Package 'polyfreqs'

December 16, 2016

Title Bayesian Population Genomics in Autopolyploids

Version 1.0.2

Description Implements a Gibbs sampling algorithm to perform Bayesian inference on biallelic SNP frequencies, genotypes and heterozygosity (observed and expected) in a population of autopolyploids. See the published paper in Molecular Ecology Resources: Blischak et al. (2016) <doi:10.1111/1755-0998.12493>.

Depends R (>= 3.0)

License GPL (>= 2)

LazyData true

Imports Rcpp

LinkingTo Rcpp

Suggests knitr, coda

VignetteBuilder knitr

URL https://github.com/pblischak/polyfreqs

BugReports https://github.com/pblischak/polyfreqs/issues

RoxygenNote 5.0.1

NeedsCompilation yes

Author Paul Blischak [aut, cre]

Maintainer Paul Blischak <blischak.4@osu.edu>

Repository CRAN

Date/Publication 2016-12-16 22:56:52

R topics documented:

`get_map_genotypes` *Maximum* a posteriori *(MAP) estmation of autopolyploid genotypes*

Description

INTERNAL: Calculates the MAP estimate of the genotypes for autopolyploid individuals using the posterior mode of the marginal posterior distribtuion of genotypes for each individual at each locus.

Usage

```
get_map_genotypes(tM, burnin = 20, geno_dir = "genotypes")
```

Arguments

`tM`	Total reads matrix: matrix containing the total number of reads mapping to each locus for each individual.
`burnin`	Percent of the posterior samples to discard as burn-in (default=20).
`geno_dir`	File path to directory containing the posterior samples of genotypes output by `polyfreqs` (default = "genotypes").

Details

The easiest way to get these estimates is to set the `genotypes` argument to `TRUE` when running `polyfreqs`.

Value

A matrix containing the maximum *a posteriori* estimates for all individuals at each locus. The MAP estimate of the genotype is simply the posterior mode.

point_Hexp *Estimation of expected heterozygosity*

Description

INTERNAL: Estimates a posterior distribution for the per locus expected heterozygosity using the unbiased estimator of Hardy (2015) and the poterior samples of allele frequencies calculated by `polyfreqs`.

Usage

```
point_Hexp(p_samp, genotypes, ploidy)
```

Arguments

`p_samp`	A posterior sample of allele frequencies from `polyfreqs`.
`genotypes`	Matrix of genotypes sampled during MCMC.
`ploidy`	The ploidy level of individuals in the population (must be >= 2).

Details

Posterior distributions for the per locus expected heterozygosity are automatically calculated and returned by the `polyfreqs` function.

Value

Returns the per locus estimates of expected heterozygosity (`per_locus_Hexp`)

References

Hardy, OJ. 2015. Population genetics of autopolyploids under a mixed mating model and the estimation of selfing rate. *Molecular Ecology Resources*, doi: 10.1111/1755-0998.12431.

point_Hobs *Estimation of observed heterozygosity*

Description

INTERNAL: Estimates a posterior distribution for the per locus observed heterozygosity using the unbiased estimator of Hardy (2015) and the poterior samples of genotypes calculated by `polyfreqs`.

Usage

```
point_Hobs(genotypes, ploidy)
```

Arguments

genotypes	A matrix of estimated genotypes returned from the function get_map_genotypes.
ploidy	The ploidy level of individuals in the population (must be >= 2).

Details

Posterior distributions for the per locus observed heterozygosity are automatically calculated and returned by the polyfreqs function.

Value

Returns per locus estimates of observed heterozygosity (per_locus_Hobs).

References

Hardy, OJ. 2015. Population genetics of autopolyploids under a mixed mating model and the estimation of selfing rate. *Molecular Ecology Resources*, doi: 10.1111/1755-0998.12431.

---

polyfreqs *Bayesian population genomics in autopolyploids*

Description

polyfreqs implements a Gibbs sampling algorithm to perform Bayesian inference on the allele frequencies (and other quantities) in a population of autopolyploids. It is the main function for conducting inference with the polyfreqs package.

Usage

```
polyfreqs(tM, rM, ploidy, iter = 1e+05, thin = 100, burnin = 20,
  print = 1000, error = 0.01, genotypes = FALSE, geno_dir = "genotypes",
  col_header = "", outfile = "polyfreqs-mcmc.out", quiet = FALSE)
```

Arguments

tM	Total reads matrix: matrix containing the total number of reads mapping to each locus for each individual.
rM	Reference reads marix: matrix containing the number of reference reads mapping to each locus for each individual.
ploidy	The ploidy level of individuals in the population (must be >= 2).
iter	The number of MCMC generations to run (default=100,000).
thin	Thins the MCMC output by sampling everything thin generations (default=100).
burnin	Percent of the posterior samples to discard as burn-in (default=20).
print	Frequency of printing the current MCMC generation to stdout (default=1000).
error	The level of sequencing error. A fixed constant (default=0.01).

genotypes	Logical variable indicating whether or not to print the values of the genotypes sampled during the MCMC (default=FALSE).
geno_dir	File path to directory containing the posterior samples of genotypes output by polyfreqs (default = "genotypes").
col_header	Optional column header tag for use in running loci in parallel (default="").
outfile	The name of the ouput file that samples from the posterior distribution of allele frequencies are written to (default="polyfreqs-mcmc.out").
quiet	Suppress the printing of the current MCMC generation to stdout (default=FALSE).

Details

Data sets run through polyfreqs must be of class "matrix" with row names representing the names of the individuals sampled. The simplest way to get data into R for running an analysis is to format the total read matrix and reference read matrix as tab delimited text files with the first column containing the individual names and one column after that with the read counts for each locus. These data can then be read in using the read.table function with the row.names argument set equal to 1. An optional tab delimited list of locus names can be included as the first row and are treated as column headers for each locus (set header=T in the read.table function). When running the polyfreqs, there are a number of options that control what the function returns. To estimate genotypes and print posterior genotype samples to file, set the genotypes argument to TRUE and select a name for the output directory geno_dir (defaults to "genotypes"). polyfreqs also prints the current MCMC generation (with a frequency set by the print_freqs argument) to the R console so that users can track run times. This print can be turned off by setting quiet=TRUE. More details on using polyfreqs can be found in the introductory vignette.

Value

Returns a list of 3 (4 if genotypes=TRUE) items:

posterior_freqs A matrix of the posterior samples of allele frequencies. These are also printed to the file with the name given by the outfile argument.

map_genotypes If genotypes=TRUE, then a fourth item will be returned as a matrix containing the maximum *a posteriori* genotype estimates accounting for burn-in.

het_obs Matrix of posterior samples of observed heterozygosity.

het_exp Matrix of posterior samples of expected heterozygosity.

Author(s)

Paul Blischak

References

Blischak PD, LS Kubatko and AD Wolfe. Accounting for genotype uncertainty in the estimation of allele frequencies in autopolyploids. *In revision*.

Examples

```
data(total_reads)
data(ref_reads)
polyfreqs(total_reads,ref_reads,4,iter=100,thin=10)
```

`polyfreqs_pps` *Posterior predictive model checks for polyfreqs*

Description

Uses the posterior distribution of allele frequences from a `polyfreqs` run to test model fit using the posterior predictive model checking procedure described in Blischak *et al*.

Usage

```
polyfreqs_pps(p_post, tM, rM, ploidy, error)
```

Arguments

`p_post`	A matrix containing the posterior samples from a `polyfreqs` run.
`tM`	Total reads matrix: matrix containing the total number of reads mapping to each locus for each individual.
`rM`	Reference reads marix: matrix containing the number of reference reads mapping to each locus for each individual.
`ploidy`	Ploidy level of individuals in the population.
`error`	The level of sequencing error. A fixed constant.

Details

The observed read count ratio (r/t) for each locus is summed across individuals and then compared to a distribution of read ratios simulated using the posterior allele frequencies by taking their difference. The criterion for passing/failing the posterior predictive check is then made on a per locus basis based on whether or not the distribution of read ratio differences contains 0 in the 95

Value

A list with two items:

ratio_diff The posterior predictive samples of the difference between the simulated read ratios and the observed read ratio summed across individuals at each locus.

locus_fit A logical vector indicating whether or not each locus passed or failed the posterior predictive model check.

References

Blischak PD, LS Kubatko and AD Wolfe. Accounting for genotype uncertainty in the estimation of allele frequencies in autopolyploids. *In revision*.

ref_reads *Reference reads matrix*

Description

A dataset of 10 individuals sampled at 2 loci with reference read counts simulated from a binomial distribution (Eq. 1 in Blischak *et al*.) with an underlying allele frequency of 0.4. Used for package testing.

Usage

```
data(ref_reads)
```

Format

A 10 x 2 matrix

References

Blischak PD, LS Kubatko and AD Wolfe. Accounting for genotype uncertainty in the estimation of allele frequencies in autopolyploids. *In revision*.

simple_freqs *Point estimation of allele frequencies based on read counts*

Description

`simple_freqs` estimates allele frequencies based on read count ratios.

Usage

```
simple_freqs(tM, rM)
```

Arguments

`tM`	Total reads matrix: matrix containing the total number of reads mapping to each locus for each individual.
`rM`	Reference reads marix: matrix containing the number of reference reads mapping to each locus for each individual.

Value

A vector of allele frequencies, one for each locus. Named `allele_freqs_hat`.

Author(s)

Paul Blischak

`sim_reads` *Simulation of sequencing read counts and genotypes*

Description

Simulates genotypes and read counts under the model of Blischak *et al*.

Usage

```
sim_reads(pVec, N_ind, coverage, ploidy, error)
```

Arguments

`pVec`	A vector of allele frequencies strung together using the concatenate function.
`N_ind`	The number of individuals to simulate.
`coverage`	The average number of sequences simulated per individual per locus (Poisson distributed).
`ploidy`	The ploidy level of individuals in the population.
`error`	The level of sequencing error. A fixed constant.

Details

Total reads are simulated using a Poisson distribution with mean equal to the `coverage` set by the user. Next, genotypes are simulated for the specified number of individuals using the vector of allele frequencies provided to the function. The number of loci simulated is equal to the number of elements supplied by the vector of allele frequencies. The number of reference reads is then simulated using Eq. 1 from Blischak *et al*. using the total reads, genotypes and sequencing error.

Value

A list of 3 matrices:

`genos` A matrix of the simulated genotypes.

`tot_read_mat` A matrix of the simulated number of total reads.

`ref_read_mat` A matrix of the simulated number of reference reads.

References

Blischak PD, Kubatko LS, Wolfe AD. 2015. Accounting for genotype uncertainty in the estimation of allele frequencies in autopolyploids. *In review*. bioRxiv, **doi**:####.

`total_reads` *Total reads matrix*

Description

A dataset of 10 individuals sampled at 2 loci with 20 reads per individual per locus. Used for package testing.

Usage

```
data(total_reads)
```

Format

A 10 x 2 matrix.

Index