` (Optional) Minimum time cutof in generations. Any lineages younger than tmin will be excluded from the analysis. Default: t = 0.

`t` (Optional) Time cutoff in generations. Default: Inf

`transitions` (Optional) Set this to FALSE to exclude transition SNPs. Only meaningful with use_muts

`use_muts` (Optional) Calculate traditional f2 statistics by only using mutations mapped to Relate trees. Default: false.

`minMAF` (Optional) Minimum frequency cutoff. Default: 1 (i.e. excl singletons)

`dump_blockpos` (Optional) Filename of blockpos file.

`apply_corr` (Optional) Use small sample size correction. Default: true.

Value

3d array of dimension #groups x #groups x #blocks. Analogous to output of f2_from_geno in admixtools.

Examples

```
file_anc  <- system.file("sim/msprime_ad0.8_split250_1_chr1.anc.gz", package = "twigstats")
file_mut  <- system.file("sim/msprime_ad0.8_split250_1_chr1.mut.gz", package = "twigstats")
poplabels <- system.file("sim/msprime_ad0.8_split250_1.poplabels", package = "twigstats")
file_map  <- system.file("sim/genetic_map_combined_b37_chr1.txt.gz", package = "twigstats")

#Calculate f2s between all pairs of populations
f2_blocks <- f2_blocks_from_Relate(file_anc, file_mut, poplabels, file_map)
f4_ratio(f2_blocks, popX="PX", popI="P1", pop1="P2", pop2="P3", popO="P4")

#Use a cutoff of 500 generations
f2_blocks <- f2_blocks_from_Relate(file_anc, file_mut, poplabels, file_map, t = 500)
f4_ratio(f2_blocks, popX="PX", popI="P1", pop1="P2", pop2="P3", popO="P4")
```

f2_blocks_from_RelateAges

Function to calculate f2 statistics from plink files, ascertained using mutation ages in Relate trees.

Description

This function will calculate f2 statistics in blocks of prespecified size for all pairs of populations specified in the input files. Input is assumed to be in PLINK binary format https://www.cog-genomics.org/plink/1.9/formats#bed and mutation ages are in Relate mut format https://myersgroup.github.io/relate/. The output is in a format that is directly accepted by the admixtools R package to calculate f2, f3, f4, f4ratio, D statistics and more (https://uqrmaie1.github.io/admixtools/).

Usage

```
f2_blocks_from_RelateAges(
  pref,
  file_mut,
  blgsize = NULL,
  transitions = NULL,
  maxmiss = NULL,
  fam = NULL,
  pops = NULL,
  chrs = NULL,
  tmin = NULL,
  t = NULL,
  include_undated = NULL,
  minMAF = NULL,
  apply_corr = NULL,
  debug_mode = 0L
)
```

Arguments

`pref` Prefix of PLINK binary files, assuming filenames of form ${pref}.bed, ${pref}.bim, ${pref}.fam.

`file_mut` (Optional) Prefix of filenames of mut files, assuming filenames of form ${file_mut}_chr1.mut(.gz). Chromosome names have to be consistent to those in the PLINK files. If no file is specified, all mutations in the PLINK file are used.

`blgsize` (Optional) SNP block size in Morgan. Default is 0.05 (5 cM). If blgsize is 100 or greater, if will be interpreted as base pair distance rather than centimorgan distance.

`transitions` (Optional) Set this to FALSE to exclude transition SNPs

`maxmiss` (Optional) Discard SNPs which are missing in a fraction of populations higher than maxmiss `fam` (Optional) 1d-array assigning individuals to populations. Corresponds to the first column in the fam file and is useful if you want to change population assignments.

`pops` (Optional) Populations for which data should be extracted. Names need to match the first column in the fam file (or fam option below)

`chrs` (Optional) List chromosome names to use.

`tmin` (Optional) Minimum time cutof in generations. Any mutations younger than tmin will be excluded from the analysis. Default: t = 0.

`t` (Optional) Time cutoff in generations. Any mutations older that t will be excluded from the analysis. Default: t = Inf.

`include_undated`
(Optional) Include mutations that are not dated. Default: false.

`minMAF` (Optional) minimum minor allele count. Default: 1.

`apply_corr` (Optional) Use small sample size correction. Default: true.

`debug_mode` (Optional) Prints progress used for debugging.

Value

3d array of dimension #groups x #groups x #blocks containing f2 statistics. Analogous to output of f2_from_geno in admixtools.

Examples

```
path <- paste0(system.file("sim/", package = "twigstats"),"/")
pref <- "msprime_ad0.8_split250_1"
file_plink <- paste0(path,pref) #only need prefix
file_mut  <- paste0(path,pref) #only need prefix (here same name as plink file but can be different)

system(paste0("gunzip ", file_plink, ".bim.gz"))

#Compute f2 statistics between all pairs of populations. You can use pops to only calculate f2s between specified po
f2_blocks <- f2_blocks_from_RelateAges(pref = file_plink, file_mut)
f4_ratio(f2_blocks, popX="PX", popI="P1", pop1="P2", pop2="P3", popO="P4")

#Use a cutoff of 500 generations
f2_blocks <- f2_blocks_from_RelateAges(pref = file_plink, file_mut, t = 500)
f4_ratio(f2_blocks, popX="PX", popI="P1", pop1="P2", pop2="P3", popO="P4")
```

`f4_ratio` *Function implementing the F4-ratio statistic*

Description

This function computes the admixture proportion given five populations. A population history following ((PI,P1),P2,PO) is assumed, and the target is assumed to be a mixture of proximal sources P1 and P2, i.e. PX = alpha*P2 + (1-alpha)*P1

This function implements a jackknife on the input data.

Usage

```
f4_ratio(f2_blocks, pop0, popI, pop1, pop2, popX, mode = 1)

jackknife(df_jack)
```

Arguments

f2_blocks	A 3d array of blocked f2 statistics
pop0	Name of outgroup population
popI	Name of ingroup population
pop1	Name of source that clusters with ingroup
pop2	Name of other source
popX	Name of target group.
data	frame. Three columns called blockID, hj, Dj, storing block ID, weight of block, and statistic without that block

Value

Returns a data frame with admixture proportion estimates and jacknifed standard errors.

Returns a data table with columns val and se.

Examples

```
#These lines assign file names to variables file_anc, file_mut, poplabels, file_map.
#see https://myersgroup.github.io/relate/getting_started.html#Output for file formats
file_anc  <- system.file("sim/msprime_ad0.8_split250_1_chr1.anc.gz", package = "twigstats")
file_mut  <- system.file("sim/msprime_ad0.8_split250_1_chr1.mut.gz", package = "twigstats")
#see https://myersgroup.github.io/relate/input_data.html for file formats
poplabels <- system.file("sim/msprime_ad0.8_split250_1.poplabels", package = "twigstats")
file_map  <- system.file("sim/genetic_map_combined_b37_chr1.txt.gz", package = "twigstats") #recombination map (t

#Calculate regular f2s between all pairs of populations
f2_blocks1 <- f2_blocks_from_Relate(file_anc = file_anc, file_mut = file_mut, poplabels = poplabels, file_map = fil
f4_ratio(f2_blocks1, popX="PX", popI="P1", pop1="P2", pop2="P3", pop0="P4")

#Use a twigstats cutoff of 500 generations
f2_blocks2 <- f2_blocks_from_Relate(file_anc = file_anc, file_mut = file_mut, poplabels = poplabels, file_map = fil
f4_ratio(f2_blocks2, popX="PX", popI="P1", pop1="P2", pop2="P3", pop0="P4")
```

Fst_blocks_from_Relate

Function to calculate Fst from Relate trees for pairs of populations specified in poplabels.

Description

This function will calculate Fst in blocks of prespecified size for all pairs of populations specified in the poplabels file. Please refer to the Relate documentation for input file formats (https://myersgroup.github.io/relate/). The output is in the same format as for f2_blocks_from_Relate.

Usage

```
Fst_blocks_from_Relate(
  file_anc,
  file_mut,
  poplabels,
  file_map = NULL,
  chrs = NULL,
  blgsize = NULL,
  mu = NULL,
  tmin = NULL,
  t = NULL,
  transitions = NULL,
  use_muts = NULL,
  minMAF = NULL,
  Fst = NULL,
  dump_blockpos = NULL,
  apply_corr = NULL
)
```

Arguments

`file_anc` Filename of anc file. If chrs is specified, this should only be the prefix, resulting in filenames of ${file_anc}_chr${chr}.anc(.gz).

`file_mut` Filename of mut file. If chrs is specified, this should only be the prefix, resulting in filenames of ${file_anc}_chr${chr}.anc(.gz).

`poplabels` Filename of poplabels file

`file_map` (Optional) File prefix of recombination map. Not needed if blgsize is given in base-pairs, i.e. blgsize > 100

`chrs` (Optional) Vector of chromosome IDs

`blgsize` (Optional) SNP block size. Default is 10000 bases. If blgsize is <100 then it will be interpreted in Morgan. If blgsize is 100 or greater, if will be interpreted as base pair distance. If blgsize is negative, every tree is its own block.

`mu` (Optional) Per base per generation mutation rate to scale f2 values. Default: 1.25e-8

`tmin` (Optional) Minimum time cutof in generations. Any lineages younger than tmin will be excluded from the analysis. Default: t = 0.

`t` (Optional) Time cutoff in generations. Default: Inf

`transitions` (Optional) Set this to FALSE to exclude transition SNPs. Only meaningful with use_muts `use_muts` (Optional) Calculate traditional f2 statistics by only using mutations mapped to Relate trees. Default: false.

`minMAF` (Optional) Minimum frequency cutoff. Default: 1 (i.e. excl singletons)

`dump_blockpos` (Optional) Filename of blockpos file.

`apply_corr` (Optional) Use small sample size correction. Default: true.

Value

3d array of dimension #groups x #groups x #blocks. Analogous to output of f2_from_geno in admixtools.

Examples

```
file_anc  <- system.file("sim/msprime_ad0.8_split250_1_chr1.anc.gz", package = "twigstats")
file_mut  <- system.file("sim/msprime_ad0.8_split250_1_chr1.mut.gz", package = "twigstats")
poplabels <- system.file("sim/msprime_ad0.8_split250_1.poplabels", package = "twigstats")
file_map  <- system.file("sim/genetic_map_combined_b37_chr1.txt.gz", package = "twigstats")

#Calculate f2s between all pairs of populations
Fst_blocks <- Fst_blocks_from_Relate(file_anc, file_mut, poplabels, file_map)

#Use a cutoff of 500 generations
Fst_blocks <- Fst_blocks_from_Relate(file_anc, file_mut, poplabels, file_map, dump_blockpos = "test.pos", t = 500)
```

`Painting` *Chromosome painting using genealogies.*

Description

This function outputs the first coalescence with an individual from a pre-specified group identity along the genome. If the first such coalescnece involves several copying candidates, a random haplotype is chosen. Output is in GLOBEtrotter format.

Usage

```
Painting(
  file_anc,
  file_mut,
  file_map,
  file_out,
  poplabels,
  blgsize = NULL,
  pops = NULL,
  chrs = NULL
)
```

Arguments

`file_anc` Filename of anc file. If chrs is specified, this should only be the prefix, resulting in filenames of ${file_anc}_chr${chr}.anc(.gz).

`file_mut` Filename of mut file. If chrs is specified, this should only be the prefix, resulting in filenames of ${file_anc}_chr${chr}.anc(.gz).

`file_map` File prefix of recombination map.

`file_out` File prefix of output files

`poplabels` Filename of poplabels file

`blgsize` (Optional) SNP block size in Morgan. Default is 0.05 (5 cM). If blgsize is 1 or greater, if will be interpreted as base pair distance rather than centimorgan distance.

`pops` (Optional) Populations for which data should be extracted. Names need to match the second column of the poplabels file

`chrs` (Optional) Vector of chromosome IDs

Value

void. Write three files idfile, paint, rec to disc.

Examples

```
file_anc  <- system.file("sim/msprime_ad0.8_split250_1_chr1.anc.gz", package = "twigstats")
file_mut  <- system.file("sim/msprime_ad0.8_split250_1_chr1.mut.gz", package = "twigstats")
poplabels <- system.file("sim/msprime_ad0.8_split250_1.poplabels", package = "twigstats")
file_map  <- system.file("sim/genetic_map_combined_b37_chr1.txt.gz", package = "twigstats")

#define populations to paint against:
pops <- c("P1","P2","P3","P4")

Painting(file_anc, file_mut, file_map, file_out = "test", poplabels, blgsize = 1e-5)
```

PaintingNNLS

Function to run an NNLS on genome-wide copying proportions to estimate admixture proportions.

Description

This function computes admixture proprtions using a non-negative least squares on painting profiles obtained from the function PaintingProfile

Usage

```
PaintingNNLS(df_sum, target_pops, source_pops = NULL)
```

Arguments

`df_sum` Output of PaintingProfile

`target_pops` Vector of populations to be fitted

`source_pops` (Optional) Vector of putative source populations. If not provided, all remaining populations are used.

Value

Returns a data frame with admixture proportions.

Examples

```
library(dplyr)
if (!requireNamespace(”tidyr”, quietly = TRUE)) {
  stop(”This example needs 'tidyr'. Please install it.”, call. = FALSE)
}
if (!requireNamespace(”lsei”, quietly = TRUE)) {
  stop(”This example needs 'lsei'. Please install it.”, call. = FALSE)
}

#This path stores files precomputed using Painting().
path      <- paste0(system.file(”test/”, package = ”twigstats”),”/”)
prefix    <- ”test” #prefix of files under path

#compute the painting profiles with 100 bootstrap samples and a blocksize of 5cM (5000*0.001)
df <- PaintingProfile(c(paste0(path,prefix,”_painting.txt.gz”)), paste0(path,prefix,”_idfile.txt.gz”), nboot = 1

#compute the NNLS to get admixture proportions
df <- PaintingNNLS(df, target_pops = c(”PX”), source_pops = c(”P2”,”P3”))

#Now you can summarize the bootstrap samples
df %>% group_by(target, POP) %>% summarize(mean_ancestry = mean(ancestry), sd_ancestry = sd(ancestry)) -> df
print(df)
```

PaintingProfile *Function to compute genome-wide copying proportions.*

Description

This function takes the output of the function Painting and computes the genome-wide ’copying vectors’, i.e. the proportion of the genome copied from each other reference population..

Usage

```
PaintingProfile(
  filename_painting,
  filename_idfile,
  nboot,
  blocksize,
  use_IDs = FALSE
)
```

Arguments

`filename_painting`
: Vector containing filenames of painting profiles. Output of Painting.

`filename_idfile`
: Filename of idfile. Output of Painting.

`nboot`
: Number of bootstrap samples.

`blocksize`
: Number of blocks to combine for the bootstrap. E.g. if Painting was run with a blgsize of 1e-5 Morgans, blocksize should be 5000 to achieve a blocksize of 5cM.

`use_IDs`
: If TRUE, compute profile for each sample. If FALSE (default), compute profile for each group as specified in the second column of the idfile.

Value

Returns a data frame with copying proportions per bootstrap sample.

Examples

```
#This path stores files precomputed using Painting().
path      <- paste0(system.file("test/", package = "twigstats"),"/")
prefix    <- "test" #prefix of files under path

#compute the painting profiles with 10 bootstrap samples and a blocksize of 5cM (5000*0.001)
df <- PaintingProfile(c(paste0(path,prefix,"_painting.txt.gz")), paste0(path,prefix,"_idfile.txt.gz"), nboot = 1
head(df)
```

theoretical_zscore   *Function to compute a theoretical z-score given source split times and admixture proportions.*

Description

This function computes the theoretical z-score of an f4-statistic of the form f4(PO,P2,PX,P1) as a function of the admixture time a, source (P1 & P2) split time s, and admixture proportion alpha. It assumes a single lineage is sampled from each population.

Usage

```
theoretical_zscore(t, a, s, alpha)
```

Arguments

`t`	Twigstats cutoff time in units of 2Ne generations
`a`	Admixture time in units of 2Ne generations
`s`	Source split time in units of 2Ne generations
`alpha`	Admixture proportions. Proportion of P2 in PX.

Value

Returns the theoretical z-score value.

`TwigScan` *Function implementing TwigScan*

Description

This function computes f2s or Fst in blocks along the genome and returns a data frame with columns blockID, pos, pop1, pop2, and the f2 or Fst values.

Usage

```
TwigScan(
  file_anc,
  file_mut,
  poplabels,
  file_map,
  file_out,
  blgsize = 10000,
  t = Inf,
  use_muts = FALSE,
  Fst = FALSE
)
```

Arguments

`file_anc`	Filename of anc file. If chrs is specified, this should only be the prefix, resulting in filenames of ${file_anc}_chr${chr}.anc(.gz).
`file_mut`	Filename of mut file. If chrs is specified, this should only be the prefix, resulting in filenames of ${file_anc}_chr${chr}.anc(.gz).
`poplabels`	Filename of poplabels file
`file_map`	File prefix of recombination map.
`file_out`	File prefix of output files
`blgsize`	(Optional) SNP block size in Morgan. Default is 0.05 (5 cM). If blgsize is 1 or greater, if will be interpreted as base pair distance rather than centimorgan distance.

`t` (Optional) Time cutoff in generations. Default: Inf

`use_muts` (Optional) Calculate traditional f2 statistics by only using mutations mapped to Relate trees. Default: False.

`Fst` (Optional) If TRUE, compute Fst. Default: FALSE

Value

Returns a data frame with Fst or f2 values.

Examples

```
#These lines assign file names to variables file_anc, file_mut, poplabels, file_map.
#see https://myersgroup.github.io/relate/getting_started.html#Output for file formats
file_anc  <- system.file(”sim/msprime_ad0.8_split250_1_chr1.anc.gz”, package = ”twigstats”)
file_mut  <- system.file(”sim/msprime_ad0.8_split250_1_chr1.mut.gz”, package = ”twigstats”)
#see https://myersgroup.github.io/relate/input_data.html for file formats
poplabels <- system.file(”sim/msprime_ad0.8_split250_1.poplabels”, package = ”twigstats”)
file_map  <- system.file(”sim/genetic_map_combined_b37_chr1.txt.gz”, package = ”twigstats”) #recombination map (t

df <- TwigScan(file_anc = file_anc,
     file_mut = file_mut,
     poplabels = poplabels,
     file_map = file_map,
     file_out = ”test”,
     blgsize = 10000,  #optional
     use_muts = FALSE, #optional
     t = 1000,        #optional
     Fst = TRUE        #optional
 )

print(head(df))
```

Index