

Package ‘HACSim’

January 7, 2020

Type Package

Title Iterative Extrapolation of Species' Haplotype Accumulation Curves for Genetic Diversity Assessment

Version 1.0.5

Date 2019-12-21

Author Jarrett D. Phillips [aut, cre], Steven H. French [ctb]

Maintainer Jarrett D. Phillips <phillipsjarrett1@gmail.com>

Description Performs iterative extrapolation of species' haplotype accumulation curves using a non-parametric stochastic (Monte Carlo) optimization method for assessment of specimen sampling completeness based on the approach of Phillips et al. (2015) <doi:10.1515/dna-2015-0008>, Phillips et al. (2019) <doi:10.1002/ece3.4757> and Phillips et al. (2020) <doi: 10.7717/peerj-cs.243>. Any genomic marker can be targeted to assess likely required specimen sample sizes. The method is well-suited to assess sampling sufficiency for DNA barcoding initiatives.

License GPL-3

NeedsCompilation yes

Imports ape (>= 5.2), graphics (>= 3.5.1), pegas (>= 0.11), Rcpp (>= 1.0.0), stats (>= 3.5.1), utils (>= 3.5.1)

LinkingTo Rcpp, RcppArmadillo

RoxygenNote 6.1.1

Repository CRAN

Date/Publication 2020-01-07 21:50:05 UTC

R topics documented:

HACSim-package	2
accumulate	4
envr	4
HAC.sim	5
HAC.simrep	6
HACClass	8
HACHypothetical	8
HACReal	10

HACSim-package	<i>Iterative Extrapolation of Species' Haplotype Accumulation Curves for Genetic Diversity Assessment</i>
----------------	---

Description

HACSim (**H**aplotype **A**ccumulation **C**urve **S**imulator) employs a novel nonparametric stochastic (Monte Carlo) optimization method of iteratively generating species' haplotype accumulation curves through extrapolation to assess sampling completeness based on the approach outlined in Phillips et al. (2015) <doi:10.1515/dna-2015-0008> and Phillips et al. (2019) <doi:10.1002/ece3.4757>. Any genomic marker can be targeted to assess likely required specimen sample sizes. The method is well-suited to assess sampling sufficiency for DNA barcoding initiatives. The package outputs a number of useful summary statistics of sampling coverage ("Measures of Sampling Closeness"), including an estimate of the likely required sample size (along with desired level confidence intervals) necessary to recover a given number/proportion of observed unique species' haplotypes.

Details

The DESCRIPTION file:

```
Package:          HACSim
Type:             Package
Title:            Iterative Extrapolation of Species' Haplotype Accumulation Curves for Genetic Diversity Assessment
Version:          1.0.5
Date:             2019-12-21
Author:           Jarrett D. Phillips [aut, cre], Steven H. French [ctb]
Maintainer:      Jarrett D. Phillips <phillipsjarrett1@gmail.com>
Description:      Performs iterative extrapolation of species' haplotype accumulation curves using a nonparametric stochastic method
License:          GPL-3
NeedsCompilation: yes
Imports:          ape (>= 5.2), graphics (>= 3.5.1), pegas (>= 0.11), Rcpp (>= 1.0.0), stats (>= 3.5.1), utils (>= 3.5.1)
LinkingTo:       Rcpp, RcppArmadillo
RoxygenNote:     6.1.1
Packaged:        2019-10-23 16:00:58 UTC; jarrettphillips
```

Index of help topics:

```
HAC.sim           Internal R code
HAC.simrep        Run a simulation of haplotype accumulation
                  curves for hypothetical or real species
HACClass          Internal R code
HACHypothetical   Set up an object to simulate haplotype
                  accumulation curves for a hypothetical species
HACReal           Set up an object to simulate haplotype
                  accumulation curves for a real species
```

HACSim-package	Iterative Extrapolation of Species' Haplotype Accumulation Curves for Genetic Diversity Assessment
accumulate	Internal C++ code
envr	Simulation variable storage environment

Author(s)

Jarrett D. Phillips [aut, cre], Steven H. French [ctb]

Maintainer: Jarrett D. Phillips <phillipsjarrett1@gmail.com>

References

Phillips, J.D., Gwiazdowski, R.A., Ashlock, D. and Hanner, R. (2015). An exploration of sufficient sampling effort to describe intraspecific DNA barcode haplotype diversity: examples from the ray-finned fishes (Chordata: Actinopterygii). *DNA Barcodes*, 3: 66-73.

Phillips, J.D., Gillis, D.J. and Hanner, R.H. (2019). Incomplete estimates of genetic diversity within species: Implications for DNA barcoding. *Ecology and Evolution*, 9(5): 2996-3010.

Phillips, J.D., Gillis, D.J. and Hanner, R.H. (2020). HACSim: An R package to estimate intraspecific sample sizes for genetic diversity assessment using haplotype accumulation curves. *PeerJ Computer Science*

Examples

```
## Simulate hypothetical species ##

N <- 100 # total number of sampled individuals
Hstar <- 10 # total number of haplotypes
probs <- rep(1/Hstar, Hstar) # equal haplotype frequency distribution

HACSObj <- HACHypothetical(N = N, Hstar = Hstar,
probs = probs, filename = "output") # outputs a CSV
# file called "output.csv"

## Simulate hypothetical species - subsampling ##
HACSObj <- HACHypothetical(N = N, Hstar = Hstar,
probs = probs, perms = 1000, p = 0.95,
subsample = TRUE, prop = 0.25, conf.level = 0.95,
filename = "output")

## Simulate hypothetical species and all parameters changed - subsampling ##
HACSObj <- HACHypothetical(N = N, Hstar = Hstar, probs = probs,
perms = 1000, p = 0.90, subsample = TRUE, prop = 0.15,
conf.level = 0.95, filename = "output")

HAC.simrep(HACSObj) # runs a simulation

## Simulate real species ##
```

```

## Simulate real species ##
# outputs file called "output.csv"
HACSObj <- HACReal(filename = "output")

## Simulate real species - subsampling ##
HACSObj <- HACReal(subsample = TRUE, prop = 0.15,
  conf.level = 0.95, filename = "output")

## Simulate real species and all parameters changed - subsampling ##
HACSObj <- HACReal(perms = 10000, p = 0.90, subsample = TRUE,
  prop = 0.15, conf.level = 0.99, filename = "output")

# user prompted to select appropriate FASTA file
HAC.simrep(HACSObj)

```

accumulate	<i>Internal C++ code</i>
------------	--------------------------

Description

accumulate comprises internal C++ code employed by HAC.sim. It is not directly called by the user.

envr	<i>Simulation variable storage environment</i>
------	--

Description

envr is a new (initially empty) environment that is created when HACSim is loaded.

Value

When a simulation is run via HAC.simrep, envr will contain 20 elements as follows:

conf.level	The desired confidence level. Default is conf.level = 0.95.
d	A dataframe with three columns: specimens (specs), accumulated haplotypes (means), and standard deviation (sd)
filename	The name of the file where results are to be saved. Default is NULL.
high	The upper endpoint of the desired level confidence interval for the 'true' required sample size
Hstar	Number of unique species' haplotypes
input.seqs	Should DNA sequences be inputted? Default is FALSE.

<code>iters</code>	The number of iterations required to reach convergence
<code>low</code>	The upper endpoint of the desired level confidence interval for the 'true' required sample size
<code>N</code>	The starting sample size used to initialize the algorithm
<code>Nstar</code>	The final (extrapolated) sample size
<code>num.iters</code>	Number of iterations to compute. <code>num.iters = NULL</code> by default (i.e., all iterations are computed; users can specify <code>num.iters = 1</code> for the first iteration.)
<code>p</code>	The user-specified level of haplotype recovery. Default is <code>p = 0.95</code> .
<code>perms</code>	The user-specified number of permutations (replications). Default is <code>perms = 10000</code> .
<code>probs</code>	Haplotype frequency distribution vector
<code>progress</code>	Should iteration results be outputted to the console? Default is <code>TRUE</code> .
<code>prop.haps</code>	If <code>subset.haps = TRUE</code> , the user-specified proportion of haplotype labels to recover
<code>prop.seqs</code>	If <code>subset.seqs = TRUE</code> , the user-specified proportion of DNA sequences to recover
<code>ptm</code>	A timer to track progress of the algorithm in seconds
<code>R</code>	The proportion of haplotypes recovered by the algorithm
<code>subset.haps</code>	Should a subsample of haplotype labels be taken? Default is <code>FALSE</code> .
<code>subset.seqs</code>	Should a subsample of DNA sequences be taken? Default is <code>FALSE</code> .

Examples

```
# Returns the frequencies of each haplotype in the extrapolated sample
max(envr$d$specs) * envr$probs

# Returns the extrapolated sample size corresponding to the dotted line
# in the last iteration plot
envr$d[which(envr$d$means >= envr$p * envr$Hstar), ][1, 1]
```

HAC.sim

Internal R code

Description

HAC.sim comprises internal R code used by HAC.simrep and is not directly called by the user.

HAC.simrep	<i>Run a simulation of haplotype accumulation curves for hypothetical or real species</i>
------------	---

Description

Runs the HACSIm algorithm by successively calling HAC.sim to iteratively extrapolate haplotype accumulation curves to determine likely specimen sample sizes for hypothetical or real species

The algorithm employs the following iterative methods when calculating the "Measures of Sampling Closeness":

- Mean number of haplotype sampled: H_i
- Mean number of haplotypes not sampled: $H^* - H_i$
- Proportion of haplotypes sampled: $\frac{H_i}{H^*}$
- Proportion of haplotypes not sampled: $1 - \frac{H_i}{H^*}$
- Mean value of N^* : $\frac{N_i H^*}{H_i}$
- Mean number of specimens not sampled: $\frac{N_i H^*}{H_i} - N_i$

where H_i is stochastically-determined through sampling from probs, the observed species' haplotype frequency distribution vector.

As the algorithm proceeds, H_i will approach H^* asymptotically (and hence, N_i will converge to N^*), but will likely fluctuate randomly from one iteration to the next. However, estimates of N^* found at each iteration will be monotonically-increasing.

Usage

```
HAC.simrep(HACSObject)
```

Arguments

HACSObject object containing the desired simulation parameters

Value

Iteration results are outputted to the console and graphs displayed in the plot window. Plots depict haplotype accumulation (along with shaded confidence intervals for the mean number of haplotypes found). Dashed lines correspond to the endpoint of the curve and reflect haplotype recovery for a user-defined cutoff (default $p = 0.95$, 95% haplotype diversity). Output from the first iteration is useful for judging levels of haplotype diversity and recovery found in observed intraspecific sequence datasets, reflecting current sampling depth. The required sample size is displayed in the second-last iteration. All other information corresponding to the extrapolated sample size can be found in the last iteration. Iteration results can optionally be saved to a CSV file. Subsampled DNA sequences are automatically saved to a FASTA file.

Note

When simulating real species via `HACReal(...)`, a pop-up window will appear prompting the user to select an intraspecific FASTA file of aligned/trimmed DNA sequences. The alignment must not contain missing or ambiguous nucleotides (i.e., it should only contain A, C, G or T); otherwise, haplotype diversity may be overestimated. Excluding sequences or alignment sites with missing/ambiguous data is an option.

Examples

```
## Simulate hypothetical species ##

N <- 100 # total number of sampled individuals
Hstar <- 10 # total number of haplotypes
probs <- rep(1/Hstar, Hstar) # equal haplotype frequency distribution

HACSObj <- HACHypothetical(N = N, Hstar = Hstar , probs = probs,
filename = "output") # outputs a CSV file called "output.csv"

## Simulate hypothetical species - subsampling ##
HACSObj <- HACHypothetical(N = N, Hstar = Hstar, probs = probs,
perms = 1000, p = 0.95, subsample = TRUE, prop = 0.25,
conf.level = 0.95, filename = "output")

## Simulate hypothetical species and all paramaters changed - subsampling ##
HACSObj <- HACHypothetical(N = N, Hstar = Hstar, probs = probs,
perms = 10000, p = 0.90, subsample = TRUE, prop = 0.15, conf.level = 0.95,
filename = "output")

HAC.simrep(HACSObj) # runs a simulation

## Simulate real species ##

## Simulate real species ##
# outputs file called "output.csv"
HACSObj <- HACReal(filename = "output")

## Simulate real species - subsampling ##
HACSObj <- HACReal(subsample = TRUE, prop = 0.15, conf.level = 0.95,
filename = "output")

## Simulate real species and all parameters changed - subsampling ##
HACSObj <- HACReal(perms = 10000, p = 0.90, subsample = TRUE,
prop = 0.15, conf.level = 0.99, filename = "output")

# user prompted to select appropriate FASTA file
HAC.simrep(HACSObj)
```

HACClass	<i>Internal R code</i>
----------	------------------------

Description

HACClass comprises internal R code used to generate an object used by HAC.simrep. It is not directly called by the user.

HACHypothetical	<i>Set up an object to simulate haplotype accumulation curves for a hypothetical species</i>
-----------------	--

Description

Helper function which creates an object containing necessary information to run a simulation of haplotype accumulation for a hypothetical species of interest

Usage

```
HACHypothetical(N, Hstar, probs, perms = 10000, p = 0.95,
  conf.level = 0.95, subsample = FALSE, prop = NULL,
  progress = TRUE, num.iters = NULL, filename = NULL)
```

Arguments

N	Number of individuals
Hstar	Number of unique species' haplotypes
probs	Haplotype frequency distribution vector
perms	Number of permutations (replications)
p	Proportion of haplotypes to recover
conf.level	Desired confidence level for graphical output and interval estimation
subsample	Is a subsample of haplotype labels desired?
prop	If subsample = TRUE, the proportion of haplotype labels to subsample
num.iters	Number of iterations to compute
progress	Should iteration output be printed to the R console?
filename	Name of file where simulation results are to be saved

Value

A list object of class "HAC" with 13 elements that can be passed to `HAC.simrep` as follows:

<code>input.seqs</code>	Should a FASTA file of aligned/trimmed DNA sequences be inputted? Default is FALSE
<code>subset.seqs</code>	Should a subsample of DNA sequences be taken? Default is FALSE
<code>prop.seqs</code>	Proportion of DNA sequences to subsample. Default is NULL
<code>prop.haps</code>	Proportion of haplotype labels to subsample. Default is NULL (can be altered by user)
<code>subset.haps</code>	Should a subsample of haplotype labels be taken? Default is NULL (can be altered by user)
<code>N</code>	Number of individuals. NA by default (provided by user)
<code>Hstar</code>	Number of unique species' haplotypes. NA by default (provided by user)
<code>probs</code>	Haplotype frequency distribution vector. NA by default (provided by user)
<code>p</code>	Proportion of haplotypes to recover. $p = 0.95$ by default.
<code>perms</code>	Number of permutations (replications). <code>perms = 10000</code> by default.
<code>conf.level</code>	Desired confidence level for graphical output and interval estimation. <code>conf.level = 0.95</code> by default.
<code>num.iters</code>	Number of iterations to compute. <code>num.iters = NULL</code> by default (i.e., all iterations are computed; users can specify <code>num.iters = 1</code> for the first iteration.)
<code>progress</code>	Should iteration output be printed to the R console? Default is TRUE
<code>filename</code>	Name of file where simulation results are to be saved.

Note

`N` must be greater than 1 and greater than or equal to `Hstar`.

`Hstar` must be greater than 1.

`probs` must have a length equal to `Hstar` and its elements must sum to 1.

Examples

```
## Simulate hypothetical species ##

N <- 100 # total number of sampled individuals
Hstar <- 10 # total number of haplotypes
probs <- rep(1/Hstar, Hstar) # equal haplotype frequency distribution

# outputs a CSV file called "output.csv"
HACSobj <- HACHypothetical(N = N, Hstar = Hstar, probs = probs, filename = "output")

## Simulate hypothetical species - subsampling ##
# subsamples 25% of haplotype labels
HACSobj <- HACHypothetical(N = N, Hstar = Hstar, probs = probs,
perms = 1000, p = 0.95, subsample = TRUE, prop = 0.25,
conf.level = 0.95, filename = "output")
```

```
## Simulate hypothetical species and all parameters changed - subsampling ##
HACSOBJ <- HACHypothetical(N = N, Hstar = Hstar, probs = probs,
perms = 10000, p = 0.90, subsample = TRUE, prop = 0.15, conf.level = 0.95,
num.iters = 1, filename = "output")
```

HACReal	<i>Set up an object to simulate haplotype accumulation curves for a real species</i>
---------	--

Description

Helper function which creates an object containing necessary information to run a simulation of haplotype accumulation for a real species of interest

Usage

```
HACReal(perms = 10000, p = 0.95, conf.level = 0.95,
subsample = FALSE, prop = NULL, progress = TRUE, num.iters = NULL, filename = NULL)
```

Arguments

perms	Number of permutations (replications)
p	Proportion of haplotypes to recover
conf.level	Desired confidence level for graphical output and interval estimation
subsample	Is a subsample of DNA sequences desired?
prop	If subsample = TRUE, the proportion of DNA sequences to subsample
num.iters	Number of iterations to compute
progress	Should iteration output be printed to the R console?
filename	Name of file where simulation results are to be saved

Value

A list object of class "HAC" with 13 elements that can be passed to HAC.simrep as follows:

input.seqs	Should a FASTA file of aligned/trimmed DNA sequences be inputted? Default is TRUE
subset.seqs	Should a subsample of DNA sequences be taken? Default is FALSE (can be altered by user)
prop.seqs	Proportion of DNA sequences to subsample. Default is NA (can be altered by user)
prop.haps	Proportion of haplotype labels to subsample. Default is NULL
subset.haps	Should a subsample of haplotype labels be taken? Default is NULL

N	Number of individuals. NA by default (computed automatically by algorithm)
Hstar	Number of unique species' haplotypes. NA by default (computed automatically by algorithm)
probs	Haplotype frequency distribution vector. NA by default (computed automatically by algorithm)
p	Proportion of haplotypes to recover. p = 0.95 by default.
perms	Number of permutations (replications). perms = 10000 by default.
conf.level	Desired confidence level for graphical output and interval estimation. conf.level = 0.95 by default.
num.iters	Number of iterations to compute. num.iters = NULL by default (i.e., all iterations are computed; users can specify num.iters = 1 for the first iteration.)
progress	Should iteration output be printed to the R console? Default is TRUE
filename	Name of file where simulation results are to be saved.

Examples

```
## Simulate real species ##
# outputs file called "output.csv"
HACSOBJ <- HACReal(filename = "output")

## Simulate real species - subsampling ##
# subsamples 25% of DNA sequences
HACSOBJ <- HACReal(subsample = TRUE, prop = 0.25, conf.level = 0.95,
filename = "output")

## Simulate real species and all parameters changed - subsampling ##
HACSOBJ <- HACReal(perms = 10000, p = 0.90, subsample = TRUE,
prop = 0.15, conf.level = 0.99, num.iters = 1, filename = "output")
```

Index

*Topic **package**

HACSim-package, [2](#)

accumulate, [4](#)

envr, [4](#)

HAC.sim, [5](#)

HAC.simrep, [6](#)

HACClass, [8](#)

HACHypothetical, [8](#)

HACReal, [10](#)

HACSim-package, [2](#)