Package 'HIest'

February 19, 2015

Type Package

Title Hybrid index estimation

Version 2.0

Date 2012-05-09

Author Ben Fitzpatrick

Maintainer Ben Fitzpatrick <benfitz@utk.edu>

Depends nnet, stats

Description Uses likelihood to estimate ancestry and heterozygosity. Evaluates simple hybrid classifications (parentals, F1, F2, backcrosses). Estimates genomic clines.

License GPL (>= 3)

LazyLoad yes

NeedsCompilation no

Repository CRAN

Date/Publication 2013-02-16 11:33:02

R topics documented:

`HIest-package` *Functions for estimating hybrid indices (ancestry and heterozygosity) and fitting genomic clines.*

Description

Uses likelihood to estimate ancestry, heterozygosity, and genomic cline parameters. Also evaluates simple hybrid classifications (parentals, F1, F2, backcrosses).

Details

Package:	HIest
Type:	Package
Version:	1.0
Date:	2012-02-13
License:	GPL 3.0
LazyLoad:	yes

Author(s)

Benjamin M. Fitzpatrick Maintainer: <benfitz@utk.edu>

References

Fitzpatrick, B. M. 2008. Hybrid dysfunction: Population genetic and quantitative genetic perspectives. American Naturalist 171:491-198.

Fitzpatrick, B. M. 2012. Estimating ancestry and heterozygosity of hybrids using molecular markers. BMC Evolutionary Biology 12:131. http://www.biomedcentral.com/1471-2148/12/131

Fitzpatrick, B. M. Alternative forms for genomic clines (in review)

Lynch, M. 1991. The genetic interpretation of inbreeding depression and outbreeding depression. Evolution 45:622-629.

Examples

```
## Not run:
data(Bluestone)
```

```
######################
# Fit genomic clines #
######################

data(Bluestone)
BS.fit <- Cline.fit(Bluestone[,1:12],model=c("logit.logistic","Barton"))
Cline.plot(BS.fit)

#######################################
# Estimate ancestry and heterozygosity #
#######################################

Bluestone <- replace(Bluestone,is.na(Bluestone),-9)

# parental allele frequencies (assumed diagnostic)
BS.P <- data.frame(Locus=names(Bluestone),Allele="BTS",P1=1,P2=0)

# estimate ancestry and heterozygosity
BS.est <-HIC(Bluestone)

# calculate likelihoods for early generation hybrid classes
BS.class <- HIclass(Bluestone,BS.P,type="allele.count")

# compare classification with maximum likelihood estimates
BS.test <- HItest(BS.class,BS.est,thresholds=c(2,2))

table(BS.test$c1)
# all 41 are TRUE, meaning the best classification is at least 2 log-likelihood units
# better than the next best

table(BS.test$c2)
# 2 are TRUE, meaning the MLE S and H are within 2 log-likelihood units of the best
# classification, i.e., the simple classification is rejected in all but 2 cases

table(BS.test$Best.class,BS.test$c2)
# individuals were classified as F2-like (class 3) or backcross to CTS (class 4), but
# only two of the F2's were credible

BS.test[BS.test$c2,]
# in only one case was the F2 classification a better fit (based on AIC) than the
# continuous model.

# equivalent to the AIC criterion:
BS.test <- HItest(BS.class,BS.est,thresholds=c(2,1))

#########################
# three-way hybrid zone #
#########################

# for example: make each parental, F1, F2, and backcross
G <- rbind(
rep(1,12),rep(1,12),
rep(2,12),rep(2,12),
```

```
rep(3,12),rep(3,12),
rep(1,12),rep(2,12),
rep(1:2,each=6),rep(1:2,6),
rep(1,12),rep(1:2,6),
rep(2,12),rep(1:2,6),
rep(1,12),rep(3,12),
rep(c(1,3),each=6),rep(c(1,3),6),
rep(1,12),rep(c(1,3),6),
rep(3,12),rep(c(1,3),6),
rep(2,12),rep(3,12),
rep(2:3,each=6),rep(2:3,6),
rep(3,12),rep(2:3,6),
rep(2,12),rep(2:3,6)
)

# 12 diagnostic markers
P <- data.frame(Locus=rep(1:12,each=3), allele=rep(1:3,12), P1=rep(c(1,0,0),12),
P2=rep(c(0,1,0),12), P3=rep(c(0,0,1),12))

# find MLE with simulated annealing ... takes a few minutes with default iterations
# Est <- threeway(G,P,method="SANN",surf=FALSE)

# shortcut for diagnostic markers
Est <- HIC3(G,P)
CL <- thirdclass(G,P)

## End(Not run)
```

`Bluestone` *Original marker data for hybrid tiger salamander larvae collected from Bluestone Quarry pond.*

Description

Each row is an individual, each column is a marker. This is an example type="allele.count". Genotypes are 0 (homozygous for native California Tiger Salamander allele), 1 (heterozygous), or 2 (homozygous for introduced Barred Tiger Salamander allele). There are NA's.

Usage

```
data(Bluestone)
```

Format

A data frame with 41 observations on 64 markers.

Source

Fitzpatrick, B. M., J. R. Johnson, D. K. Kump, H. B. Shaffer, J. J. Smith, and S. R. Voss. 2009. Rapid fixation of non-native alleles revealed by genome-wide SNP analysis of hybrid tiger salamanders. BMC Evolutionary Biology 9:176. http://www.biomedcentral.com/1471-2148/9/176

Examples

```
## Not run:
data(Bluestone)
BS.fit <- Cline.fit(Bluestone[,1:12], model = c("logit.logistic", "Barton"))
Cline.plot(BS.fit)

# # parental allele frequencies (assumed diagnostic)
BS.P <- data.frame(Locus=names(Bluestone),Allele="BTS",P1=1,P2=0)

# # estimate ancestry and heterozygosity
BS.est <-HIest(Bluestone,BS.P,type="allele.count")

# shortcut for diagnostic markers and allele count data:
BS.est <- HIC(Bluestone)

# # calculate likelihoods for early generation hybrid classes
BS.class <- HIclass(Bluestone,BS.P,type="allele.count")

# # compare classification with maximum likelihood estimates
BS.test <- HItest(BS.class,BS.est)

table(BS.test$c1)
# # all 41 are TRUE, meaning the best classification is at least 2 log-likelihood units
# # better than the next best

table(BS.test$c2)
# # 2 are TRUE, meaning the MLE S and H are within 2 log-likelihood units of the best
# # classification, i.e., the simple classification is rejected in all but 2 cases

table(BS.test$Best.class,BS.test$c2)
# # individuals were classified as F2-like (class 3) or backcross to CTS (class 4),
# # but only two of the F2's were credible

BS.test[BS.test$c2,]
# # in only one case was the F2 classification a better fit (based on AIC) than the
# # continuous model.

## End(Not run)
```

`Cline.fit` *Fit alternative cline functions to data from one or more genetic loci.*

Description

This function takes individual or population data for multiple genetic loci and fits genomic clines. This implementation assumes markers are diagnostic.

Usage

```
Cline.fit(Data, By = NULL, S = NULL, model,
  Start = NULL, Methods = NULL, iterations = 99, SD = NULL, headstart = TRUE,
   Grid = FALSE, ploidy = 2, trim = 0, include = 1:ncol(Data))
```

Arguments

Data	Data matrix of allele counts (allele diagnostic for one species desgnated 1, all others 0) by individual or sample (rows) and marker (columns).
By	Optional factor defining aggregation of individuals into samples (e.g., sites or populations). Length must be equal to the number of rows in Data.
S	Optional alternative values for the genome wide hybrid index. By default, the mean ancestry across all loci in Data will be used. If S is specified, the values given by the user trump the default.
model	Character vector specifying which models to fit. Valid choices are "multinom","binom","logit.logistic","B and "Richards".
Start	Optional starting values for optimization. If NULL, start values at the 1:1 line are provided by default.
Methods	Optional named list of strings indicating the optimization method for each cline model (logit-logistic, Barton, Beta, and Richards). Choices are "L-BFGS-B", "SANN", and "mcmc". Default is Methods=list(logit.logistic="L-BFGS-B",Barton="L-BFGS-B",Be
iterations	The number of MCMC generations to use if "mcmc" is used for any Methods.
SD	Dispersion parameters for the "mcmc" and "SANN" methods. In these methods, new parameter values are proposed by drawing values from normal distributions centered on the current value and with standard deviations from SD.
headstart	Logical: if TRUE and method="mcmc" or "SANN", starting values will be found by first using optim with "L-BFGS-B".
Grid	Logical: if TRUE and method="mcmc" and model="Beta", starting values for the Markov Chain will be found by finding the highest likelihood on a 100 x 100 grid made by mu <- seq(from=0.02,to=0.90,length.out=10); nu <- 2^(0:9)/10.
ploidy	Ploidy of the data, i.e., an integer (1 or 2) to multiply the sample size (number of individuals).
trim	Optional fraction of extreme values to omit from calculation of S (see mean).
include	Optional vector of column indices to include in calculation of S - can be used to omit biological outliers.

Value

A list including a named matrix for each fitted model. Each such matrix includes the parameter estimates, maximum log-likelihood, AICc, squared Mahalanobis distance D2, P-value for Mahalanobis-based outlier detection, and TRUE/FALSE declaration of whether a marker is an outlier based on a Bonferroni-adjusted critical P-value. If more than one model was fit, an additional data frame best.fit is included, giving the model with lowest AIC for each marker.

Author(s)

Benjamin M. Fitzpatrick

References

Fitzpatrick, B. M. 2012. Alternative forms for genomic clines. In prep

See Also

See `gcline.fn` for the basic fitting function. `Cline.plot` provides an easy way to visualize the output of `Cline.fit`.

Examples

```
data(Bluestone)
BS.fit <- Cline.fit(Bluestone[,1:12],model=c("logit.logistic","Barton"))
Cline.plot(BS.fit)
```

`Cline.plot` *Graph genomic clines fit to data*

Description

A quick way to summarize and view the results of `Cline.fit`

Usage

```
Cline.plot(cfit)
```

Arguments

`cfit`	An object: the output of `Cline.fit`.

Details

Creates up to sixteen graphs, depending on which models are fit to the data. For each fitted model, fitted clines are graphed first. For two-parameter models, a bivariate graph of parameter estimates is plotted next (this is left blank for four-parameter models). Finally, a Q-Q plot of squared Mahalanobis distances vs. CHi-squared quantiles is graphed. Putative outlier markers are indicated in red.

Author(s)

Benjamin M. Fitzpatrick

References

Fitzpatrick, B. M. 2012. Alternative forms for genomic clines. In prep

See Also

See `gcline.fn` for the basic fitting function. `Cline.plot` provides an easy way to visualize the output of `Cline.fit`.

Examples

```
data(Bluestone)
BS.fit <- Cline.fit(Bluestone[,1:12],model=c("logit.logistic","Barton"))
Cline.plot(BS.fit)
```

`gcline.fn` *Fit a genomic cline using maximum likelihood*

Description

Fit a genomic cline and compare it to a null expectation. Choices are logit-logistic cline, Barton cline, Beta cline, multinomial regression, binomial regression, and Richards cline. This function is used by `Cline.fit` to compare clines fit to a data set.

Usage

```
gcline.fn(x, n, y, start, model = "logit-logit", method = "L-BFGS-B", iterations = 99,
 SD = rep(0.01, length(start)), headstart = FALSE,Grid = TRUE)
```

Arguments

`x`	A numeric vector of genome-wide mean ancestry (or any independent variable on the unit interval).
`n`	A numeric vector of sample sizes for each value in `x`. E.g., for individual diploid data, `n=2`.
`y`	A numeric vector containing the dependent variable: usually an allele count for each `x`
`start`	A vector of starting values: u and v for the logit-logistic, μ and ν for the Beta cline, a and b for the Barton cline, and U, L, m, and b for the Richards cline.
`model`	Character string indicating which cline function to fit:`"logit-logit"`, `"Beta"`, `"Barton"`, `"multinom"`, `"logistic"` or `"Richards"`
`method`	Character string indicating which algorithm to use to find the MLE: `"L-BFGS-B"` and `"SANN"` are used by the native optimization function `optim`, `"mcmc"` is a Markov Chain Monte Carlo using Metropolis-Hastings sampling. If `"mcmc"` is used, the following four items are also used.

iterations	The desired number of MCMC generations. The larger this number is, the greater the chance that the chain will find the maximum likelihood.
SD	Dispersion parameters for the "mcmc" and "SANN" methods. In these methods, new parameter values are proposed by drawing values from normal distributions centered on the current value and with standard deviations from SD.
headstart	Logical: if TRUE and method="mcmc" or "SANN", starting values will be found by first using optim with "L-BFGS-B".
Grid	Logical: if TRUE and method="mcmc" and model="Beta", starting values for the Markov Chain will be found by finding the highest likelihood on a 100 x 100 grid made by mu <- seq(from=0.02,to=0.90,length.out=10); nu <- 2^(0:9)/10.

Value

A list:

model	The cline function used
method	The optimization method used
estimates	Maximum likelihood parameter estimates
lnL	The maximum likelihood and the likelihood of the data given the naive null model E(y)=x
k	The number of fitted coefficients
AICc	Akaike's information criterion with sample size correction
convergence	From optim: 0 means the algorithm thinks it did a good job, otherwise the MLE may be incorrect. If mcmc was used, this will be the full Markov Chain, which could be analyzed for convergence (e.g., see coda).

Author(s)

Benjamin M. Fitzpatrick

References

Fitzpatrick, B. M. 2012. Alternative forms for genomic clines. In prep

See Also

See Cline.fit for application to multilocus data sets and analysis of interclass heterozygosity.

Examples

```
x <- 0:50/50          # hypothetical genomic ancestry proportions
y <- rbinom(50,2,x)   # random diploid genotypes for a diagnostic marker
n=rep(2,50)   # sample size is two alleles per diploid individual

gcline.fn(x=x,n=n,y=y,start=c(.5,2),model="logit-logit")
```

HIC *Closed form maximum-likelihood estimates of ancestry and heterozygosity for diagnostic markers*

Description

For hybrid individuals genotyped with diagnostic markers (one allele fixed in each parental lineage), the ancestry index and interclass heterozygosity are calculated using closed form estimators using allele counts.

Usage

```
HIC(G)
```

Arguments

G	Data matrix of individual genotypes (individuals in rows, markers in columns), coded as 0, 1, or 2 for the number of alleles inherited from parental lineage 1.

Details

Data must be coded as allele counts and markers must be assumed diagnostic. The MLE of the ancestry index is $S = \sum(x_i)/(2n)$, where x_i is the allele count for locus i. The MLE of interclass heterozygosity is simply the observed fraction of markers that are heterozygous.

Value

A matrix with three named columns is returned.

S	The ancestry index for each individual
H	The interclass heterozygosity for each individual
logLik	The log-likelihood of the joint estimate, given the data for each individual

Author(s)

Ben Fitzpatrick

References

Fitzpatrick, B. M. 2008. Hybrid dysfunction: Population genetic and quantitative genetic perspectives. American Naturalist 171:491-198.

Fitzpatrick, B. M. 2012. Estimating ancestry and heterozygosity of hybrids using molecular markers. BMC Evolutionary Biology 12:131. http://www.biomedcentral.com/1471-2148/12/131

Lynch, M. 1991. The genetic interpretation of inbreeding depression and outbreeding depression. Evolution 45:622-629.

See Also

HIest finds maximum likelihood estimates for non-diagnostic markers.

Examples

```
## A random set of allele counts for 10 loci and 10 individuals
G <- matrix(rbinom(100,2,.5),nrow=10)
HIC(G)
```

HIC3 *Closed form maximum-likelihood estimates of ancestry and heterozygosity for diagnostic markers in a three-way hybrid zone*

Description

For hybrids with up to three parental lineages with diagnostic molecular markers (one allele fixed in each parental lineage). The function computes genomic proportions and ancestry indices using closed form estimators.

Usage

```
HIC3(G, P)
```

Arguments

G	A matrix or data frame of diploid genotypes. Each individual occupies two consecutive rows (one for each allele). Each marker is a column. For heterozygotes, it does not matter which allele is entered first.
P	A matrix or data frame identifying which alleles are fixed in each parental lineage. The first column stores locus identifiers and the second allele identifiers (one per parental lineage). The value of the third column is 1 for each allele fixed in the first parental lineage, and 0 otherwise. The value of the fourth column is 1 for each allele fixed in the second parental lineage, and 0 otherwise. The value of the fifth column is 1 for each allele fixed in the third parental lineage, and 0 otherwise.

Value

A matrix with 10 named columns for each individual

p11	Proportion of markers homozygous for lineage 1 alleles
p22	Proportion of markers homozygous for lineage 2 alleles
p33	Proportion of markers homozygous for lineage 3 alleles
p12	Proportion of markers heterozygous for lineage 1 and 2 alleles
p13	Proportion of markers heterozygous for lineage 1 and 3 alleles

p23	Proportion of markers heterozygous for lineage 2 and 3 alleles
S1	Lineage 1 ancestry index: proportion of alleles derived from parental lineage 1
S2	Lineage 2 ancestry index: proportion of alleles derived from parental lineage 2
S3	Lineage 3 ancestry index: proportion of alleles derived from parental lineage 3
logLik	log-likelihood of the genomic proportions given the individual marker data

Author(s)

Ben Fitzpatrick

References

Fitzpatrick, B. M. 2012. Estimating ancestry and heterozygosity of hybrids using molecular markers. BMC Evolutionary Biology 12:131. http://www.biomedcentral.com/1471-2148/12/131

See Also

threeway finds maximum likelihood estimates for non-diagnostic marker data. thirdclass and HItest3 evaluate simple classification of three-way hybrids into parental, F1, F2, and backcross categories. For conventional two-way hybrid zone analyses, see HIC, HIest,link{HIclass},link{HItest}.

Examples

```
## all possible 2-way crosses after 2 generations
G <- rbind(
rep(1,12),rep(1,12),               # parental 1
rep(2,12),rep(2,12),               # parental 2
rep(3,12),rep(3,12),               # parental 3
rep(1,12),rep(2,12),               # 1 x 2 F1
rep(1:2,each=6),rep(1:2,6),        # 1 x 2 F2
rep(1,12),rep(1:2,6),              # 1 x 1 x 2 BC
rep(2,12),rep(1:2,6),              # 1 x 2 x 2 BC
rep(1,12),rep(3,12),               # 1 x 3 F1
rep(c(1,3),each=6),rep(c(1,3),6),  # 1 x 3 F2
rep(1,12),rep(c(1,3),6),           # 1 x 1 x 3 BC
rep(3,12),rep(c(1,3),6),           # 1 x 3 x 3 BC
rep(2,12),rep(3,12),               # 2 x 3 F1
rep(2:3,each=6),rep(2:3,6),        # 2 x 3 F2
rep(3,12),rep(2:3,6),              # 2 x 3 x 3 BC
rep(2,12),rep(2:3,6)               # 2 x 2 x 3 BC
)

P <- data.frame(Locus=rep(1:12,each=3),allele=rep(1:3,12),P1=rep(c(1,0,0),12),
P2=rep(c(0,1,0),12),P3=rep(c(0,0,1),12))

HIC3(G,P)
```

HIclass *Calculate likelihoods for early generation hybrid genotype classes*

Description

HIclass uses genetic marker data and parental allele frequencies to calculate the likelihoods for each of the six diploid genotype classes possible in the first two generations of admixture (each parental, F1, F2, and each backcross)

Usage

```
HIclass(G, P, type)
```

Arguments

G	A matrix or data frame of genetic marker data. Each column is a locus. For type="dominant" or type="allele.count", there should be one row per individual. For type="codominant", each individual is to be represented in consecutive rows (one for each allele).
P	A matrix or data frame with the following columns (order is important!): Locus name, Allele name, P1 allele frequency, P2 allele frequency. For type="dominant" or type="allele.count", there should be one row per locus, giving the frequencies of the dominant or "1" allele. For type="codominant" there should be a separate row for each allele AND the Allele names should match the data in G.
type	A string representing the data type. The options are "codominant", "dominant", and "allele.count".

Details

Classification should be used only for systems for which two generations of admixture are credible. HIest can be used to find the joint maximum likelihood estimates of ancestry (S) and interclass heterozygosity (H), which offer a more nuanced summary of hybrid genotypes.

Value

A data frame with the following columns (one row per individual)

class100	log-likelihood for genotype class expected for pure parental P2 (P = 0, H = 0)
class010	log-likelihood for genotype class expected for F1 hybrids (P = 0.5, H = 1)
class121	log-likelihood for genotype class expected for F2 hybrids (P = 0.5, H = 0.5)
class110	log-likelihood for genotype class expected for backcrosses to parental 2 (P = 0.25, H = 0.5)
class011	log-likelihood for genotype class expected for backcrosses to parental 1 (P = 0.75, H = 0.5)
class001	log-likelihood for genotype class expected for pure parental P1 (P = 1, H = 0)

Best	The class with the highest likelihood of the six
LLD	The difference in log-likelihood between the best and second-best fit class. This can be used as a criterion for deciding whether the best fit class is enough better to reject the alternatives.

Author(s)

Ben Fitzpatrick

References

Fitzpatrick, B. M. 2008. Hybrid dysfunction: Population genetic and quantitative genetic perspectives. American Naturalist 171:491-198.

Fitzpatrick, B. M. 2012. Estimating ancestry and heterozygosity of hybrids using molecular markers. BMC Evolutionary Biology 12:131. http://www.biomedcentral.com/1471-2148/12/131

Lynch, M. 1991. The genetic interpretation of inbreeding depression and outbreeding depression. Evolution 45:622-629.

See Also

HIest for maximum likelihood estimation of S and H, HIsurf for a likelihood surface, HItest to compare the classification to the maximum likelihood, HILL for the basic likelihood function.

Examples

```
data(Bluestone)
Bluestone <- replace(Bluestone,is.na(Bluestone),-9)
# parental allele frequencies (assumed diagnostic)
BS.P <- data.frame(Locus=names(Bluestone),Allele="BTS",P1=1,P2=0)

# estimate ancestry and heterozygosity
# BS.est <-HIest(Bluestone,BS.P,type="allele.count")

# shortcut for diagnostic markers
BS.est <- HIC(Bluestone)

# calculate likelihoods for early generation hybrid classes
BS.class <- HIclass(Bluestone,BS.P,type="allele.count")

# compare classification with maximum likelihood estimates
BS.test <- HItest(BS.class,BS.est,thresholds=c(2,2))

table(BS.test$c1)
# all 41 are TRUE, meaning the best classification is at least 2 log-likelihood units
# better than the next best

table(BS.test$c2)
# 2 are TRUE, meaning the MLE S and H are within 2 log-likelihood units of the best
# classification, i.e., the simple classification is rejected in all but 2 cases

table(BS.test$Best.class,BS.test$c2)
```

```
# individuals were classified as F2-like (class 3) or backcross to CTS (class 4),
# but only two of the F2's were credible

BS.test[BS.test$c2,]
# in only one case was the F2 classification a better fit (based on AIC) than the
# continuous model.

# equivalent to the AIC criterion:
BS.test <- HItest(BS.class,BS.est,thresholds=c(2,1))
```

`HIest` *Find the joint maximum likelihood estimates of ancestry and interclass heterozygosity in a sample of hybrids.*

Description

`HIest` provides approaches to find maximum likelihood estimates of S and H, using the likelihood functions described by Fitzpatrick (2012).

Usage

```
HIest(G, P, type, method = "SANN", iterations = 1000, Cscale = NULL, surf = FALSE,
 startgrid = 99, start = c(.5,.5), control = list(fnscale = -1, maxit = iterations))
```

Arguments

`G`	A matrix or data frame of genetic marker data. Each column is a locus. For `type="dominant"` or `type="allele.count"`, there should be one row per individual. For `type="codominant"`, each individual is to be represented in consecutive rows (one for each allele).
`P`	A matrix or data frame with the following columns (order is important!): Locus name, Allele name, P1 allele frequency, P2 allele frequency. For `type="dominant"` or `type="allele.count"`, there should be one row per locus, giving the frequencies of the dominant or "1" allele. For `type="codominant"` there should be a separate row for each allele AND the Allele names should match the data in `G`.
`type`	A string representing the data type. The options are `"codominant"`, `"dominant"`, and `"allele.count"`.
`method`	Optimization method to search for maximum likelihood estimates of ancestry and heterozygosity. Alternatives are `"SANN"`, `"L-BFGS-B"`, `"surf"`, and `"mcmc"`. See details.
`iterations`	The desired number of MCMC steps to perform when `method="mcmc" or "SANN"`.
`Cscale`	An integer, controlling the the proposal distribution for `method = "SANN"` or `"mcmc"`. The default value is 100. Smaller values will cause the algorithm to search more broadly, but could make the search inefficient. See details.

surf	Logical: Should the function find starting values by evaluating likelihoods on a grid?
startgrid	Integer. This controls the size of the grid for method="surf". It is the same as the argument size in the function HIsurf.
start	A vector including the starting values of S and H.
control	A list of options to be passed to control in the optim function. Whatever else is chosen, be sure fnscale is negative to make optim search for a maximum rather than a minimum.

Details

Given two ancestral species or parental populations (P1 and P2), the ancestry index (S) is the proportion of an individual's alleles descending from alleles in the P1 population and the interclass heterozygosity (H) is the proportion of an individual's loci that have one allele from each ancestral population (Lynch 1991). The likelihood functions are described in Fitzpatrick (2012). The likelihood functions take advantage of the correspondence between ancestry and heterozygosity and the genomic proportions: p_{11} = proportion of one's genome that is homozygous for alleles inherited from parental lineage 1, $p_{12} = H$ = proportion one's genome that is heterozygous for alleles inhertied from each parental lineage, and p_{22} = proportion of one's genome that is homozygous for alleles inherited from parental lineage 2.

Currently, the function provides four methods for searching the likelihood surface. method = "SANN" is probably the best; it uses the general purpose optimization function optim with its simulated annealing algorithm. For this estimation problem, a custom proposal function is passed to the option gr. This proposal function draws new genomic proportions (p_{11}, p_{12}, p_{22}) from a three dimensional Dirichlet distribution centered on the old genomic proportions. The concentration of the proposal distribution is controlled by Cscale; the larger this value, the more the proposal distribution is concentrated near the current state.

method = "mcmc" uses a Markov-Chain Monte Carlo with Metropolis-Hastings sampling to explore the likelihood surface. It also uses the Dirichlet proposal distribution, and could be useful (with some modification of the code) for generating posterior distributions. method = "SANN" is probaby superior for simply finding the MLE.

method = "L-BFGS-B" also uses optim, but with a quasi-Newton likelihood search algorithm to look for the maximum likelihood. This method is relatively fast, but it can miss the MLE if it is near the edge of the sample space.

"surf", finds all likelihoods on a grid defined by startgrid and chooses the maximum. This is not going to find the MLE unless the MLE happens to be one of the grid points. However, using the option surf = TRUE with the SANN or mcmc methods can improve efficiency by initiating the search at the grid point nearest the MLE.

Value

A data frame with one row for each individual and three columns:

S	The maximim likelihood estimate of the ancestry index
H	The maximum likelihood estimate of the interclass heterozygosity
logLik	The maximum log-likelihood

Author(s)

Ben Fitzpatrick

References

Fitzpatrick, B. M. 2008. Hybrid dysfunction: Population genetic and quantitative genetic perspectives. American Naturalist 171:491-198.

Fitzpatrick, B. M. 2012. Estimating ancestry and heterozygosity of hybrids using molecular markers. BMC Evolutionary Biology 12:131. http://www.biomedcentral.com/1471-2148/12/131

Fitzpatrick, B. M., J. R. Johnson, D. K. Kump, H. B. Shaffer, J. J. Smith, and S. R. Voss. 2009. Rapid fixation of non-native alleles revealed by genome-wide SNP analysis of hybrid tiger salamanders. BMC Evolutionary Biology 9:176. http://www.biomedcentral.com/1471-2148/9/176

Lynch, M. 1991. The genetic interpretation of inbreeding depression and outbreeding depression. Evolution 45:622-629.

See Also

HIsurf for a likelihood surface, HIclass for likelihoods of early generation hybrid classes, HItest to compare the classification to the maximum likelihood, HILL for the basic likelihood function.

Examples

```
##-- A random codominant data set of 5 individuals and 5 markers with three alleles each
L <- 10
P <- data.frame(Locus=rep(1:L,each=3),Allele=rep(1:3,L),
P1=as.vector(rmultinom(L,10,c(.7,.2,.1)))/10,
P2=as.vector(rmultinom(L,10,c(.1,.2,.7)))/10)
G <- matrix(nrow=10,ncol=L)
for(i in 1:L){
G[,i] <- sample(c(1,2,3),size=10,replace=TRUE,prob=rowMeans(P[P$Locus==i,3:4]))
}

HI.cod <- HIest(G,P,type="codominant",iterations=99,surf=TRUE,startgrid=20)

# this is unlikely to converge on the MLE: increase iterations and/r startgrid.

# # optional plot
# plot(c(0,.5,1,0),c(0,1,0,0),type="l",xlab=expression(italic(S)),
# ylab=expression(italic(H[I])),lwd=2,cex.lab=1.5,cex.axis=1.5,bty="n")
# points(HI.cod$S,HI.cod$H,cex=1.5,lwd=2)
# axis(1,labels=FALSE,lwd=2);axis(2,labels=FALSE,lwd=2)

# # other examples

# ##-- Make it into allele count data (count "3" alleles)
# P.c <- P[seq(from=3,to=dim(P)[2],by=3),]
# G.c <- matrix(nrow=5,ncol=L)
# for(i in 1:5){
# G.c[i,] <- colSums(G[c(i*2-1,i*2),]==3)
# }
```

```
# HI.ac <- HIest(G.c,P.c,type="allele.count",iterations=500,surf=TRUE,startgrid=50)

# ##-- Make it into dominant data where allele 3 is dominant
# G.d <- replace(G.c,G.c==2,1)

# HI.dom <- HIest(G.d,P.c,type="dominant",iterations=500,surf=TRUE,startgrid=50)

# ## -- A real dataset (Fitzpatrick et al. 2009)
# data(Bluestone)
# Bluestone <- replace(Bluestone,is.na(Bluestone),-9)
# # parental allele frequencies (assumed diagnostic)
# BS.P <- data.frame(Locus=names(Bluestone),Allele="BTS",P1=1,P2=0)

# # estimate ancestry and heterozygosity
# # BS.est <-HIest(Bluestone,BS.P,type="allele.count")
# ## shortcut for diagnostic markers
# BS.est <- HIC(Bluestone)

# # calculate likelihoods for early generation hybrid classes
# BS.class <- HIclass(Bluestone,BS.P,type="allele.count")

# # compare classification with maximum likelihood estimates
# BS.test <- HItest(BS.class,BS.est)

# table(BS.test$c1)
# # all 41 are TRUE, meaning the best classification is at least 2 log-likelihood units
# # better than the next best

# table(BS.test$c2)
# # 2 are TRUE, meaning the MLE S and H are within 2 log-likelihood units of the best
# # classification, i.e., the simple classification is rejected in all but 2 cases

# table(BS.test$Best.class,BS.test$c2)
# # individuals were classified as F2-like (class 3) or backcross to CTS (class 4), but
# # only two of the F2's were credible

# BS.test[BS.test$c2,]
# # in only one case was the F2 classification a better fit (based on AIC) than the
# # continuous model.
```

HILL *Hybrid Index Log Likelihood*

Description

HILL calulates the log-likelihood of a hybrid genotype given parental allele frequencies, an ancestry index (S), and interclass heterozygosity (H). This function is used by HIest and HIclass for data analysis.

Usage

```
HILL(par = c(S, H), G, P, type)
```

Arguments

par	A numeric vector including a value of S and a value of H
G	An individual diploid genotype as matrix G. If type == "codominant", G must be a two-row matrix with one column for each locus, as in STRUCTURE (http://pritch.bsd.uchicago.edu/structure.html). If type == "dominant", G is a vector of 0,1 for absence,presence of the dominant allele. If type == "allele.count", G must be a vector of genotypes coded as 0,1,2 for the number of "j" alleles. That is, genotype 2 is homozygous for allele j, genotype 1 is heterozygous, and genotype 0 has no j alleles.
P	Parental allele frequencies. A matrix or data frame with the following columns (order is important!): Locus name, Allele name, P1 allele frequency, P2 allele frequency. For type="dominant" or type="allele.count", there should be one row per locus, giving the frequencies of the dominant or "1" allele. For type="codominant" there should be a separate row for each allele AND the Allele names should match the data in G.
type	A string representing the data type. The options are "codominant", "dominant", and "allele.count".

Details

Given two ancestral species or parental populations (P1 and P2), the ancestry index (S) is the proportion of an individual's alleles descending from alleles in the P1 population and the interclass heterozygosity (H) is the proportion of an individual's loci that have one allele from each ancestral population (Lynch 1991). The likelihood functions are described in Fitzpatrick (2012).

Value

The natural log of the likelihood of the data given S, H, and P.

Author(s)

Ben Fitzpatrick

References

Fitzpatrick, B. M. 2008. Hybrid dysfunction: Population genetic and quantitative genetic perspectives. American Naturalist 171:491-198.

Fitzpatrick, B. M. 2012. Estimating ancestry and heterozygosity of hybrids using molecular markers. BMC Evolutionary Biology 12:131. http://www.biomedcentral.com/1471-2148/12/131

Lynch, M. 1991. The genetic interpretation of inbreeding depression and outbreeding depression. Evolution 45:622-629.

See Also

HIest for maximum likelihood estimation of S and H, HIsurf for a likelihood surface, HIclass for likelihoods of early generation hybrid classes, HItest to compare the classification to the maximum likelihood.

Examples

```
##-- A random codominant data set of 1 individual with 5 markers with three possible alleles each
P <- data.frame(Locus=rep(1:5,each=3),Allele=rep(1:3,5),
P1=as.vector(rmultinom(5,10,c(.7,.2,.1)))/10,
P2=as.vector(rmultinom(5,10,c(.1,.2,.7)))/10)
G <- matrix(nrow=2,ncol=5)
for(i in 1:5){
G[,i] <- sample(c(1,2,3),size=2,replace=TRUE,prob=rowMeans(P[P$Locus==i,3:4]))
}

HILL(par=c(0.5,0.5),G,P,type="codominant")

##-- Make it into allele count data (count "3" alleles)
P.c <- P[seq(from=3,to=15,by=3),]
G.c <- colSums(G==3)

HILL(par=c(0.5,0.5),G.c,P.c,type="allele.count")

##-- Make it into dominant data where allele 3 is dominant
G.d <- replace(G.c,G.c==2,1)

HILL(par=c(0.5,0.5),G.d,P.c,type="dominant")
```

HIsurf *Describe the joint likelihood surface of ancestry and heterozygosity for a hybrid genotype.*

Description

HIsurf calculates the log likelihood of points on a bivariate grid to describe the joint likelihood surface of ancestry and interclass heterozygosity for a genotype given parental allele frequencies.

Usage

```
HIsurf(G, P, type, size)
```

Arguments

G	An individual diploid genotype as matrix G. If type == "codominant", G must be a two-row matrix with one column for each locus, as in STRUCTURE (http://pritch.bsd.uchicago.edu/structure.html). If type == "dominant", G is a vector of 0,1 for absence,presence of the dominant allele. If type == "allele.count",

G must be a vector of genotypes coded as 0,1,2 for the number of "j" alleles. That is, genotype 2 is homozygous for allele j, genotype 1 is heterozygous, and genotype 0 has no j alleles.

P — Parental allele frequencies. A matrix or data frame with the following columns (order is important!): Locus name, Allele name, P1 allele frequency, P2 allele frequency. For `type="dominant"` or `type="allele.count"`, there should be one row per locus, giving the frequencies of the dominant or "1" allele. For `type="codominant"` there should be a separate row for each allele AND the Allele names should match the data in G.

type — A string representing the data type. The options are `"codominant"`, `"dominant"`, and `"allele.count"`.

size — An integer giving the desired number of gridlines in each direction. The function will calculate the likelihood for all `(size^2)/2` combinations of S and H that fall within the triangular sample space.

Details

Given two ancestral species or parental populations (P1 and P2), the ancestry index (S) is the proportion of an individual's alleles descending from alleles in the P1 population and the interclass heterozygosity (H) is the proportion of an individual's loci that have one allele from each ancestral population (Lynch 1991). The likelihood functions are described in Fitzpatrick (2012).

Value

A `size x size` matrix of log likelihoods for all combinations of ancestry (S) and interclass heterozygosity (H). Rows correspond to the `size` values of S, and columns the `size` values of H. For impossible combinations (`H > min(2*S,2-2*S)`), NA's are returned.

Author(s)

Ben Fitzpatrick

References

Fitzpatrick, B. M. 2008. Hybrid dysfunction: Population genetic and quantitative genetic perspectives. American Naturalist 171:491-198.

Fitzpatrick, B. M. 2012. Estimating ancestry and heterozygosity of hybrids using molecular markers. BMC Evolutionary Biology 12:131. http://www.biomedcentral.com/1471-2148/12/131

Lynch, M. 1991. The genetic interpretation of inbreeding depression and outbreeding depression. Evolution 45:622-629.

See Also

`HIest` for maximum likelihood estimation of S and H, `HIclass` for likelihoods of early generation hybrid classes, `HItest` to compare the classification to the maximum likelihood, `HILL` for the basic likelihood function.

Examples

```
data(Bluestone)
Bluestone <- replace(Bluestone,is.na(Bluestone),-9)
# parental allele frequencies (assumed diagnostic)
BS.P <- data.frame(Locus=names(Bluestone),Allele="BTS",P1=1,P2=0)

# a small surface to view in the console
BS.surf.5 <- HIsurf(Bluestone[21,],BS.P,type="allele.count",size=5)
BS.surf.5 # the maximum likelihood is very near the center (S ~ H ~ 0.5)

# # a more finely sampled surface to visualize with image
# BS.surf <- HIsurf(Bluestone[21,],BS.P,type="allele.count",size=99)

# image(-BS.surf,col=gray(seq(from=0,to=1,length.out=6)),
# breaks=seq(from=min(-BS.surf,na.rm=TRUE),by=2,length.out=7),
# cex.axis=1.5,bty="n",xaxs="r",yaxs="r")
# # the breaks option is set so that each level of shading corresponds to 2 log-likelihood
# # units (for one unit increments, set by=1).
# # now make it pretty:
# image(is.na(BS.surf),col="light blue",breaks=c(.5,1),add=TRUE)
# axis(1,labels=FALSE,lwd=2); axis(2,labels=FALSE,lwd=2)
# title(xlab=expression(italic(S)),ylab=expression(italic(H[I])),cex.lab=1.5)
# lines(c(0,.5,1,0),c(0,1,0,0),lty=2,lwd=2)
```

`HItest` *Compare the likelihood of hybrid classification to MLE estimates of ancestry and heterozygosity.*

Description

`HItest` compares the best fit of six early generation diploid hybrid genotypes (parental, F1, F2, backcross) to the maximum likelihood genotype decribed by ancestry (S) and interclass heterozygosity (H).

Usage

```
HItest(class, MLE, thresholds = c(2, 1))
```

Arguments

`class`	Output from `HIclass`: a data frame summarizing the fit of each individual to the six genotype classes.
`MLE`	Output from `HIest`: a data frame giving the MLE S and H and associated log-likelihood.
`thresholds`	Criteria for classification. The first criterion (`thresholds[1]`) is a cutoff for the difference in log-likelihood for the best vs. second best genotype class. The second criterion (`thresholds[2]`) is a cutoff for the difference in log-likelihood for the best genotype class vs. the MLE.

Details

As a quick-and-dirty rule of thumb, one might accept a putative classification as credible if the log-likelihood of the best-fit class was over 2 units greater than the log-likelihood of the second best-fit class AND within 2 units of the maximum log-likelihood. The first criterion is based on the approximate equivalence of a 2 x log-likelihood interval to a 95 percent confidence interval for some distributions (Hudson 1971; Hillborn and Mangel 1997). The second is based on the conventional penalty of two log-likelihood units for an additional estimated parameter in model selection (Edwards 1972; Burnaham and Anderson 2004). The classification model can be viewed as having one free parameter (once the best-fit class is set to "chosen", the other five are constrained to "not chosen"), while the continuous model has two (S and H). This approach has the disadvantage of effectively treating the classification as a null model, which is not biologically justified.

A better approach might be to accept the classification only if its AIC is lower than the AIC of the MLE, i.e., if `dAIC` is negative (Fitzpatrick 2012). Note that `dAIC` cannot be less than -2 (the case where the MLE is identical to the expectation for a class).

Value

A data frame with one row per individual. Columns are:

`S`	The maximim likelihood estimate of the ancestry index from `HIest`.
`H`	The maximum likelihood estimate of the interclass heterozygosity from `HIest`.
`Best.class`	The class with the highest likelihood of the six from `HIclass`.
`LL.class`	The log-likelihood of the data for the best-fit class from `HIclass`.
`LLD.class`	The difference in log-likelihood between the best and second-best fit class from `HIclass`.
`LL.max`	The maximum log-likelihood from `HIest`.
`dAIC`	The difference in AIC between the continuous model MLE (2 estimated parameters) and the best-fit class (1 estimated parameter).
`c1`	Logical: `TRUE` if the best-fit class is supported by more than `thresholds[1]` log-likelihood units over the second best.
`c2`	Logical: `TRUE` if the best-fit class is WITHIN `thresholds[2]` log-likelihood units of the MLE.

Author(s)

Ben Fitzpatrick

References

Burnham, K. P., and D. R. Anderson. 2004. Multimodel inference: understanding AIC and BIC in model selection. Sociological Methods and Research 33:261-304.

Edwards, A. W. F. 1972. Likelihood. Cambridge University Press, Cambridge.

Fitzpatrick, B. M. 2008. Hybrid dysfunction: Population genetic and quantitative genetic perspectives. American Naturalist 171:491-198.

Fitzpatrick, B. M. 2012. Estimating ancestry and heterozygosity of hybrids using molecular markers. BMC Evolutionary Biology 12:131. http://www.biomedcentral.com/1471-2148/12/131

Hilborn, R., and M. Mangel. 1997. The ecological detective: Confronting models with data. Princeton University Press, New Jersey.

Hudson, D. J. 1971. Interval estimation from the likelihood function. Journal of the Royal Statistical Society, Series B 33: 256-262.

Lynch, M. 1991. The genetic interpretation of inbreeding depression and outbreeding depression. Evolution 45:622-629.

See Also

HIest for maximum likelihood estimation of S and H, HIsurf for a likelihood surface, HIclass for likelihoods of early generation hybrid classes, HILL for the basic likelihood function.

Examples

```
## Not run:
data(Bluestone)
Bluestone <- replace(Bluestone,is.na(Bluestone),-9)
# parental allele frequencies (assumed diagnostic)
BS.P <- data.frame(Locus=names(Bluestone),Allele="BTS",P1=1,P2=0)

# estimate ancestry and heterozygosity
BS.est <-HIC(Bluestone)

# calculate likelihoods for early generation hybrid classes
BS.class <- HIclass(Bluestone,BS.P,type="allele.count")

# compare classification with maximum likelihood estimates
BS.test <- HItest(BS.class,BS.est)

table(BS.test$c1)
# all 41 are TRUE, meaning the best classification is at least 2 log-likelihood units
# better than the next best

table(BS.test$c2)
# 2 are TRUE, meaning the MLE S and H are within 2 log-likelihood units of the best
# classification, i.e., the simple classification is rejected in all but 2 cases.

table(BS.test$Best.class,BS.test$c2)
# individuals were classified as F2-like (class 3) or backcross to CTS (class 4), but
# only two of the F2's were credible

BS.test[BS.test$c2,]
# in only one case was the F2 classification a better fit (based on AIC) than the
# continuous model.

## End(Not run)
```

HItest3 *Compare the likelihood of hybrid classification to MLE estimates of ancestry and heterozygosity for three-way hybrid zones.*

Description

HItest3 compares the best fit of fifteen early generation diploid hybrid genotypes (parental, F1, F2, backcross between all three pairs of parental lineages) to the maximum likelihood genotype proportions estimated by threeway or HIC3.

Usage

```
HItest3(class, MLE, thresholds = c(2, 8))
```

Arguments

class	Object containing output from thirdclass
MLE	Object containing output from threeway or HIC3.
thresholds	Vector of length 2, containing criteria for classification. The first criterion (thresholds[1]) is a cutoff for the difference in log-likelihood for the best vs. second best genotype class. The second criterion (thresholds[2]) is a cutoff for the difference in log-likelihood for the best genotype class vs. the MLE. For the three-way model, thresholds[2] = 8 corresponds to the criterion that the best classification must have a lower AIC than the MLE to favor classification over the continuous model.

Details

The AIC for the continuous model accounts for k = 5 estimated paramaters. For the classification model, on a per individual basis, there is only one estimated parameter. Classification could be expanded to include more complex genotypic combinations, but it is not clear how (or whether) to account for additional complexity in model comparison. It is probably advisable to consider testing only classifications reflecting a clear biological question, such as whether F1 hybrids are completely sterile.

Value

A data matrix with 7 named columns and one row per individual.

Best.class	Most likely classification of the individual given the data. Classes are indicated by their expected genomic proportions in order: p11, p22, p33, p12, p13, p23. For example, a pure parental 1 would be c100000, an F1 between parentals 1 and 3 would be c000010, an F2 between parentals 2 and 3 would be c011002, etc. NOTE, if two or more classes are equally likely, only the first will be reported here, but LLD.class will be 0 or NaN.
LL.class	The log-likelihood of the best classification given the data. This will be -Inf if no classification is theoretically possible.

LLD.class	The difference in log-likelihood between the best and second best classification.
LL.max	The log-liklihood of the MLE genomic proportions, not constrained to fit a class (continuous model).
dAIC	Difference in AIC between the MLE and best classification. This will be negative when the AIC for classification (1 parameter) is lower than the AIC for the continuous model MLE (5 parameters).
c1	Logical: TRUE if the best-fit class is supported by more than thresholds[1] log-likelihood units over the second best.
c2	Logical: TRUE if the best-fit class is WITHIN thresholds[2] log-likelihood units of the continuous model MLE.

Author(s)

Ben Fitzpatrick

References

Fitzpatrick, B. M. 2012. Estimating ancestry and heterozygosity of hybrids using molecular markers. BMC Evolutionary Biology 12:131. http://www.biomedcentral.com/1471-2148/12/131

See Also

threeway finds maximum likelihood estimates for non-diagnostic marker data. thirdclass evaluates simple classification of three-way hybrids into parental, F1, F2, and backcross categories. For conventional two-way hybrid zone analyses, see HIC, HIest, HIclass, HItest.

Examples

```
## Not run:
## all possible 2-way crosses after 2 generations
G <- rbind(
rep(1,12),rep(1,12),               # parental 1
rep(2,12),rep(2,12),               # parental 2
rep(3,12),rep(3,12),               # parental 3
rep(1,12),rep(2,12),               # 1 x 2 F1
rep(1:2,each=6),rep(1:2,6),        # 1 x 2 F2
rep(1,12),rep(1:2,6),              # 1 x 1 x 2 BC
rep(2,12),rep(1:2,6),              # 1 x 2 x 2 BC
rep(1,12),rep(3,12),               # 1 x 3 F1
rep(c(1,3),each=6),rep(c(1,3),6),  # 1 x 3 F2
rep(1,12),rep(c(1,3),6),           # 1 x 1 x 3 BC
rep(3,12),rep(c(1,3),6),           # 1 x 3 x 3 BC
rep(2,12),rep(3,12),               # 2 x 3 F1
rep(2:3,each=6),rep(2:3,6),        # 2 x 3 F2
rep(3,12),rep(2:3,6),              # 2 x 3 x 3 BC
rep(2,12),rep(2:3,6)               # 2 x 2 x 3 BC
)

P <- data.frame(Locus=rep(1:12,each=3),allele=rep(1:3,12),P1=rep(c(1,0,0),12),
P2=rep(c(0,1,0),12),P3=rep(c(0,0,1),12))
```

```
Est <- HIC3(G,P)
Class <- thirdclass(G,P)
HItest3(Class,Est)

## now for some three-way mixes
G3 <- matrix(1+rbinom(200,2,.5),ncol=10)
Est3 <- HIC3(G3,P)
Class3 <- thirdclass(G3,P)
HItest3(Class3,Est3)    # usually all classifications will be impossible because all
# individuals will have nonzero contributions from each of the
# three parentals

## bias toward parental 1
G3 <- matrix(1+rbinom(200,2,.25),ncol=10)
Est3 <- HIC3(G3,P)
Class3 <- thirdclass(G3,P)
HItest3(Class3,Est3)  # now you might have a few that look like F2's
# between 1 and 2 (c110200)

## End(Not run)
```

`longX2` *Longs heterogeneity test for admixture proportions in a single population.*

Description

For hybrid individuals genotyped with diagnostic markers (one allele fixed in each parental lineage), the null hypothesis that variation in allele frequencies is explained by sampling and drift (no selection) is evaluated with a chi-squared test statistic.

Usage

```
longX2(Freqs)
```

Arguments

`Freqs`	A vector of allele frequencies (one estimate per locus in a single population).

Value

A list with the following items is returned.

`test`	a list including the global test statistic, degrees of freedom, and p-value.
`chisq.res`	Chi-squared residuals for each marker

Author(s)

Ben Fitzpatrick

References

Long, J. C. (1991). The genetic structure of admixed populations. Genetics, 127:417-428. Fitzpatrick, B. M., Johnson, J. R., Kump, D. K., Shaffer, H. B., Smith, J. J., and Voss, S. R. (2009). Rapid fixation of non-native alleles revealed by genome-wide snp analysis of hybrid tiger salamanders. BMC Evolutionary Biology, 9:176.

Examples

```
longX2(c(0.95,rbeta(20,2,5)))
```

`spatial.AD` *Simulate admixture in continuous, 2-dimensional space*

Description

Simulate admixture dynamics with or without selection on a few loci. Although space matters for local density-dependent competition, mating and dispersal are random (uniform) with respect to space.

Usage

```
spatial.AD(minX ,minY, maxX, maxY, XY, Genotypes,
beta=0,sel=0, mid=0,
h=0, DM = matrix(0,ncol=3,nrow=3),
sigmac, R, M, gens,
immigrants=FALSE,plotgrowth=FALSE,m=0.10)
```

Arguments

`minX, minY, maxX, maxY`	Limits of the model space in x and y dimensions.
`XY`	Matrix of initial x,y coordinates of individual organinisms.
`Genotypes`	Matrix of genotypes of initial organisms. Each genotype should be coded as 0, 0.5, or 1 for the frequency of alleles derived from one parental population. Rows are organisms, columns are unlinked loci. The first four loci can cause fitness variation.
`beta`	Steepness of an environmental gradient affecting the first locus.
`sel`	Strength of environmental selection affecting the first locus.
`mid`	Midpoint of the environmental gradient affecting the first locus.
`h`	Selection on heterozygotes at the second locus.
`DM`	Matrix of 2-locus fitness values for the 3rd and 4th loci (see `details`).
`sigmac`	Local competition parameter: Standard deviation of Gaussian competition function.
`R`	Instantaneous growth rate of the Beverton-Holt model.

M	Determines the local carrying capacity of the Beverton-Holt (K = (R-1)*M).
gens	Number of generations to simulate.
immigrants	If FALSE, the model space is closed to immigration and all boundaries are reflecting. If TRUE, the model is open to immigrants from pure parental populations at each edge of the x-dimension. If TRUE, m controls the edge dynamic (see below).
plotgrowth	If TRUE, the population size at each generation will be plotted.
m	Immigration parameter. If immigrants=TRUE, any individual within m/2 of each edge in the x-dimension will be replaced by a pure parental genotype.

Details

For the DM incompatibility, the matrix of fitnesses is 3x3, with rows corresponding to the first DM locus and columns correspinding to the second DM locus. Entries are W[i,j], where i and j index genotypes 0, 1, and 2 at the first and second locus, respectively. See example.

Value

A list with

XY	The x,y coordinates of the diploid individuals in the final generation.
Genotypes	The genotypes of the diploid individuals (rows) in the final generation.
mothers	The genotypes of the successful mothers in the next-to-last generation (roughly, an "after selection" sample from that generation).

Author(s)

Benjamin M. Fitzpatrick

References

Fitzpatrick, B. M. Alternative forms for genomic clines. In review

See Also

See spatial.HZ for a version with limited dispersal and mating distances. The simulated data can be analyzed with Cline.fit, but the genotypes must be multiplied by 2.

Examples

```
## Not run:
# define space:
minX <- minY <- -3
maxX <- maxY <- 3
# 100 individuals randomly placed:
XY <- cbind(runif(100,minX,maxX),runif(100,minY,maxY))
# simulate secondary contact by sorting aling the x dimension and assigning parental genotypes on each side of the
XY <- XY[order(XY[,1]),]
Genotypes <- rbind(matrix(0,nrow=sum(XY[,1]<=0),ncol=10),matrix(1,nrow=sum(XY[,1]>0),ncol=10))
```

```
# competition parameters:
 sigmac <- 0.2; R <- 1.75; M <- 5

# selection, inlcuding heterozygote disadvantage at locus 2 and a DM incompatibility between 3 and 4:
beta <- 0
sel <- 0
mid <- 0
h <- 0.4
DM <- rbind(
c(0,0.2,0.4),
c(0,0.0,0.2),
c(0,0.0,0.0))

# simulate 10 generations, open to immigration:
G10 <- spatial.AD(minX,minY,maxX,maxY,XY,Genotypes,beta,sel,mid,h,DM,sigmac,R,M,gens=10,immigrants=TRUE)


## End(Not run)
```

`spatial.HZ` *Simulate a hybrid zone in continuous, 2-dimensional space*

Description

Simulate hybrid zone dynamics with or without selection on a few loci.

Usage

```
spatial.HZ(minX ,minY, maxX, maxY, XY, Genotypes,
beta=0,sel=0, mid=0,
h=0, DM = matrix(0,ncol=3,nrow=3),
sigmad, sigmac, sigmam, R, M, gens,
immigrants=FALSE,plotgrowth=FALSE,m=0.10)
```

Arguments

`minX, minY, maxX, maxY`	Limits of the model space in x and y dimensions.
`XY`	Matrix of initial x,y coordinates of individual organinisms.
`Genotypes`	Matrix of genotypes of initial organisms. Each genotype should be coded as 0, 0.5, or 1 for the frequency of alleles derived from one parental population. Rows are organisms, columns are unlinked loci. The first four loci can cause fitness variation.
`beta`	Steepness of an environmental gradient affecting the first locus.
`sel`	Strength of environmental selection affecting the first locus.
`mid`	Midpoint of the environmental gradient affecting the first locus.
`h`	Selection on heterozygotes at the second locus.

DM	Matrix of 2-locus fitness values for the 3rd and 4th loci (see details).
sigmad	Dispersal parameter: Standard deviation of mother-offspring distance.
sigmac	Local competition parameter: Standard deviation of Gaussian competition function.
sigmam	Mating parameter: Standard deviation of distance between mates.
R	Instantaneous growth rate of the Beverton-Holt model.
M	Determines the local carrying capacity of the Beverton-Holt (K = (R-1)*M).
gens	Number of generations to simulate.
immigrants	If FALSE, the model space is closed to immigration and all boundaries are reflecting. If TRUE, the model is open to immigrants from pure parental populations at each edge of the x-dimension. If TRUE, m controls the edge dynamic (see below).
plotgrowth	If TRUE, the population size at each generation will be plotted.
m	Immigration parameter. If immigrants=TRUE, any individual within m/2 of each edge in the x-dimension will be replaced by a pure parental genotype.

Details

For the DM incompatibility, the matrix of fitnesses is 3x3, with rows corresponding to the first DM locus and columns correspinding to the second DM locus. Entries are W[i,j], where i and j index genotypes 0, 1, and 2 at the first and second locus, respectively. See example.

Value

A list with

XY	The x,y coordinates of the diploid individuals in the final generation.
Genotypes	The genotypes of the diploid individuals (rows) in the final generation.
mothers	The genotypes of the successful mothers in the next-to-last generation (roughly, an "after selection" sample from that generation).

Author(s)

Benjamin M. Fitzpatrick

References

Fitzpatrick, B. M. Alternative forms for genomic clines. In review

See Also

See spatial.AD for a version with uniformly random mating and dispersal (panmixia). The simulated data can be analyzed with Cline.fit, but the genotypes must be multiplied by 2.

Examples

```
## Not run:
# define space:
minX <- minY <- -3
maxX <- maxY <- 3
# 100 individuals randomly placed:
XY <- cbind(runif(100,minX,maxX),runif(100,minY,maxY))
# simulate secondary contact by sorting aling the x dimension and assigning parental genotypes on each side of the
XY <- XY[order(XY[,1]),]
Genotypes <- rbind(matrix(0,nrow=sum(XY[,1]<=0),ncol=10),matrix(1,nrow=sum(XY[,1]>0),ncol=10))
# dispersal and competition parameters:
sigmad <- 0.3; sigmac <- 0.2; sigmam <- 0.3; R <- 1.75; M <- 5

# selection, inlcuding heterozygote disadvantage at locus 2 and a DM incompatibility between 3 and 4:
beta <- 0
sel <- 0
mid <- 0
h <- 0.4
DM <- rbind(
c(0,0.2,0.4),
c(0,0.0,0.2),
c(0,0.0,0.0))

# simulate 10 generations, open to immigration:
G10 <- spatial.HZ(minX,minY,maxX,maxY,XY,Genotypes,beta,sel,mid,h,DM,sigmad,sigmac,sigmam,R,M,gens=10,immigrant

## End(Not run)
```

`thirdclass` *Calculate likelihoods for early generation hybrid genotype classes in a three-way hybrid zone.*

Description

`thirdclass` uses genetic marker data and parental allele frequencies to calculate the likelihoods for each of the 15 diploid genotype classes possible in the first two generations of admixture (each parental, F1, F2, and each backcross) for each pair of parental lineages.

Usage

```
thirdclass(G, P, type = "codominant")
```

Arguments

`G`	A matrix or data frame of genetic marker data. Each column is a locus. For `type="dominant"`, there should be one row per individual. For `type="codominant"`, each individual is to be represented in consecutive rows (one for each allele).

P	A matrix or data frame with the following columns (order is important!): Locus name, Allele name, P1 allele frequency, P2 allele frequency, P3 allele frequency. For type="dominant", there should be one row per locus, giving the frequencies of the dominant or "1" allele. For type="codominant" there should be a separate row for each allele AND the Allele names should match the data in G.
type	A string representing the data type. The options are "codominant" and "dominant".

Details

The function evaluates genotype classes including only two parental lineages in the context of the possibility of a third parental lineage. It would be straightforward but tedious to extend it to include complex classes such as the offspring of a 1x2 F1 with a 2x3 F1. However, it is not entirely clear how to account for the additional complexity introduced by allowing more classes.

Value

A data matrix with the following columns (one row per individual)

c100000	log-likelihood for genotype class expected for pure parental P1
c010000	log-likelihood for genotype class expected for P2
c001000	log-likelihood for genotype class expected for P3
c000100	log-likelihood for genotype class expected for 1x2 F1 hybrids
c110200	log-likelihood for genotype class expected for 1x2 F2 hybrids
c100100	log-likelihood for genotype class expected for 1x1x2 backcrosses
c010100	log-likelihood for genotype class expected for 1x2x2 backcrosses
c000010	log-likelihood for genotype class expected for 1x3 F1 hybrids
c101020	log-likelihood for genotype class expected for 1x3 F2 hybrids
c100010	log-likelihood for genotype class expected for 1x1x3 backcrosses
c001010	log-likelihood for genotype class expected for 1x3x3 backcrosses
c000001	log-likelihood for genotype class expected for 2x3 F1 hybrids
c011002	log-likelihood for genotype class expected for 2x3 F2 hybrids
c010001	log-likelihood for genotype class expected for 2x2x3 backcrosses
c001001	log-likelihood for genotype class expected for 2x3x3 backcrosses
Best	The class with the highest likelihood of the 15
LLD	The difference in log-likelihood between the best and second-best fit class. This can be used as a criterion for deciding whether the best fit class is enough better to reject the alternatives.

Author(s)

Ben Fitzpatrick

References

Fitzpatrick, B. M. 2012. Estimating ancestry and heterozygosity of hybrids using molecular markers. BMC Evolutionary Biology 12:131. http://www.biomedcentral.com/1471-2148/12/131

See Also

threeway finds maximum likelihood estimates for non-diagnostic marker data. HItest3 compares classification to the continuous model MLE. For conventional two-way hybrid zone analyses, see HIC, HIest, HIclass, HItest.

Examples

```
## Not run:
## all possible 2-way crosses after 2 generations
G <- rbind(
rep(1,12),rep(1,12),               # parental 1
rep(2,12),rep(2,12),               # parental 2
rep(3,12),rep(3,12),               # parental 3
rep(1,12),rep(2,12),               # 1 x 2 F1
rep(1:2,each=6),rep(1:2,6),        # 1 x 2 F2
rep(1,12),rep(1:2,6),              # 1 x 1 x 2 BC
rep(2,12),rep(1:2,6),              # 1 x 2 x 2 BC
rep(1,12),rep(3,12),               # 1 x 3 F1
rep(c(1,3),each=6),rep(c(1,3),6),  # 1 x 3 F2
rep(1,12),rep(c(1,3),6),           # 1 x 1 x 3 BC
rep(3,12),rep(c(1,3),6),           # 1 x 3 x 3 BC
rep(2,12),rep(3,12),               # 2 x 3 F1
rep(2:3,each=6),rep(2:3,6),        # 2 x 3 F2
rep(3,12),rep(2:3,6),              # 2 x 3 x 3 BC
rep(2,12),rep(2:3,6)               # 2 x 2 x 3 BC
)

P <- data.frame(Locus=rep(1:12,each=3),allele=rep(1:3,12),P1=rep(c(1,0,0),12),
P2=rep(c(0,1,0),12),P3=rep(c(0,0,1),12))

Est <- HIC3(G,P)
Class <- thirdclass(G,P)
HItest3(Class,Est)

## now for some three-way mixes
G3 <- matrix(1+rbinom(200,2,.5),ncol=10)
Est3 <- HIC3(G3,P)
Class3 <- thirdclass(G3,P)
HItest3(Class3,Est3)    # usually all classifications will be impossible because all
# individuals will have nonzero contributions from each of the
# three parentals

## bias toward parental 1
G3 <- matrix(1+rbinom(200,2,.25),ncol=10)
Est3 <- HIC3(G3,P)
Class3 <- thirdclass(G3,P)
HItest3(Class3,Est3)  # now you might have a few that look like F2's
# between 1 and 2 (c110200)

## End(Not run)
```

threeway *Find joint maximum-likelihood estimates of ancestry and heterozygosity for a sample of hybrids in a three-way hybrid zone*

Description

For hybrids with up to three parental lineages, the function estimates genomic proportions and ancestry indices by numerically searching likelihood space.

Usage

```
threeway(G, P, type = "codominant", surf = TRUE, method = "SANN", iterations = 500,
start = rep(1/6, 6), Cscale = NULL, props = NULL,
control = list(fnscale = -1, maxit = iterations))
```

Arguments

G	A matrix or data frame of genetic marker data. Each column is a locus. For type="dominant", there should be one row per individual. For type="codominant", each individual is to be represented in consecutive rows (one for each allele).
P	A matrix or data frame with the following columns (order is important!): Locus name, Allele name, P1 allele frequency, P2 allele frequency, P3 allele frequency. For type="dominant", there should be one row per locus, giving the frequencies of the dominant or "1" allele. For type="codominant" there should be a separate row for each allele AND the Allele names should match the data in G.
type	A string representing the data type. The options are "codominant" or "dominant".
surf	Logical: should the function find starting values by evaluating likelihoods on a grid?
method	Optimization method to search for maximum likelihood estimates of ancestry and heterozygosity. Alternatives are "SANN", "L-BFGS-B", "surf", and "mcmc". See details.
iterations	The desired number of MCMC steps to perform when method="mcmc" or "SANN".
start	A vector including the starting values of the six genomic proportions. See details.
Cscale	An integer, controlling the the proposal distribution for method = "SANN" or "mcmc". Smaller values will cause the algorithm to search more broadly, but could make the search inefficient. See details.
props	Optional: a matrix of genomic proportions to evaluate if surf = TRUE or method = "surf". Columns correspond to the six genomic proportions in order: p11, p22, p33, p12, p13, p23.
control	A list of options to be passed to control in the optim function. Whatever else is chosen, be sure fnscale is negative to make optim search for a maximum rather than a minimum. Specifying maxit will override iterations for method = "SANN".

Details

Given three ancestral species or parental populations (P1, P2, and P3), the genome of a hybrid can be described by six genomic proportions: p_{11} = proportion of one's genome that is homozygous for alleles inherited from P1, p_{22} = proportion of one's genome that is homozygous for alleles inherited from P2, p_{33} = proportion of one's genome that is homozygous for alleles inherited from P3, p_{12} = proportion of one's genome that is heterozygous for alleles inhertied from P1 and P2, p_{13} = proportion of one's genome that is heterozygous for alleles inhertied from P1 and P3, p_{23} = proportion of one's genome that is heterozygous for alleles inhertied from P2 and P3.

Currently, the function provides four methods for searching the likelihood space (a 5-simplex with vertices wherever each genomic proportion is equal to 1.0). `method = "SANN"` is probably the best; it uses the general purpose optimization function `optim` with its simulated annealing algorithm. For this estimation problem, a custom proposal function is passed to the option `gr`. This proposal function draws new genomic proportions from a 6-dimensional Dirichlet distribution centered on the old genomic proportions. The concentration of the proposal distribution is controlled by `Cscale`; the larger this value, the more the proposal distribution is concentrated near the current state.

`method = "mcmc"` uses a Markov-Chain Monte Carlo with Metropolis-Hastings sampling to explore the likelihood space. It also uses the Dirichlet proposal distribution, and could be useful (with some modification of the code) for generating posterior distributions. `method = "SANN"` is probaby superior for simply finding the MLE.

`method = "L-BFGS-B"` also uses `optim`, but with a quasi-Newton likelihood search algorithm to look for the maximum likelihood. This method is relatively fast, but it can miss the MLE if it is near an edge of the sample space.

`"surf"`, finds all likelihoods on a 6-dimensional grid defined by `props` and chooses the maximum. By default, `props` is all possible combinations of 10 equally spaced proportions (from 0 to 1), subject to the constraint that they add to 1. By itself, this method is not going to find the MLE unless the MLE happens to be one of the grid points. However, using the option `surf = TRUE` with the `SANN` or `mcmc` methods can improve efficiency by initiating the search at the grid point nearest the MLE.

Value

A matrix with 10 named columns for each individual, containing estimated genomic proportions, ancestry indices, and the log-likelihood:

`p11`	Proportion of markers homozygous for lineage 1 alleles
`p22`	Proportion of markers homozygous for lineage 2 alleles
`p33`	Proportion of markers homozygous for lineage 3 alleles
`p12`	Proportion of markers heterozygous for lineage 1 and 2 alleles
`p13`	Proportion of markers heterozygous for lineage 1 and 3 alleles
`p23`	Proportion of markers heterozygous for lineage 2 and 3 alleles
`S1`	Lineage 1 ancestry index: proportion of alleles derived from parental lineage 1
`S2`	Lineage 2 ancestry index: proportion of alleles derived from parental lineage 2
`S3`	Lineage 3 ancestry index: proportion of alleles derived from parental lineage 3
`logLik`	log-likelihood of the genomic proportions given the individual marker data

Author(s)

Ben Fitzpatrick

References

Fitzpatrick, B. M. 2012. Estimating ancestry and heterozygosity of hybrids using molecular markers. BMC Evolutionary Biology 12:131. http://www.biomedcentral.com/1471-2148/12/131

See Also

HIC3 calculates closed-form maximum likelihood estimates for diagnostic marker data. thirdclass and HItest3 evaluate simple classification of three-way hybrids into parental, F1, F2, and backcross categories. For conventional two-way hybrid zone analyses, see HIC, HIest, HIclass, HItest.

Examples

```
## Not run:
## all possible 2-way crosses after 2 generations
G <- rbind(
rep(1,12),rep(1,12),               # parental 1
rep(2,12),rep(2,12),               # parental 2
rep(3,12),rep(3,12),               # parental 3
rep(1,12),rep(2,12),               # 1 x 2 F1
rep(1:2,each=6),rep(1:2,6),        # 1 x 2 F2
rep(1,12),rep(1:2,6),              # 1 x 1 x 2 BC
rep(2,12),rep(1:2,6),              # 1 x 2 x 2 BC
rep(1,12),rep(3,12),               # 1 x 3 F1
rep(c(1,3),each=6),rep(c(1,3),6),  # 1 x 3 F2
rep(1,12),rep(c(1,3),6),           # 1 x 1 x 3 BC
rep(3,12),rep(c(1,3),6),           # 1 x 3 x 3 BC
rep(2,12),rep(3,12),               # 2 x 3 F1
rep(2:3,each=6),rep(2:3,6),        # 2 x 3 F2
rep(3,12),rep(2:3,6),              # 2 x 3 x 3 BC
rep(2,12),rep(2:3,6)               # 2 x 2 x 3 BC
)

P <- data.frame(Locus=rep(1:12,each=3),allele=rep(1:3,12),
P1=rep(c(1,0,0),12),P2=rep(c(0,1,0),12),P3=rep(c(0,0,1),12))

mle.o <- threeway(G,P,surf=FALSE,iterations=99)
mle.c <- HIC3(G,P)

# compare the optimization (mle.o) to the closed-form (mle.c):
# 99 iterations is not enough to converge on the known true values.
# Try setting surf=TRUE and/or increasing iterations.

## End(Not run)
```

Index