Package ‘bgchm’

October 10, 2024

Title Bayesian Analyses of Hybrid Zones with Hamiltonian Monte Carlo

Version 0.0.0.9000

Description This is an R package for Bayesian analyses of population genomic data from hybrid zones, including Bayesian genomic cline analysis, estimation of hybrid indexes and ancestry classes, some geographic cline analyses, and accessory plotting functions. This package using Hamiltonian Monte Carlo (HMC) for sampling posterior distributions, with HMC sampling implemented via Stan.

License GPL (>= 3)

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

Biarch true

Depends R (>= 3.4.0)

Imports methods,
Rcpp (>= 0.12.0),
RcppParallel (>= 5.0.1),
rstan (>= 2.18.1),
rstantools (>= 2.3.1.1)

LinkingTo BH (>= 1.66.0),
Rcpp (>= 0.12.0),
RcppEigen (>= 0.3.3.3.0),
RcppParallel (>= 5.0.1),
rstan (>= 2.18.1),
StanHeaders (>= 2.18.0)

SystemRequirements GNU make

R topics documented:

`bgchm-package` *The 'bgchm' package.*

Description

This is an R package for Bayesian analyses of population genomic data from hybrid zones, including Bayesian genomic cline analysis, estimation of hybrid indexes and ancestry class proportions, some geographic cline analyses, and accessory plotting functions. This package using Hamiltonian Monte Carlo (HMC) for sampling posterior distributions, with HMC sampling implemented via Stan.

References

Gompert Z, DeRaad D, Buerkle CA. A next generation of hierarchical Bayesian analyses of hybrid zones enables model-based quantification of variation in introgression in R. bioRxiv 2024.03.29.587395. Stan Development Team (2022). RStan: the R interface to Stan. R package version 2.21.7. https://mc-stan.org

`calc_pp` *Function to calcualte posterior probabilities of extreme cline parameter values*

Description

Computes the posterior probability that cline parameter values exceed a user defined null or reference value.

Usage

```
calc_pp(hmc = NULL, param = NULL, greater = TRUE, refval = NULL)
```

Arguments

`hmc`	list object produced by est_genocl or est_geocl.
`param`	character or string specifying the cline parameter to summarize.
`greater`	Boolean indicating whether to compute the posterior probability the value is (i) greater than (TRUE) or (ii) less than (FALSE) the null or reference value.
`refval`	null or reference value for posterior probability calculation.

Details

This function computes the posterior probability the a parameter value exceeds (or is less than) a null or reference value using samples from the parameter posterior distribution. This works with geographic or genomic cline output and for any model parameter in the HMC output based on the name of the parameter in the HMC object (i.e., Stan model). For genomic clines, this includes cline centers (center), slopes (v), cline standard deviations (sc and sv) and cline means (muc and muv) (not all parameters are components of all models). For geographic clines, this includes cline centers (cent), slopes (slope), cline widths (w), the mean cline width (mu), and the standard deviation in cline slopes (sigma). In both cases, the name of the parameter in the HMC object is given in parentheses and should be provided as a character or string to the param argument. Different reference or null values are relevant for different questions or parameters. For example, the expected average genomic cline slope, v, is 1, and thus one could identify loci with steeper (or shallower) genomic clines than the genome-average by computing the posterior probability that the slope, v, exceeds (or is less than) 1.

Value

A vector of posterior probabilities (of length > 1 for vector parameters, such as cline slopes with multiple loci).

References

Gompert Z, DeRaad D, Buerkle CA. A next generation of hierarchical Bayesian analyses of hybrid zones enables model-based quantification of variation in introgression in R. bioRxiv 2024.03.29.587395.

See Also

'rstan::stan' for details on HMC with stan and the rstan HMC output object.

`est_genocl` *Estimate genomic clines using Bayesian HMC*

Description

Uses Hamiltonian Monte Carlo (HMC) for Bayesian inference of genomic clines from genetic data. This fits a hierarchical logit-logistic model for genomic clines with two key parameters, cline center and cline gradient (i.e., slope, inversely proportional to cline width). The model also estimates the cline standard deviations (SDs) across loci, SDc = variation in logit centers and SDv = variation in log10 gradients.

Usage

```
est_genocl(
  Gx = NULL,
  G0 = NULL,
  G1 = NULL,
  p0 = NULL,
  p1 = NULL,
  H = NULL,
  model = "genotype",
  ploidy = "diploid",
  pldat = NULL,
  hier = TRUE,
  SDc = NULL,
  SDv = NULL,
  estMu = FALSE,
  sd0 = 1,
  mu0 = 1,
  n_chains = 4,
  n_iters = 2000,
  p_warmup = 0.5,
  n_thin = 1,
  n_cores = NULL,
  full = TRUE
)
```

Arguments

`Gx` genetic data for putative hybrids in the form of a matrix for known genotypes (rows = individuals, columns = loci), a list of matrixes for genotype likelihoods (same dimensions but one matrix per genotype), or matrix of ancestry for the ancestry model (rows = individuals, columns = loci).

`G0` genetic data for parental reference set 0 formatted as described for Gx.

`G1` genetic data for parental reference set 1 formatted as described for Gx.

`p0` vector of allele frequencies for parental reference set 0 (one entry per locus).

`p1` vector allele frequencies for parental reference set 1 (one entry per locus).

`H` vector of hybrid indexes for the putative hybrids (one entry per locus).

`model` for genetic data, either 'genotype' for known gentoypes, 'glik' for genotype likelihoods, or 'ancestry' for known ancestry.

`ploidy` species ploidy, either all 'diploid' or 'mixed' for diploid and haploid loci or individuals.

`pldat` matrix or list of matrixes with ploidy data for mixed ploidy (rows = individuals, columns = loci) indicating ploidy (2 = diploid, 1 = haploid).

`hier` Boolean, fit hierarchical model (TRUE) that estimates cline SDs or non-hierarchical model that assumes cline SDs are known (FALSE).

`SDc` known cline center SD on logit scale.

`SDv` known cline gradient SD on log10 scale.

`estMu` boolean, estimate genomic cline means (TRUE) or assume a prior mean of 0 (FALSE, the default).

`sd0` standard deviation for the normal prior on the cline standard deviations, default is 1.

`mu0` standard deviation for the normal prior on the cline mean, default is 1.

`n_chains` number of HMC chains for posterior inference.

`n_iters` A positive integer specifying the number of iterations for each chain (including warmup), default is 2000.

`p_warmup` proportion (between 0 and 1) of n_iters to use as warmup (i.e., burnin), default is 0.5.

`n_thin` positive integer, save every n_thin HMC iterations for the posterior, default is 1.

`n_cores` number of cores to use for HMC, leave as NULL to automatically detect the number of cores present (no more than n_chains cores can be used even if more cores are available).

`full` boolean denoting whether (TRUE, the default) or not (FALSE) to return the HMC Stan object with the full set of samples from the posterior.

Details

Clines can be inferred based on known genotypes (model = 'genotype'), genotype likelihoods, (model = 'glik') or known (estimated) ancestry (model = 'ancestry'). Genotypes should be encoded as 0 (homozygote), 1 (heterozygote) and 2 (alternative homozygote). No specific polarization (e.g., minor allele, reference allele, etc.) of 0 vs 2 is required. For haploid loci, use 0 and 1. Genotype likelihoods should be on their natural scale (not on the phred scaled) and the values for each locus for an individual should sum to 1 (i.e., the likelihoods are scaled to be probabilities). The data should be provided as a list of three matrixes, with the matrixes giving the likelihoods for genotypes 0, 1 and 2 respectively. Thus, each matrix will have one row per individual and one column per locus. For haploid loci with genotype likelihoods, you must use the 0 and 1 matrixes to store the likelihoods of the two possible states. For the ancestry model, clines are inferred directly from known local (locus-specific) ancestry rather than from genotype data. Users are free to use whatever software they prefer for local ancestry inference (many exist). In this case, each entry in the individual (rows) by locus (columns) matrix should denote the number of gene copies inherited from parental population 1 (where pure parent 1 corresponds with a hybrid index of 1 and pure parent 0 corresponds with a hybrid index of 0). Haploids can be encoded using 0 and 1. For all models, missing data can be encoded by setting the ploidy for an individual/locus to 0 (this indicates no information, whereas genotype likelihoods encode uncertainty in genotypes).

Hybrid indexes must be provided. Genetic (or ancestry) data for putative hybrids are also always required. For genotype or genotype likelihood models, users must either provide pre-estimated parental allele frequencies or parent genetic (genotypes or genotype likelihoods) that can be used to infer allele frequencies with est_p. Parental data are not required for the ancestry model.

Ploidy data are only required for the mixed ploidy data. In this case, there should be one matrix for the hybrids or a list of matrixes for the hybrids (1st matrix) and each parent (2nd and 3rd matrixes, with parent 0 first). The latter is required for the genotype or genotype likelihood models if parental allele frequencies are not provided. The matrixes indicate whether each locus (column) for each individual (row) is diploid (2) or haploid (1) (use 0 for missing data).

Value

A list of parameter estimates and full HMC results from stan, this includes cline parameters (center and gradient), and, for hierarchical models, standard deviations describing variability in clines across loci (SDc and SDv). Mean values for cline center (Muc) and slope (Muv) are also provided if estMu is set to TRUE. Parameter estimates are provided as a point estimate (median of the posterior) and 90% equal-tail probability intervals (5th and 95th quantiles of the posterior distribution). These are provided as a vector or matrix depending on the dimensionality of the parameter. The full HMC output from rstan is provided as the final element in the list. This can be used for HMC diagnostics and to extract other model outputs not provided by default.

References

Gompert Z, DeRaad D, Buerkle CA. A next generation of hierarchical Bayesian analyses of hybrid zones enables model-based quantification of variation in introgression in R. bioRxiv 2024.03.29.587395.

See Also

'rstan::stan' for details on HMC with stan and the rstan HMC output object.

Examples

```
## Not run:
## load the data set
data(genotypes)
## this includes three objects, GenHybrids, GenP0, and GenP1

## estimate parental allele frequencies
p_out<-est_p(G0=GenP0,G1=GenP1,model="genotype",ploidy="diploid")

## estimate hybrid indexes, uses default HMC settings
## and uses point estimates (posterior medians) of allele frequencies
h_out<-est_hi(Gx=GenHybrids,p0=p_out$p0[,1],p1=p_out$p1[,1],model="genotype",ploidy="diploid")

## fit a hierarchical genomic cline model for all 51 loci using the estimated
## hybrid indexes and parental allele frequencies (point estimates)
## use 4000 iterations and 2000 warmup to make sure we get a nice effective sample size
gc_out<-est_genocl(Gx=GenHybrids,p0=p_out$p0[,1],p1=p_out$p1[,1],H=h_out$hi[,1],model="genotype",ploidy="diplo

## End(Not run)
```

est_geocl *Function to estimate geographic clines for a set of genetic loci or hybrid index*

Description

Uses Hamiltonian Monte Carlo (HMC) for Bayesian inference of geographic clines from genetic data.

Usage

```
est_geocl(
  G = NULL,
  P = NULL,
  Geo = NULL,
  Ids = NULL,
  model = "genotype",
  ploidy = "diploid",
  pldat = NULL,
  hier = TRUE,
  prec = 0.001,
  y_lb = -2,
  y_ub = 2,
  gamma_a = 0.1,
  gamma_b = 0.01,
  n_chains = 4,
  n_iters = 2000,
  p_warmup = 0.5,
  n_thin = 1,
  n_cores = NULL,
  full = TRUE
)
```

Arguments

`G` genetic data for the sampled hybrid zone in the form of a matrix for known genotypes (rows = individuals, columns = loci), a list of matrixes for genotype likelihoods (same dimensions but one matrix per genotype), or matrix of ancestry for the ancestry model (rows = individuals, columns = loci).

`P` matrix of allele frequencies for the hybrid zone with rows denoting sampled populations (localities) and columns denoting loci.

`Geo` vector of geographic coordinates for the sampled populations.

`Ids` indexes designating which population each individual belongs to. This should be provided as a single vector (length equal to the number of sampled individuals). The indexes should range from 1 to the number of populations.

`model` for genetic data, either 'genotype' for known genotypes or 'glik' for genotype likelihoods.

`ploidy` specifies ploidy, either all 'diploid' or 'mixed' for diploid and haploid loci or individuals.

`pldat` matrix or list of matrixes of ploidy data for mixed ploidy (rows = individuals, columns = loci) indicating ploidy (2 = diploid, 1 = haploid).

`hier` Boolean, fit hierarchical model (TRUE) that estimates cline slope SDs or non-hierarchical model for one locus with (relatively) uninformative priors (FALSE).

`prec` approximate precision for known (or estimated) allele frequencies, which should be set to about 1/2N (use the average 2N across populations). Do not set this to 0; this can result in errors when working with the log of the allele frequencies.

`y_lb` minimum value of logit allele frequencies to include in the model (the default of -2 is a good choice).

`y_ub` minimum value of logit allele frequencies to include in the model (the default of 2 is a good choice).

`gamma_a` alpha parameter for gamma priors on the error standard deviation and standard deviation on the prior for slope (default = 0.1).

`gamma_b` beta parameter for gamma priors on the error standard deviation and standard deviation on the prior for slope (default = 0.01).

`n_chains` number of HMC chains for posterior inference.

`n_iters` A positive integer specifying the number of iterations for each chain (including warmup), default is 2000.

`p_warmup` proportion (between 0 and 1) of n_iters to use as warmup (i.e., burnin), default is 0.5.

`n_thin` positive integer, save every n_thin HMC iterations for the posterior, default is 1.

`n_cores` number of cores to use for HMC, leave as NULL to automatically detect the number of cores present (no more than n_chains cores can be used even if more cores are available).

`full` boolean denoting whether (TRUE, the default) or not (FALSE) to return the HMC Stan object with the full set of samples from the posterior.

Details

Geographic clines are estimated from population (deme) allele frequencies. This is done using a linear model for the logit of the allele frequencies (a sigmoid cline on the natural scale becomes linear on the logit scale). Users can provide allele frequency estimates or the allele frequencies can be estimate from genotypic data. In the latter case, allele frequencies are first estimated based on known genotypes (model = 'genotype') or genotype likelihoods (model = 'glik') using an analytical solution for the posterior. Genotypes should be encoded as 0 (homozygote), 1 (heterozygote) and 2 (alternative homozygote). No specific polarization (e.g., minor allele, reference allele, etc.) of 0 vs 2 is required. For haploid loci, use 0 and 1. Genotype likelihoods should be on their natural scale (not phred scaled) and the values for each locus for an individual should sum to 1 (i.e., the likelihoods are scaled to be probabilities). The data should be provided as a list of three matrixes, with the matrixes giving the likelihoods for genotypes 0, 1 and 2 respectively. Thus, each matrix will have one row per individual and one column per locus. For haploid loci with genotype likelihoods, use the 0 and 1 matrixes to store the likelihoods of the two possible states. If provided directly, allele frequencies should be given as a matrix, with one column per locus (assumes bi-allelic SNPs or the equivalent) and one row per population (deme).

Ploidy data are only required for mixed ploidy genotypic data (they are not used if allele frequencies are provided directly). In this case, there should be one matrix with the same dimensions as the genetic data. The values in the matrix indicate whether each locus (column) for each individual (row) is diploid (2) or haploid (1). Ploidy can be set to 0 to denote missing data.

The model assumes organisms have been sampled from populations (demes) along a 1D transect through a hybrid zone. Various approaches exist for approximating a 2D sampling scheme in 1D and can be used to transform coordinates to a single dimension. No specific coordinate units are expected, and coordinates are always centered (given a mean of 0) prior to analysis (this is done internally if not done by the user). If population allele frequencies are given directly, populations are assumed to be in the same order in the Geo vector and allele frequency matrix. If genotypic data are provided, and additional object, Ids, is required that indicates which population (numbered 1 to the number of populations and following the order in Geo) each individual belongs to.

The model works with the logit of the allele frequencies. Consequently, allele frequencies of 0 are not allowed (these will cause an error). The value specified by prec will be added to allele frequencies of 0 and subtracted from allele frequencies of 1. This prevents problems with taking logs and also is meant to reflect that fact that one cannot be certain an allele is not present in a population. We recommend setting prec to 1/2N, where N is the mean (or median) sample size across demes. Single and multilocus clines should be well approximated by a linear function for the logit allele frequencies near the center of the hybrid zone; this is the reason for only analyzing populations with intermediate allele frequencies (logit p between y_lb and y_ub, -2 and 2 by default, or p of about 0.11 to 0.88)

Value

A list of parameter estimates and full HMC results from stan. Estimates are provided for the cline width on the natural scale for each locus (w), the slope on the logit scale (slope), the cline center based on the centered geographic coordinates on the logit scale (center), and the mean (mu) and standard deviation (sigma) for cline widths on the logit scale (the latter two only apply to the hierarchical model). Parameter estimates are provided as a point estimate (median of the posterior), 90% equal-tail probability intervals (2.5th and 97.5th quantiles of the posterior distribution), and 95% equal-tail probability intervals (2.5th and 97.5th quantiles of the posterior distribution). These are provided as a vector or matrix depending on the dimensionality of the parameter. The full HMC output from rstan is provided as the final element in the list. This can be used for HMC diagnostics and to extract other model outputs not provided by default.

References

Gompert Z, DeRaad D, Buerkle CA. A next generation of hierarchical Bayesian analyses of hybrid zones enables model-based quantification of variation in introgression in R. bioRxiv 2024.03.29.587395.

See Also

’rstan::stan’ for details on HMC with stan and the rstan HMC output object.

Examples

```
## Not run:
## load the data set
data(pfreqs)
## this includes one object, a matrix P with allele frequencies
## 110 rows = demes, 51 columns = loci

## use standardized deme numbers as geographic coordinates
x<-1:110
geo<-(x-mean(x))/sd(x)

## fit the geographic cline model
o<-est_geocl(P=P,Geo=geo,prec=0.01,y_lb=-2,y_ub=2,hier=TRUE,n_iters=5000)
```

```
## plot clines on logit scale, which should be linear
plot(geo,o$cent[1,1] + o$slope[1,1] * geo,type='l',ylim=c(-15,15),ylab=”Logit allele frequency”,xlab=”Deme number
axes=FALSE)
axis(1,at=geo[seq(5,110,5)],x[seq(5,110,5)])
axis(2)
box()
for(i in 2:51){
lines(geo,o$cent[i,1] + o$slope[i,1] * geo)
}

## End(Not run)
```

est_hi *Function to estimate hybrid index*

Description

Uses Hamiltonian Monte Carlo (HMC) for Bayesian inference of hybrid indexes from genetic data.

Usage

```
est_hi(
  Gx = NULL,
  G0 = NULL,
  G1 = NULL,
  p0 = NULL,
  p1 = NULL,
  model = ”genotype”,
  ploidy = ”diploid”,
  pldat = NULL,
  n_chains = 4,
  n_iters = 2000,
  p_warmup = 0.5,
  n_thin = 1,
  n_cores = NULL,
  full = TRUE
)
```

Arguments

`Gx` genetic data for putative hybrids in the form of a matrix for known genotypes (rows = individuals, columns = loci), a list of matrixes for genotype likelihoods (same dimensions but one matrix per genotype), or matrix of ancestry for the ancestry model (rows = individuals, columns = loci).

`G0` genetic data for parental reference set 0 formatted as described for Gx.

`G1` genetic data for parental reference set 1 formatted as described for Gx.

`p0` vector of allele frequencies for parental reference set 0 (one entry per locus).

`p1` vector of allele frequencies for parental reference set 1 (one entry per locus).

`model` for genetic data, either 'genotype' for known genotypes, 'glik' for genotype likelihoods, or 'ancestry' for known ancestry.

`ploidy` species ploidy, either all 'diploid' or 'mixed' for diploid and haploid loci or individuals.

`pldat` matrix or list of matrixes of ploidy data for mixed ploidy (rows = individuals, columns = loci) indicating ploidy (2 = diploid, 1 = haploid).

`n_chains` number of HMC chains for posterior inference.

`n_iters` A positive integer specifying the number of iterations for each chain (including warmup), default is 2000.

`p_warmup` proportion (between 0 and 1) of n_iters to use as warmup (i.e., burnin), default is 0.5.

`n_thin` positive integer, save every n_thin HMC iterations for the posterior, default is 1.

`n_cores` number of cores to use for HMC, leave as NULL to automatically detect the number of cores present (no more than n_chains cores can be used even if available).

`full` boolean denoting whether (TRUE, the default) or not (FALSE) to return the HMC Stan object with the full set of samples from the posterior.

Details

Hybrid indexes can be estimated from known genotypes (model = 'genotype'), genotype likelihoods (model = 'glik'), or known (estimated) ancestry (model = 'ancestry'). Genotypes should be encoded as 0 (homozygote), 1 (heterozygote) and 2 (alternative homozygote). No specific polarization (e.g., minor allele, reference allele, etc.) of 0 vs 2 is required. For haploid loci, use 0 and 1. Genotype likelihoods should be on their natural scale (not phred scaled) and the values for each locus for an individual should sum to 1. The data should be provided as a list of three matrixes, with the matrixes giving the likelihoods for genotypes 0, 1 and 2 respectively. Thus, each matrix will have one row per individual and one column per locus. For haploid loci with genotype likelihoods, you must use the 0 and 1 matrixes to store the likelihoods of the two possible states. For the ancestry model, hybrid indexes are inferred directly from known local (locus-specific) ancestry rather than from genotype data. Users are free to use whatever software they prefer for local ancestry inference (many exist). In this case, each entry in the individual (rows) by locus (columns) matrix should denote the number of gene copies inherited from parental population 1 (where pure parent 1 corresponds with a hybrid index of 1 and pure parent 0 corresponds with a hybrid index of 0). Haploids can be encoded using 0 and 1. For all models, missing data can be encoded by setting the ploidy for an individual/locus to 0 (this indicates no information, whereas genotype likelihoods encode uncertainty in genotypes).

Hybrid genetic (or ancestry) data are always required. For genotype or genotype likelihood models, users must either provide pre-estimated parental allele frequencies or parent genetic (genotypes or genotype likelihoods) that can be used to infer allele frequencies. Parental data are not required for the ancestry model

Ploidy data are only required for the mixed ploidy data. In this case, there should be one matrix for the hybrids or a list of matrixes for the hybrids (1st matrix) and each parent (2nd and 3rd matrixes, with parent 0 first). The latter is required for the genotype or genotype likelihood models if parental allele frequencies are not provided. The matrixes indicate whether each locus (column) for each individual (row) is diploid (2) or haploid (1) (use 0 for missing data).

Value

A list of parameter estimates (hi = hybrid indexes) and full HMC results from stan. Parameter estimates are provided as a point estimate (median of the posterior), 90% equal-tail probability intervals (5th and 95th quantiles of the posterior distribution), 95% equal-tail probability intervals (2.5th and 97.5th quantiles of the posterior distribution). These are provided as a vector or matrix depending on the dimensionality of the parameter. The full HMC output from rstan is provided as the final element in the list. This can be used for HMC diagnostics and to extract other model outputs not provided by default.

A list of parameter estimates and full HMC results from stan, this includes cline parameters (center and gradient), and, for hierarchical models, standard deviations describing variability in clines across loci (SDc and SDv). Parameter estimates are provided as a point estimate (median of the posterior), 90% equal-tail probability intervals (5th and 95th quantiles of the posterior distribution), and 95% equal-tail probability intervals (2.5th and 97.5th quantiles of the posterior distribution). These are provided as a vector or matrix depending on the dimensionality of the parameter. The full HMC output from rstan is provided as the final element in the list. This can be used for HMC diagnostics and to extract other model outputs not provided by default.

References

Gompert Z, DeRaad D, Buerkle CA. A next generation of hierarchical Bayesian analyses of hybrid zones enables model-based quantification of variation in introgression in R. bioRxiv 2024.03.29.587395.

See Also

'rstan::stan' for details on HMC with stan and the rstan HMC output object.

Examples

```
## Not run:
## load the data set
data(genotypes)
## this includes three objects, GenHybrids, GenP0, and GenP1

## estimate parental allele frequencies, uses default HMC settings
p_out<-est_p(G0=GenP0,G1=GenP1,model="genotype",ploidy="diploid")

## estimate hybrid indexes, uses default HMC settings
## and uses point estimates (posterior medians) of allele frequencies
h_out<-est_hi(Gx=GenHybrids,p0=p_out$p0[,1],p1=p_out$p1[,1],model="genotype",ploidy="diploid")

## End(Not run)
```

est_p *Function to estimate parental allele frequencies*

Description

Bayesian inference parental allele frequencies from genetic data.

Usage

```
est_p(
  G0 = NULL,
  G1 = NULL,
  model = "genotype",
  ploidy = "diploid",
  pldat = NULL,
  n_chains = 4,
  n_iters = 2000,
  p_warmup = 0.5,
  n_thin = 1,
  n_cores = NULL,
  HMC = FALSE,
  full = TRUE
)
```

Arguments

`G0` genetic data for parental reference set 0 in the form of a matrix for known genotypes (rows = individuals, columns = loci) or a list of matrixes for genotype likelihoods (same dimensions but one matrix per genotype).

`G1` genetic data for parental reference set 1 formatted as described for G0.

`model` for genetic data, either 'genotype' for known gentoypes or 'glik' for genotype likelihoods.

`ploidy` species ploidy, either all 'diploid' or 'mixed' for diploid and haploid loci or individuals.

`pldat` list of matrixes of ploidy data for mixed ploidy (rows = individuals, columns = loci) indicating ploidy (2 = diploid, 1 = haploid); matrix 2 in the list is for parental reference set 0 and matrix 3 is for parental reference set 1 (matrix 1 is reserved for the hybrids, which are not included in this analysis).

`n_chains` number of HMC chains for posterior inference.

`n_iters` A positive integer specifying the number of iterations for each chain (including warmup), default is 2000.

`p_warmup` proportion (between 0 and 1) of n_iters to use as warmup (i.e., burnin), default is 0.5.

`n_thin` positive integer, save every n_thin HMC iterations for the posterior, default is 1.

`n_cores`	number of cores to use for HMC, leave as NULL to automatically detect the number of cores present (no more than n_chains cores can be used even if available).
`HMC`	boolean denotes whether to use HMC (TRUE) or an analytical solution (FALSE) for the posterior for the genotype likelihood model (the analytical solution is always used for known genotypes), default is TRUE.
`full`	boolean denoting whether (TRUE, the default) or not (FALSE) to return the HMC Stan object with the full set of samples from the posterior (only relevant if HMC = TRUE).

Details

Parental allele frequencies can be inferred based on known genotypes (model = 'genotype') or genotype likelihoods (model = 'glik'). With known genotypes, the posterior is computed exactly (it is a beta distribution given the binomial likelihood and a beta prior). With genotype likelihoods, the user can decide to compute the exact posterior based on a point estimate of the allele counts from the genotype likelihoods (HMC=FALSE), or to conduct the full Hamiltonian Monte Carlo analyses (HMC=TRUE). The latter provides a better characterization of uncertainty, but will take much more time. Thus, the analytical approach should be chosen for large (more than a few thousand loci) data sets.

Genotypes should be encoded as 0 (homozygote), 1 (heterozygote) and 2 (alternative homozygote). No specific polarization (e.g., minor allele, reference allele, etc.) of 0 vs 2 is required. For haploid loci, use 0 and 1. Genotype likelihoods should be on their natural scale (not phred scaled) and the values for each locus for an individual should sum to 1. The data should be provided as a list of three matrixes, with the matrixes giving the likelihoods for genotypes 0, 1 and 2 respectively. Thus, each matrix will have one row per individual and one column per locus. For haploid loci with genotype likelihoods, use the 0 and 1 matrixes to store the likelihoods of the two possible states. For all models, missing data can be encoded by setting the ploidy for an individual/locus to 0 (this indicates no information, whereas genotype likelihoods encode uncertainty in genotypes).

Ploidy data are only required for the mixed ploidy data. In this case, there should be a list of matrixes (the 1st matrix is for the hybrid so not used here), including one for each parent (2nd and 3rd matrixes, with parent 0 first). The matrixes indicate whether each locus (column) for each individual (row) is diploid (2) or haploid (1) (use 0 for missing data).

Value

A list of parameter estimates and full HMC results from stan (only when HMC is used). Parameter estimates are provided as a point estimate (median of the posterior), 90% equal-tail probability intervals (5th and 95th quantiles of the posterior distribution), and 95% equal-tail probability intervals (2.5th and 97.5th quantiles of the posterior distribution). These are provided as a vector or matrix depending on the dimensionality of the parameter. In this case, there are two matrixes, p0 and p1, for parent 0 and parent 1 allele frequencies. The full HMC output from rstan is provided as the final element in the list. This can be used for HMC diagnostics and to extract other model outputs not provided by default.

See Also

'rstan::stan' for details on HMC with stan and the rstan HMC

Examples

```
## Not run:
## load the data set
data(genotypes)
## this includes three objects, GenHybrids, GenP0, and GenP1

## estimate parental allele frequencies, uses the analytical solution to the posterior
p_out<-est_p(G0=GenP0,G1=GenP1,model="genotype",ploidy="diploid")

## End(Not run)
```

`est_Q` *Function to estimate ancestry class proportions (Q)*

Description

Uses Hamiltonian Monte Carlo (HMC) for Bayesian inference of ancestry classes proportions from genetic data. Ancestry classes denote the proportion of an individual's genome where both gene copies come from source 1 (Q11), both gene copies come from source 0 (Q00), or where one gene copy comes from source 1 and one from source 0 (Q10).

Usage

```
est_Q(
  Gx = NULL,
  G0 = NULL,
  G1 = NULL,
  p0 = NULL,
  p1 = NULL,
  model = "genotype",
  ploidy = "diploid",
  pldat = NULL,
  n_chains = 4,
  n_iters = 2000,
  p_warmup = 0.5,
  n_thin = 1,
  n_cores = NULL,
  full = TRUE
)
```

Arguments

Gx	genetic data for putative hybrids in the form of a matrix for known genotypes (rows = individuals, columns = loci), a list of matrixes for genotype likelihoods (same dimensions but one matrix per genotype), or matrix of ancestry for the ancestry model (rows = individuals, columns = loci).
G0	genetic data for parental reference set 0 formatted as described for Gx.

`G1` genetic data for parental reference set 1 formatted as described for Gx.

`p0` vector of allele frequencies for parental reference set 0 (one entry per locus).

`p1` vector allele frequencies for parental reference set 1 (one entry per locus).

`model` for genetic data, either 'genotype' for known genotypes, 'glik' for genotype likelihoods, or 'ancestry' for known ancestry.

`ploidy` species ploidy, either all 'diploid' or 'mixed' for diploid and haploid loci or individuals.

`pldat` matrix or list of matrixes of ploidy data for mixed ploidy (rows = individuals, columns = loci) indicating ploidy (2 = diploid, 1 = haploid).

`n_chains` number of HMC chains for posterior inference.

`n_iters` A positive integer specifying the number of iterations for each chain (including warmup), default is 2000.

`p_warmup` proportion (between 0 and 1) of n_iters to use as warmup (i.e., burnin), default is 0.5.

`n_thin` positive integer, save every n_thin HMC iterations for the posterior, default is 1.

`n_cores` number of cores to use for HMC, leave as NULL to automatically detect the number of cores present (no more than n_chains cores can be used even if available).

`full` boolean denoting whether (TRUE, the default) or not (FALSE) to return the HMC Stan object with the full set of samples from the posterior.

Details

Ancestry class proportions can be estimated from known genotypes (model = 'genotype'), genotype likelihoods, (model = 'glik') or known (estimated) ancestry (model = 'ancestry'). Genotypes should be encoded as 0 (homozygote), 1 (heterozygote) and 2 (alternative homozygote). No specific polarization (e.g., minor allele, reference allele, etc.) of 0 vs 2 is required. For haploid loci, use 0 and 1. Genotype likelihoods should be on their natural scale (not phred scaled) and the values for each locus for an individual should sum to 1. The data should be provided as a list of three matrixes, with the matrixes giving the likelihoods for genotypes 0, 1 and 2 respectively. Thus, each matrix will have one row per individual and one column per locus. For haploid loci with genotype likelihoods, you must use the 0 and 1 matrixes to store the likelihoods of the two possible states. For the ancestry model, hybrid indexes are inferred directly from known local (locus-specific) ancestry rather than from genotype data. Users are free to use whatever software they prefer for local ancestry inference (many exist). In this case, each entry in the individual (rows) by locus (columns) matrix should denote the number of gene copies inherited from parental population 1 (where pure parent 1 corresponds with a hybrid index of 1 and pure parent 0 corresponds with a hybrid index of 0). Haploids can be encoded using 0 and 1. For all models, missing data can be encoded by setting the ploidy for an individual/locus to 0 (this indicates no information, whereas genotype likelihoods encode uncertainty in genotypes).

Hybrid genetic (or ancestry) data are always required. For genotype or genotype likelihood models, users must either provide pre-estimated parental allele frequencies or parent genetic (genotypes or genotype likelihoods) that can be used to infer allele frequencies. Parental data are not required for the ancestry model

Ploidy data are only required for the mixed ploidy data. In this case, there should be one matrix for the hybrids or a list of matrixes for the hybrids (1st matrix) and each parent (2nd and 3rd matrixes, with parent 0 first). The latter is required for the genotype or genotype likelihood models if parental allele frequencies are not provided. The matrixes indicate whether each locus (column) for each individual (row) is diploid (2) or haploid (1) (use 0 for missing data). Haploid loci are useful for estimating the proportion of the genome inherited from each reference population and thus are relevant for this model, but do not specifically provide information about how this is partitition into heterozygous vs homozygous ancestry classes.

Value

A list of parameter estimates and full HMC results from stan, this includes Q (ancestry class proportions) and hybrid indexes, which are derived from Q. Parameter estimates are provided as a point estimate (median of the posterior) and 95% equal-tail probability intervals (2.5th and 97.5th quantiles of the posterior distribution). These are provided as a vector or matrix depending on the dimensionality of the parameter. The full HMC output from rstan is provided as the final element in the list. This can be used for HMC diagnostics and to extract other model outputs not provided by default.

References

Gompert Z, DeRaad D, Buerkle CA. A next generation of hierarchical Bayesian analyses of hybrid zones enables model-based quantification of variation in introgression in R. bioRxiv 2024.03.29.587395.

See Also

'rstan::stan' for details on HMC with stan and the rstan HMC output object.

Examples

```
## Not run:
## load the data set
data(genotypes)
## this includes three objects, GenHybrids, GenP0, and GenP1
## estimate parental allele frequencies, uses default HMC settings
p_out<-est_p(G0=GenP0,G1=GenP1,model="genotype",ploidy="diploid")

## estimate interspecific ancestry, this can
## be especially informative about the types of hybrids present
q_out<-est_Q(Gx=GenHybrids,p0=p_out$p0[,1],p1=p_out$p1[,1],model="genotype",ploidy="diploid")

## End(Not run)
```

gencline_plot *Plots genomic clines for a set of loci*

Description

Plots a set of genomic clines.

Usage

```
gencline_plot(
  center = NULL,
  v = NULL,
  pdf = TRUE,
  outf = "cline_plot.pdf",
  cvec = NULL,
  xvec = NULL,
  ...
)
```

Arguments

`center`	vector of cline centers (from est_gencline)
`v`	vector of cline gradients (from est_gencline)
`pdf`	a logical specifying whether results should be output to a pdf file; if false the plot is sent to the default graphics device.
`outf`	a character string specifying the name of the output file if 'pdf=TRUE' default = cline_plot.pdf.
`cvec`	vector of colors to use for clines (one entry per locus)
`xvec`	vector of line widths (relative to 1) to use for clines (one entry per locus)
`...`	additional arguments for plotting, see options in par and plot.

Details

This function plots genomic clines for set of loci, that is the probability of local ancestry from parental population 1 at a locus given hybrid index (the overall proportion of an individual's genome inherited from population 1). The clines for all loci are shown on a single plot, with one line per locus. A 1:1 dashed line denotes the null expected ancestry probability if all loci exhibit dynamics precisely equal to the genome-wide average intogression.

Value

A plot is produced, but there is no return value.

pp_plot *Plots bivariate posterior distribution of hierarchical cline parameters*

Description

Creates bivariate plots of combinations of cline standard deviations and means for one or multiple sets of loci.

Usage

```
pp_plot(
  objs = NULL,
  param1 = "muc",
  param2 = "sdc",
  probs = c(0.5, 0.75, 0.95),
  colors = "black",
  addPoints = TRUE,
  palpha = 0.3,
  pdf = TRUE,
  outf = "pp_plot.pdf",
  ...
)
```

Arguments

`objs` the output from est_gencline, or a list of such objects

`param1` a string specifying the genomic cline parameter (muc, muv, sc or sv) for the x-axis on the plot.

`param2` a string specifying the genomic cline parameter (muc, muv, sc or sv) for the y-axis on the plot.

`probs` vector specifying which confidence ellipses to plot (e.g., 0.95 for an ellipse that encloses 95% of the posterior).

`colors` single color or vector of colors for plotting, if more than one color is supplied different colors will be used for different posteriors (in the order specified by objs).

`addPoints` Boolean indicating whether points denoting samples from the posterior should be added to the plot (default = TRUE).

`palpha` transparency (alpha) value for the color of points, must be between 0 and 1, smaller values make points more transparent.

`pdf` a logical specifying whether results should be output to a pdf file; if false the plot is sent to the default graphics device.

`outf` a character string specifying the name of the output file if 'pdf=TRUE' default = tri_plot.pdf.

`...` additional arguments for plotting, see options in par and plot.

Details

This function produces a plot illustrating the bivariate posterior probability distribution for the cline mean or standard deviation parameters for one or more sets of genetic loci. Bivariate plots can be produced for any combination of higher-level cline parameters, that is the mean center (muc), mean slope (muv), standard deviation of centers (sc) or standard deviation of slopes (sv) (these are the names of the parameters in the HMC objects and should be used for param1 and param2). Plots include confidence ellipses capturing user specified proportions of the posterior (set with probs) and optionally points representing the samples from the posterior. Ellipses are produced by approximating the bivariate posterior with a bivariate normal distribution (these are meant to serve an illustrative purpose only). Posteriors from multiple sets of loci (multiple outputs from the est_gencline function) can be placed on the same plot. To plot a single posterior, pass the entire object (a list) produced by est_gencline to the function for the objs argument. To plot multiple, pass a list of these objects (a list of lists) to the objs function.

Users can specify colors for plotting posteriors. When more than one color is provided, different colors will be used for each object (set of loci). The transparency of the points (if included) can be adjusted with the palpha parameter. Transparency of the ellipses is set automatically by setting the alpha (transparency) value for each ellipses to 1-probs.

Value

A plot is produced, but there is no return value.

`sum2zero` *Impose hard sum-to-zero constraints on cline estimates*

Description

Re-calculates cline parameter estimates to ensure that the average cline parameter corresponds with expectations for genome-average introgression.

Usage

```
sum2zero(center = NULL, v = NULL, hmc = NULL, transform = TRUE, ci = 0.95)
```

Arguments

`center` vector or matrix of cline centers (from est_gencline). If a vector, there should be one element per locus; if a matrix, there should be one row per loucs and one column each for the point estimate and lower and upper bounds of the credible interval.

`v` vector or matrix of cline gradients (from est_gencline). If a vector, there should be one element per locus; if a matrix, there should be one row per locus and one column each for the point estimate and lower and upper bounds of the credible interval.

`hmc` HMC object from est_gencline.

`transform` Boolean variable indicating whether to apply the constraint on the natural scale of v and center (FALSE) or on the transformed scale of log(v) and logit(center). We recommend applying these transformations.

`ci` size of the credible interval, specifically, the equal-tail probability interval, to generate from the HMC object (if supplied) default = 0.95.

Details

Genomic clines model locus-specifc patterns of introgression relative to genome-average introgression or ancestry. Thus, if the same loci are used to estimate hybrid index and genomic clines (or if the former are a random sample of the latter), we would expect the average deviations from genome-average admixture to be 0 across loci (i.e., the deviations should cancel out). This is suggested by prior structure of the hierarchical Bayesian model, where the mean for log10(v) and logit(center) are set to 0. This is a soft sum-to-zero constraint (the prior pulls the sum towards zero, but a sum-to-zero constraint is not enforced). This function instead enforced a hard sum-to-zero constraint. This is done either by working with the full HMC output or parameter estimates (just point estimates or point estimates and credible intervals). In the former case, the mean cline parameters (on the log10 or logit scale, as appropriate) at each HMC iteration are forced to be 0 by subtracting off the mean. Point estimates and credible intervals are then re-calculated. In the case of parameter estimates, the mean point estimate (on the log10 or logit scale) is subtracted from the point estimate and credible intervals. Using the full HMC object is generally preferable.

As an alternative, the constraint can be applied on the natural scale rather than the log or logit scale, that is, the mean cline center can be constrained to 0.5 and the mean gradient (v) to 1. For center, the difference will often be trivial. For gradient the difference could be greater, and my current suggestion is to use the log10 scale as v is a ratio.

Value

A list with two vectors (if only a parameter estimate vector was provided) or a matrixes (if a matrix of the full HCM object was given) with the re-calculated, constrained parameter estimates. If the HMC object was given, a point estimate and the bounds of hte specified credible intervals are given.

References

Gompert Z, DeRaad D, Buerkle CA. A next generation of hierarchical Bayesian analyses of hybrid zones enables model-based quantification of variation in introgression in R. bioRxiv 2024.03.29.587395.

See Also

'rstan::stan' for details on HMC with stan and the rstan HMC output object.

`tri_plot` *Plots interpopulation ancestry (Q10) as a function of hybrid index*

Description

Creates a triangle plot of hybrid index versus interpopulation ancestry.

Usage

```
tri_plot(hi = NULL, Q10 = NULL, pdf = TRUE, outf = "tri_plot.pdf", ...)
```

Arguments

`hi` a vector of hybrid index estimates (from est_h or est_Q).

`Q10` a vector of interpopulation ancestry estimates (from est_Q).

`pdf` a logical specifying whether results should be output to a pdf file; if false the plot is sent to the default graphics device.

`outf` a character string specifying the name of the output file if 'pdf=TRUE' default = tri_plot.pdf.

`...` additional arguments for plotting, see options in par and plot.

Details

This function generates a scatterplot of interpopulation (a.k.a. interclass or interspecies) ancestry as a function of hybrid index. In other words, this shows the proportion of the genome where each putative hybrid inherited a gene copy from both parents, versus the proportion of the genome inherited from parent 1. Theoretical maxima for interpopulation ancestry given a value of hybrid index are shown as a triangle. Individuals with maximal values of interpopulation ancestry given their hybrid index have one or more non-hybrid parents, meaning they are F1s (hybrid index = 0.5 and interpopulation ancestry = 1) or backcrosses (other cases of maximal interpopulation ancestry). Of course, uncertainty in these admixture parameters can pull point estimates away from these theoretical expectations.

Value

A plot is produced, but there is no return value.

Index