Package ‘VarEff’

October 22, 2014

Type Package

Title Estimation of effective population size and variation in the time scale

Version 1.2

Date 2014-10-22

Author Natacha Nikolic and Claude Chevalet

Maintainer Natacha Nikolic <documents_57@hotmail.com> or <natachanikolic@hotmail.com>

Description Functions to estimate the effective population sizes with coalescent approach from present to ancestral time using microsatellite markers.

License GPL (>= 2)

LazyLoad yes

Depends mcmc, graphics, grDevices

URL https://qgp.jouy.inra.fr

R topics documented:

VarEff-package *Overview: Estimation of effective sizes from present to ancestral time*

Description

This package is a model called `VarEff` meaning 'Variation of Effective population size'.

It estimates the evolution of effective population size (Ne) with coalescent approach from microsatellite markers.

The estimation is done on simulated demographic histories modelled by steps of constant size for which the posterior probabilities are derived using an approximation of likelihood.

Details

Package:	VarEff
Type:	Package
Version:	1.1
Date:	2014-01-19
License:	GPL (>= 2)
LazyLoad:	yes
Depends:	mcmc
URL:	URL: https://qgp.jouy.inra.fr

Overview

This package depends on package MCMC (C.J. Geyer, 2009, version 0.9-2), so you have to load the library mcmc: library(mcmc).

To use this package and explore the results go through four steps:

1. Data preparation
2. Variable input
3. Output files
4. Explore the results

1. Data preparation

The data file describes the genotypes of a population at microsatellite markers.

The model assumes alleles defined by their lengths (number of microsatellite repeats).

The format of the file is close to MSVAR infile (Beaumont 1999). To convert a MSVAR file in `VarEff` file go to the Web site: `https://qgp.jouy.inra.fr`

Infile:

Each marker is described by 2 lines.

The first line gives the number of alleles (allelic classes) at the locus.

The second line gives the numbers of alleles at each corresponding length of the locus.

Caution:

You have to specify all potential alleles between those of minimum and maximum lengths.

It means that if you have a locus with 2 types of alleles at the lengths 10 and 12 (number of repeat motifs), you have to mention the unobserved allele with 11 motifs.

So if the alleles 10 and 12 have been observed at frequencies 24 and 6 respectively, you have to describe the locus by:

3

24 0 6

In this package the infile test is called `InputTest`.

2. Input

Because this model follows a Bayesian approach, you have to give priors on effective sizes (current and ancestral) and age of the population, specifying means and variances on the logarithmic scale.

Estimations are concerned with reduced population sizes (on the Theta = 4 * N * u scale) and reduced time (product of generation times (T) and mutation rate (u)).

The mutation rate is not estimated in this package. The package uses the mutation rate as a scale parameter to recover actual census size and actual times (generation numbers) from the results.

If you wish to get estimates of Ne in diploid numbers rather than in theta (4*Ne*u) you have to estimate previously the mutation rate (u) with existing method (Ex: MSVAR (Beaumont 1999)).

Concerning the prior on the current effective size use the function Theta().

The other parameters use in the package are to visualize the results - the time (generations) that you want to go back and the times you want to watch.

Call the package `VarEff` then answer the questions:

- parafile (Name that you give to the job and to the output files created by the model).

- infile (Name of the data file).

- NBLOC (Number of Loci).

- JMAX (Number of times when the effective size has changed, used to generate step functions simulating the past demography. Ex: JMAX=2, if you think that the population took 3 different main and global effective sizes in the past).

- MODEL (choose one mutation model - S = Single Step Model, T = Two Phase Model, or G = Geometric Model, with an additional coefficient (C) for T and G models).

- MUTAT (Mutation rate, assumed the same for all loci).

- NBAR (Global prior mean of effective size).

- VARP1 (Variance of the prior log-distribution of effective sizes. Ex: VARP1=3 allows for searches with 20- to 40-fold relative variations of effective size).

- RHOCORN (Coefficient of correlation between effective sizes in successive intervals).

- GBAR (Number of generations since the assumed origin of the population).

- VARP2 (Variance of the prior log-distribution of time intervals during which the population is assumed of constant size).

- DMAXPLUS = DMAX+1 (DMAX is the maximal distance between alleles (number of microsatellite motifs) that is used in the estimation algorithm).

- Diagonale (A smoothing parameter to balance the observed covariance structure with a theoretical diagonal variance matrix and avoid numerical instability. Diagonale = 0.5 is a robust choice).

- NumberBatch (number of batch (nbatch) for metrop in MCMC).

- LengthBatch (length of batch (blen) for metrop in MCMC).

- SpaceBatch (space of batch (nspac) for metrop in MCMC).

- Burnin (Length of the burnin period).

- AccRate(Acceptation rate).

You can also directly give the parameters into R console.

Exemple with data `InputTest`:

VarEff(infile=system.file("data/InputTest.txt", package = "VarEff"), parafile = 'job', NBLOC=20, JMAX=3, MODEL = 'S', MUTAT=0.01, NBAR=1000, VARP1=3, RHOCORN=0, GBAR=5000, VARP2=3, DMAXPLUS=12, Diagonale=0.5, NumberBatch = 1000, LengthBatch = 10, SpaceBatch = 10, Burnin=10000, AccRate=0.25)

3. Output files

At the end of the calculations, VarEff() returns the effective size estimates, the summaries of adjustment criteria of data to model, and the distribution of posterior probabilities.

The main result of VarEff() is the .Batch file, which reports a list of demographic evolutions described by step functions. Each line includes:

Column 1: the number i of the simulated state (from 1 to Numberbatch).

Column 2: quadratic deviation of data from the i-th simulated state.

Column 3: natural logarithm of the prior probability the i-th state.

Columns 4 to JMAX+4: the JMAX+1 population sizes in the i-th state.

Columns JMAX+5 to 2 JMAX+4: times of size changes in the i-th state.

Columns 2 JMAX + 5: value of the C parameter of the mutation model.

Results are kept in the .Batch files in reduced scales:

Theta's for population sizes, products of generation numbers times mutation rate for times of size changes. The additional C parameter is set to 0 for the Single Step Mutation Model, positive for geometrical model or negative for the Two Phase Model.

`NatSizeDist`

To obtain the distributions of effective size at a number of generations in the past, from the time of sampling to an ancestral time, use the function called NatSizeDist().

This function provides 2 files with the results in the Ne scale:

-job.Nstat

-job.Ndist

`LogSizeDist`

To obtain the distributions of logarithm of effective size at a number of generations in the past, from the time of sampling to an ancestral time, use the function called LogSizeDist().

This function provides 2 files with the results in the Log(Ne) scale:

-job.Lstat

-job.Ldist

Format of Nstat or Lstat file

Column 1: Time in generations (if MUTAT is not 0) or the corresponding reduced time.

Columns 2: Arithmetic Mean of Ne or Log(Ne).

Columns 3: Harmonic means of Ne (not provided for Log(Ne), set to 0 in .Lstat file).

Columns 4: Mode of Ne or Log(Ne).

Columns 5: Median of Ne or Log(Ne).

Columns 6: Quantile 5 percent of Ne or Log(Ne).

Columns 7: Quantile 95 percent of Ne or Log(Ne).

Columns 8: Standard deviation of Ne or Log(Ne) (added in Version 2).

`LogSizeDist`

To obtain the distributions of logarithm of effective size at a number of generations in the past, from the time of sampling to an ancestral time, use the function called LogSizeDist().

This function provides 2 files with the results in the Log(Ne) scale:

-job.Lstat

-job.Ldist

Format of Ndist or Ldist file:

Posterior densities of Ne or Log(Ne) at past times (fitted distribution using the density R function).

File with (Nbinstants+1) lines and 514 or 1025 columns.

Lines: Instants when the distribution of N(T(i-1)) was calculated (1<i<Nbinstants+1; 0<T(i-1)<Tempsmax).

File Ndist:

Columns in line i:

Column 1: Value of T(i-1).

Columns 2: Size of each of the intervals (=TMAX/511) in the abcsissa (Ne scale).

Columns 3 to 514: Ordinates (densities of Ne at 512 points).

File Ldist:

Columns in line i:

Column 1: Value of T(i-1).

Columns 2 to 513: Abscissa (Log(Ne) values).

Columns 514 to 1025: Ordinates (densities of these Log(Ne)).

File Nsize and Lsize:

It is a matrix with as many lines as simulated states (NumberBatch parameter) and NBT+1 columns. This file give the Nbatch series of population size values of Ne or log10(Ne), generated in the simulated states, at the required NBT+1 times.. Example: log10(N(0)), log10(N(1)) ... log10(N(NBT)).

`Tmrca`

To obtain the distributions of posterior distribution of (logarithm), the Time to the Most Recent Ancestor (TMRCA).

This function provides two files:

-A figure on the screen.

-job.Tmrca (a file recording the trials with the probabilities that TMRCA is less than the chosen upper times).

Author(s)

Natacha Nikolic <documents_57@hotmail.com> and Claude Chevalet <claude.chevalet@toulouse.inra.fr>

Maintainer Natacha Nikolic: Who to complain to <documents_57@hotmail.com>

References

Nikolic N, Chevalet C. 2014. Detecting the evolution of coalescent effective population size. Evolutionary Applications, 7(6):663-81.

Chevalet C & Nikolic N. 2010. Distribution of coalescent times and distances between microsatellite alleles with changing effective population size. Theoretical Population Biology, 77(3): 152-163.

See Also

Summary: `VarEff`
Exemple: `InputTest`
HelpData: `HelpData`
Functions to built output files: `NatSizeDist`
or `LogSizeDist`
Functions to visualize and plot the results: `plotNdistrib`
and `NTdist`
Web site: `https://qgp.jouy.inra.fr`

`HelpData` *Help: Infile to test the model*

Description

The infile called InputTest provides an overview of possible results. It is a simulation of a population that underwent a huge bottleneck 200 generations ago. The past effective population size was 1000 and the present size is 100.

Usage

```
data(InputTest)
```

Format

The format is: num [1:240] 9 4 12 2 0 14 21 15 3 9 ...

Details

The data file `InputTest.txt` is an example which can be used directly into the R console.

How to use the example

Copy into R console:

library(mcmc)

Run Theta function Example: `VarEff-package`

and

VarEff function Overview: `VarEff-package`
Example: `VarEff`

Meaning of input variables

The variables to run one of two functions (Theta or VarEff) correspond to the input arguments:

- parafile (Name that you give to the job and to the output files created by the model)

- infile (Name of the data file)

- NBLOC (Number of Loci)

- JMAX (Number of times when the effective size has changed, used to generate step functions simulating the past demography. Ex: , JMAX=2, if we think that the population took 3 different effective sizes in the past)

- MODEL (choice one mutation model in: S = Single Step Model, T = Two Phase Model, G = Geometric Model, and provide an additional coefficient (C) for T and G models)

- MUTAT (Mutation rate, assumed the same for all loci)

- NBAR (Prior value for the effective size)

- VARP1 (Variance of the prior log-distribution of effective sizes. Ex: VARP1=3 allows for searches with 20- to 40-fold relative variations of effective size)

- RHOCORN (Coefficient of correlation between effective sizes in successive intervals)

- GBAR (Number of generations since the assumed origin of the population)

- VARP2 (Variance of the prior log-distribution of time intervals during which the population is assumed of constant size)

- DMAXPLUS = DMAX+1 (DMAX is the maximal distance between alleles (number of microsatellite motifs) that is used in the estimation algorithm)

- Diagonale (A smoothing parameter to balance the observed covariance structure with a theoretical diagonal variance matrix and avoid numerical instability. Diagonale = 0.5 is a robust choice)

- NumberBatch (number of batch (nbatch) for metrop in MCMC)

- LengthBatch (length of batch (blen) for metrop in MCMC)

- SpaceBatch (space of batch (nspac) for metrop in MCMC)

- Burnin (number of preliminary simulations, to get equilibrium distributions)

- AccRate (the acceptation rate used to select valuable states)

The input example corresponds to a population genotyped with 20 microsatellite markers (NBLOC):

We supposed 3 huge events (JMAX) which have affected this population. JMAX has no impact on the actual effective size but it can affect the distribution of past effective size. Hence check different JMAX (from 1 to 10) and keep the one in which the mean, median and mode are the nearest.

Generally the mutation rate (MUTAT) is around 0.01 to 0.0001, so we gave a prior of 0.01.

Concerning the other priors, we supposed an actual effective size (NBAR) of 1000 with a large variance (VARP1=3).

We also assumed no correlation (RHOCORN=0) between the successive effective sizes from present to ancestral time.

We supposed a time since origin (GBAR) of 5000 generations with a large variance (VARP2=3).

Even if the maximum number of differences of alleles' length (number of repeat motifs) in InputTest is equal at 18, we only took 12 motifs (DMAXPLUS) because they are representative of 95 percent of data (above the red line in graph Fk).

Here, the metrop (method of Monte Carlo Markov Chain) parameters are 1000 (number of batch, NumberBatch), 10 (length of batch, LengthBatch) and 10 (space of batch, SpaceBatch) for a quick run. Burnin length is fixed at 10000 and acceptation rate at 0.25. The recommanded parameters by the function metrop are 10000 (number of batch, NumberBatch), 10 (lenght of batch, LengthBatch) and 10 (space of batch, SpaceBatch) i.e. 1 million of iterations.

From the core file (.Batch) use the function `NatSizeDist`
or `LogSizeDist`
to estimate the effective size (Ne) (or Log(Ne)) at a number of times from 0 to a certain time ago (given by the user).

To visualize the results use the functions: `NTdist`
`plotNdistrib`

Source

This data is a simulated data obtained by the model DemoDivMs from Nikolic et al. 2009. `https://qgsp.jouy.inra.fr`

References

Nikolic N., Butler J., Bagliniere JL., Laughton R., McMyn I.A.G, Chevalet C. 2009. An examination of genetic diversity and effective population size in Atlantic salmon populations, and applications for conservation and management. Genetics Research. 91: 1-18.

Examples

```
data(InputTest)
```

`InputTest` *Infile: Population simulated*

Description

This dataset corresponds to a population that underwent a huge bottleneck 200 generations ago. The past effective population size was 1000 and the present size is 100.

The genetic data of this population involve 20 microsatellite markers .

Usage

```
data(InputTest)
```

Format

The format is:

num [1:240] 9 4 12 2 0 14 21 15 3 9 ...

Details

The infile is called InputTest.txt and it provides the microsatellite data at 20 markers.

Infile format

Each markers is described by 2 lines and the unobserved alleles are mentionned by zeros:

9 (number of alleles at the first locus)

4 12 2 0 14 21 15 3 9 (counts of number of alleles with the same length)

12 (number of alleles at next locus)

.

.

.

14 (number of alleles at last locus)

1 4 0 0 0 38 18 7 9 1 0 0 1 1

It means that if you have a locus with 2 types of alleles of lenghts 10 and 12 (number of repeat motifs) you have to mention that the allele 11 is missing.

So if alleles 10 and 12 have frequencies 24 and 6 respectively, you have to describe the locus by:

3

24 0 6

This file is close to MSVAR file and you can convert a MSVAR file to VarEff file at the website `https://qgp.jouy.inra.fr`

Source

This data is a simulated data obtained by the model DemoDivMS from Nikolic et al. 2009 `https://qgsp.jouy.inra.fr`

References

Nikolic N., Butler J., Bagliniere JL., Laughton R., McMyn I.A.G, Chevalet C. 2009. An examination of genetic diversity and effective population size in Atlantic salmon populations, and applications for conservation and management. Genetics Research. 91: 1-18.

Examples

```
data(InputTest)
```

`job.Batch` *File created by the functions*

Description

The first job created by the core function VarEff is a job.Batch

Format

Core of estimation by the model VarEff

Details

This infile created by the function VarEff is used by the others functions: NatSizeDist, LogSizeDist, NTDist.

Source

See SUMMARY Manual for VAREFF package: R software from Nikolic and Chevalet. `https://qgp.jouy.inra.fr`

References

Nikolic Natacha, Chevalet Claude (2014). VarEff. Variation of Effective size. Software VAREFF (package R in file.zip) and the documentation. http://dx.doi.org/10.13155/28781. Nikolic N., Chevalet C. 2012. SUMMARY Manual for VAREFF package: R software.

`job.Ndist` *File created by the function NatSizeDist*

Description

Posterior distribution of effective size (Ne).

Format

Big file containing the detailed densities of the posterior distributions of Ne, with 512 points.

Details

This infile created by the function VarEff is used by the function: plotNdistrib.

Source

See SUMMARY Manual for VAREFF package: R software from Nikolic and Chevalet. `https://qgp.jouy.inra.fr`

References

Nikolic Natacha, Chevalet Claude (2014). VarEff. Variation of Effective size. Software VAREFF (package R in file.zip) and the documentation. http://dx.doi.org/10.13155/28781 Nikolic N., Chevalet C. 2012. SUMMARY Manual for VAREFF package: R software.

`LogSizeDist` *Reveals the estimates of Log(Ne)*

Description

Calculates estimates of effective size in Logarithmic scale (Log10(Ne)) at a number of times from 0 to a certain time ago (given by the user), plots and saves results on files.

Usage

```
LogSizeDist(NameBATCH = NULL, MUTAT = NULL, TMAX = NULL, NBT = NULL)
```

Arguments

`NameBATCH`	[matrix]: File .Batch created by the function VarEff
`MUTAT`	[numeric]: Mutation rate
`TMAX`	[integer]: Length of the period for which the distributions of Log10(Ne) in the past are generated
`NBT`	[integer]: Number of time intervals. Example: If TMAX=1000 generations, and NBT=100, the estimates are calculated every 10th generation until 1000 generations ago

Details

To obtain the distribution of effective size (Ne) in natural scale (census or Theta values) at a number of generations in the past, run the function NatSizeDist().

Both functions work the same. They use the job.Batch file created by VarEff(). Results can be shown in natural census values of population sizes, or in reduced theta scale if mutation rate is set to 0 in these functions.

Lstat file:

Column 1: Time in generations (if MUTAT is not 0) or the corresponding reduced time.

Column 2: Arithmetic Mean of Log10(Ne).

Column 3: Zer0

Column 4: Mode of Log10(Ne).

Column 5: Median of Log10(Ne).

Column 6: Quantile 5 percent of Log10(Ne).

Column 7: Quantile 95 percent of Log10(Ne).

Column 8: Standard deviation of Log10(Ne).

Ldist file:

Posterior densities of Log10(Ne) at past times (fitted distribution using the density R function).

File with (Nbinstants+1) lines and 1025 columns.

Lines: Instants when the distribution of N(T(i-1)) was calculated. 1 <= i <= NBT+1, 0 <= T(i-1) <= TMAX

- TMAX (Length of the period for which the distributions of Log(Ne) in the past are generated)

- NBT (Number of time intervals. Example: If TMAX=1000 generations, and NBT=100, estimates are calculated every 10th generation until 1000 generations ago)

Columns in line i:

Column 1: Values of T(i-1)

Columns 2 to 513: Abscissa (N values)

Columns 514 to 1025: Ordinates (densities of these N).

Lsize file:

A file giving the Nbatch series of population size values log10(N(0)), log10(N(1)) ... log10(N(NBT)), it is a matrix with as many lines as simulated states (NumberBatch parameter) and NBT+1 columns.

Value

`job.Lstat` Summary statistics of posterior effective population size in logarithm (Log(Ne)). Matrix(ncol=7,nrow=generation time): Time in generations, Arithmetic mean of Log(Ne), Mode of Log(Ne), Median of Log(Ne), Quantile 5 percent of Log(Ne), Quantile 95 percent of Log(Ne)

`job.Ldist` containing the detailed densities of the posterior distributions of Ne)

`job.Lsize` containing the values of log10(Ne) generated by the Nbatch simulations at NBT+1 times)

Note

More details on the model can be found on the website: `https://qgsp.jouy.inra.fr`

Author(s)

Natacha Nikolic <documents_57@hotmail.com> and Claude Chevalet <claude.chevalet@toulouse.inra.fr>

References

Nikolic N., Chevalet C. 2014. Detecting past changes of effective population size. Evolutionary Applications, 7(6):663-81.
Chevalet C., Nikolic N. 2010. Distribution of coalescent times and distances between microsatellite alleles with changing effective population size. Theoretical Population Biology, 77(3): 152-163.

See Also

Overview: `VarEff-package`
Exemple: `InputTest`
HelpData: `HelpData`

Examples

```
library(VarEff)
LogSizeDist (NameBATCH = system.file("extdata/job.Batch", package = "VarEff")
, MUTAT = 0.01, TMAX = 200, NBT = 10)
```

`NatSizeDist` *Reveal the estimates of Ne*

Description

Calculates estimates of effective size (Ne) at a number of times from 0 to a certain time ago (given by the user), plots and saves results on files.

Usage

```
NatSizeDist(NameBATCH = NULL, MUTAT = NULL, TMAX = NULL, NBT = NULL)
```

Arguments

`NameBATCH` [matrix]:File .Batch create by the function VarEff

`MUTAT` [numeric]:Mutation rate

`TMAX` [integer]:Length of the period for which the distributions of Ne or Log(Ne) in the past are generated

`NBT` [integer]:Number of time intervals.

Example: If TMAX=1000 generations, and NBT=100, the estimates are calculated every 10th generation until 1000 generations ago

Details

Results can be shown in natural census values of population sizes, or in reduced theta scale if mutation rate is set to 0 in these functions.

To obtain the distribution of Log10(Ne) at a number of generations in the past, run the function LogSizeDist().

Both functions work the same. They use the job.Batch file created by VarEff().

Nstat file:

Column 1: Time in generations (if MUTAT is not 0) or the corresponding reduced time.

Columns 2: Arithmetic Mean of Ne.

Columns 3: Harmonic means of Ne (not provided by LogSizeDist() for Log10(Ne)).

Columns 4: Mode of Ne.

Columns 5: Median of Ne.

Columns 6: Quantile 5 percent of Ne.

Columns 7: Quantile 95 percent of Ne.

Column 8: Standard deviation of Ne.

Ndist file:

Posterior densities of Ne or Log(Ne) at past times (fitted distribution using the density R function).

File with (Nbinstants+1) lines and 514 columns.

Lines: Instants when the distribution of N(T(i-1)) was calculated. 1 <= i <= NBT+1, 0 <= T(i-1) <= TMAX

- TMAX (Length of the period for which the distributions of Ne in the past are generated).

- NBT (Number of time intervals. Example: If TMAX=1000 generations, and NBT=100, estimates are calculated every 10th generation until 1000 generations ago).

Columns in line i:

Column 1: Values of T(i-1).

Columns 2: interval between two N abscissa of the distribution

Columns 3 to 514: Ordinates (densities of 512 Ne values).

Nsize file:

A file giving the Nbatch series of population size values N(0), N(1) ... N(NBT), it is a matrix with as many lines as simulated states (NumberBatch parameter) and NBT+1 columns.

Value

`job.Nstat` Summary statistics of posterior effective population size. Matrix(ncol=7,nrow=generation time): Time in generations, Arithmetic mean of Ne, Harmonic mean of Ne, Mode of Ne, Median of Ne, Quantile 5 percent of Ne, Quantile 95 percent of Ne.

`job.Ndist` Containing the detailed densities of the posterior distributions of Ne.

`job.Nsize` Containing the values of Ne generated by the Nbatch simulations at NBT+1 times.

Note

More details on the model can be found on the website: `https://qgsp.jouy.inra.fr`

Author(s)

Natacha Nikolic <documents_57@hotmail.com> and Claude Chevalet <claude.chevalet@toulouse.inra.fr>

References

Nikolic N., Chevalet C. 2014. Detecting past changes of effective population size. Evolutionary Applications, 7(6):663-81.

Chevalet C., Nikolic N. 2010. Distribution of coalescent times and distances between microsatellite alleles with changing effective population size. Theoretical Population Biology, 77(3): 152-163.

See Also

Overview: `VarEff-package`
Exemple: `InputTest`
HelpData: `HelpData`

Examples

```
library(VarEff)
NatSizeDist (NameBATCH = system.file("extdata/job.Batch", package = "VarEff")
, MUTAT = 0.01, TMAX = 200, NBT = 10)
```

`NTdist` *Graphical summary*

Description

Graphical summary of the posterior distribution of estimates of Log(Ne) in the past (a 2D plot).

Usage

```
NTdist(NameBATCH = NULL, MUTAT = NULL, TMAX = NULL)
```

Arguments

`NameBATCH`	[matrix]: File .Batch create by the function VarEff
`MUTAT`	[numeric]:Mutation rate
`TMAX`	[integer]:Length of the period for which the distributions of Ne or Log(Ne) in the past are generated

Details

The function NTdist() uses the job.Batch file and makes a figure to summarise the posterior distribution of Ne or Log(Ne) in a certain period (TMAX) given by the user. The graph is plotted and saved as a text.

Value

`job.2D`	Into R console: Graphical summary of the posterior distribution of estimates of Log(Ne) in the past (a 2D plot)

Note

More details on the model can be found on the website: `https://qgsp.jouy.inra.fr`

Author(s)

Natacha Nikolic <documents_57@hotmail.com> and Claude Chevalet <claude.chevalet@toulouse.inra.fr>

References

Nikolic N., Chevalet C. 2014. Detecting past changes of effective population size. Evolutionary Applications, 7(6):663-81.

Chevalet C., Nikolic N. 2010. Distribution of coalescent times and distances between microsatellite alleles with changing effective population size. Theoretical Population Biology, 77(3): 152-163.

See Also

Overview: `VarEff-package`
Exemple: `InputTest`
HelpData: `HelpData`

Examples

```
library(VarEff)
NTdist (NameBATCH = system.file("extdata/job.Batch", package = "VarEff")
, MUTAT = 0.01, TMAX = 200)
```

`plotNdistrib` *Plots the posterior distributions of Ne or Log(Ne))*

Description

Plots the posterior distributions of the estimates of Ne (or Log10(Ne)) at a number of times in the past.

Usage

```
plotNdistrib(infile = NULL, nbcases=NULL)
```

Arguments

`infile`	[matrix]:job.Ndist created by the function NatSizeDist or job.Ldist create by the function LogSizeDist
`nbcases`	[integrer]: number of times (suggestion: <=5) you wish to plot the distribution

Details

The function plotNdistrib() makes use of files .Ndist or .Ldist previously built by NatSizeDist() or LogSizeDist(). It allows the user to exhibit the density of the posterior distribution of Ne (or of Log10(Ne)) at several times in the past.

Compared to NTdist() that gives a rough but global view of these densities, plotNdistrib() gives a precise view of these densities at a small number of times. The global 2D plot given by NTdis() may help choosing the times when plotNdistrib() is used.

The function is interactive, allowing the user to check several plots.

Value

`job_NDIST`	Figure of the posterior distributions of Ne at R console
`job_LDIST`	Figure of the posterior distributions of Log10(Ne) at R console

Note

More details on the model can be found on the website: `https://qgsp.jouy.inra.fr`

Author(s)

Natacha Nikolic <documents_57@hotmail.com> and Claude Chevalet <claude.chevalet@toulouse.inra.fr>

References

Nikolic N., Chevalet C. 2014. Detecting past changes of effective population size. Evolutionary Applications, 7(6):663-81. Chevalet C., Nikolic N. 2010. Distribution of coalescent times and distances between microsatellite alleles with changing effective population size. Theoretical Population Biology, 77(3): 152-163.

See Also

Overview: `VarEff-package`
Exemple: `InputTest`
HelpData: `HelpData`

Examples

```
plotNdistrib (NameBATCH = system.file("extdata/job.Ndist", package = "VarEff")
, nbcases=0)
```

`Theta` *Preliminary analysis of data to choose Theta priors and DMAX from microsatellites data*

Description

This function allows to get global Theta (4*Ne*u) using 3 estimators correlated to the mean of Present (Theta_0), Intermediate (Theta_1) and Ancestral (Theta_2) population sizes.

The results are written to a .Thet0 file.

It is also provide Imbalance indices ln(Theta_1/Theta_0) and ln(Theta_2/Theta_0), then the minimum of census Ne (Ne0=Theta_0/4*u) and the maximum Ne (Ne_2=Theta_2/4*u).

Values of .Thet0 file: Theta_0, Theta_1, Theta_2 Imbalance indices ln(Theta_1/Theta_0) and ln(Theta_2/Theta_0)
MinNe, MaxNe
These results provide an overview of priors necessary for VarEff function.

The programm also allows one to build a file .R to process the model VarEff with the function VarEff().

Usage

```
Theta(parafile = NULL,infile  = NULL, NBLOC = NULL, JMAX = NULL,
MODEL = NULL, MUTAT = NULL, NBAR = NULL, VARP1 = NULL, RHOCORN = NULL,
GBAR = NULL, VARP2 = NULL, DMAXPLUS = NULL, Diagonale = NULL, NumberBatch = NULL
 LengthBatch = NULL, SpaceBatch = NULL, Burnin = NULL, AccRate = NULL)
```

Arguments

`parafile`	[character]:job name
`infile`	[character]:Input data
`NBLOC`	[integer]:Number of markers
`JMAX`	[integer]:Number of time that the effective size has changed
`MODEL`	[numeric]:Mutation model
`MUTAT`	[numeric]:Mutation rate
`NBAR`	[integer]:Global mean of effective size
`VARP1`	[integer]:Variance of the prior log-distribution of effective sizes
`RHOCORN`	[numeric]:Coefficient of correlation between effective sizes
`GBAR`	[integer]:Number of generations since the origin of the population
`VARP2`	[integer]:Variance of the prior log-distribution of time intervals
`Diagonale`	[numeric]:A smoothing parameter to balance the observed covariance structure with a theoretical diagonal variance
`DMAXPLUS`	[integer]:Range of allele distances to be analysed
`NumberBatch`	[integer]:number of batch (nbatch) for metrop in MCMC
`LengthBatch`	[integer]:length of batch (blen) for metrop in MCMC
`SpaceBatch`	[integer]:space between batch outputs (nspac) for metrop in MCMC
`Burnin`	[integer]:number of preliminary simulations, to get equilibrium distributions
`AccRate`	[integer]:the acceptation rate used to select valuable states

Details

This function provides:
-Global Theta0, Theta1 and Theta2 estimates, with Theta0 (mean of current Theta(4*Ne*u)), Theta1 (mean of intermediate Theta), Theta2 (mean of ancestral Theta).

-Imbalance indices ln(Theta1/Theta0) and ln(Theta2/Theta0).

-Expected range of Ne values, from the minimum and maximum global Theta estimates.
Secondly, the Theta file provides the realized acceptation rate, means, and standard deviations over simulations of the quadratic deviations of data from simulated state and of natural logarithm of the prior probabilities of the simulated states.
It also builds a script to run VarEff (job.R).

Value

`job.Thet0`	File with global preliminary statistics
`job.R`	script to run VarEff with the chosen parameters

Note

More details on the model can be found on the website: `https://qgsp.jouy.inra.fr`

Author(s)

Natacha Nikolic <documents_57@hotmail.com> and Claude Chevalet <claude.chevalet@toulouse.inra.fr>

References

Nikolic N., Chevalet C. 2014. Detecting past changes of effective population size. Evolutionary Applications, 7(6):663-81. Chevalet C., Nikolic N. 2010. Distribution of coalescent times and distances between microsatellite alleles with changing effective population size. Theoretical Population Biology, 77(3): 152-163.

See Also

Overview: `VarEff-package`
Exemple: `InputTest`
HelpData: `HelpData`

Examples

```
library(VarEff)
Theta (infile=system.file("data/InputTest.txt", package = "VarEff"),
parafile = 'job',
        NBLOC=20,
        JMAX=3,
        MODEL = 'S',
        MUTAT=0.01,
        NBAR=1000,
        VARP1=3,
        RHOCORN=0,
        GBAR=5000,
        VARP2=3,
        DMAXPLUS=12,
        Diagonale=0.5,
        NumberBatch = 100,
        LengthBatch = 10,
        SpaceBatch = 10,
        Burnin = 100,
        AccRate = 0.25)
```

Tmrca *Calculates and plots the posterior distribution of the Time to Most Recent Common Ancestor*

Description

Tmrca (Time of coalescent event) The function Tmrca() derives the posterior distribution of the Time to the Most Recent Common Ancestor of two alleles drawn from the current population. The function makes use of the job.Batch file and makes a figure to summarize the posterior distribution of TMRCA or Log(TMRCA) in a certain period (TMAX) given by the user. Like plotNdistrib(), it is an interactive function that allows the user to check several options.

Usage

```
Tmrca(batchfile=NULL,MUTAT=NULL,TMAXin=NULL,LogTMAXin=NULL)
```

Arguments

batchfile	[matrix]:File .Batch create by the function VarEff.
MUTAT	[numeric]:Mutation rate.
TMAXin	[integer]:Length of the period for which the distributions of density of the posterior distribution of time (generations) are generated in the past
LogTMAXin	[integer]:Length of the period for which the Log distributions of density of the posterior distribution of time (generations) are generated in the past

Details

The function Tmrca works with the file .Batch. It will produce a file .Tmrca and a figure on the screen. The user can produce a figure of the density of time in Log or not.

Value

job.Tmrca	Records the trials with the probabilities that TMRCA is less than the chosen upper times.

Note

More details on the model can be found on the website: https://qgp.jouy.inra.fr

Author(s)

Natacha Nikolic <documents_57@hotmail.com> and Claude Chevalet <claude.chevalet@toulouse.inra.fr>

References

Nikolic N., Chevalet C. 2014. Detecting past changes of effective population size. Evolutionary Applications, 7(6):663-81.

Chevalet C., Nikolic N. 2010. Distribution of coalescent times and distances between microsatellite alleles with changing effective population size. Theoretical Population Biology, 77(3): 152-163.

See Also

Overview: `VarEff-package`
Exemple: `InputTest`
HelpData: `HelpData`

Examples

```
Tmrca(batchfile=system.file("extdata/job.Batch", package = "VarEff"),MUTAT=0.01,TMAXin=0,
```

VarEff *Function VarEff from the package called also VarEff*

Description

`VarEff` (Variation of effective size) is a model to estimate the evolution of effective population size with coalescent approach. The estimation is done on simulated demographies modelled by steps of constant size for which the posterior probabilities are derived using an approximation of likelihood.

The user is asked the number of different steps he wishes to use to model the variation of effective size in the past.

Here, the function VarEff is given with all parameters to solve the model `VarEff`.

It produces 2 output files: summary statistics (.Theta) and the list of detailed states simulated by the MCMC chain (.Batch).

Further, files are produced by the additional functions that allows results to be visualized.

For an overview of the package, see VarEff-package.

Caution: Source the MCMC package of C.J. Geyer (2009): version 0.9-2.

Usage

```
VarEff(parafile =NULL, infile = NULL, NBLOC = NULL, JMAX = NULL,
MODEL = NULL, MUTAT = NULL, NBAR = NULL, VARP1 = NULL, RHOCORN = NULL,
GBAR = NULL, VARP2 = NULL, DMAXPLUS = NULL, Diagonale = NULL,
NumberBatch = NULL, LengthBatch = NULL, SpaceBatch = NULL, Burnin = NULL, AccRat
```

Arguments

`parafile`	[character]:Output data.
`infile`	[character]:Input data.
`NBLOC`	[integer]:Number of markers.
`JMAX`	[integer]:Number of times when the effective size has changed.
`MODEL`	[numeric]:Mutation model.
`MUTAT`	[numeric]:Mutation rate.
`NBAR`	[integer]:Global prior mean of effective size.
`VARP1`	[integer]:Logarithmic variance of effective size.
`RHOCORN`	[numeric]:Coefficient of correlation between effective sizes in successive steps.
`GBAR`	[integer]:Number of generations since the origin of the population.
`VARP2`	[integer]:Logarithmic variance of time interval.
`DMAXPLUS`	[integer]:Range of allele distances to be analysed.
`Diagonale`	[numeric]:Smoothing parameter.
`NumberBatch`	[integer]:Number of batch.
`LengthBatch`	[integer]:Length of batch.
`SpaceBatch`	[integer]:Space of batch.
`Burnin`	[integer]:Length of the burnin period.
`AccRate`	[integer]:Acceptation rate.

Details

The package named `VarEff` works with the main function called VarEff.

This function is the corpus of coalescent estimation of effective sizes from the time of sampling to ancestor population size.

The user can get the posterior distribution of effective population size at any time in the past, as asked in further functions.

Value

`job.Theta`	Summary statistics
`job.Batch`	Core of program to produce all files .Nstat, Ndist, etc.

Summary

To summarize the main part of this package go thought these five sections:

1. INFILE
2. INPUT
3. BATCH FILE
4. OUTPUTS
5. PLOT

1. INFILE

The infile is the microsatellite data of a population. It is near of MSVAR (Beaumont 1999) infile. Go to https://qgp.jouy.inra.fr to convert a MSVAR file in VarEff file.

In this package an exemple can be found in the directory data called InputTest.

2. INPUT

- parafile (Name that you give to the job and to the output files created by the model).

- infile (Name of the data file).

- NBLOC (Number of Loci).

- JMAX (Number of times when the effective size has changed, used to generate step functions simulating the past demography. Ex: JMAX=2, if you think that the population took 3 different effective sizes in the past).

- MODEL (choose one mutation model in: S = Single Step Model, T = Two Phase Model, G = Geometric Model, and provide an additional coefficient (C) for T and G models).

- MUTAT (Mutation rate, assumed the same for all loci).

- NBAR (Prior value for the effective size).

- VARP1 (Variance of the prior log-distribution of effective sizes. Ex: VARP1=3 allows for searches with 20- to 40-fold relative variations of effective size).

- RHOCORN (Coefficient of correlation between effective sizes in successive intervals).

- GBAR (Number of generations since the assumed origin of the population).

- VARP2 (Variance of the prior log-distribution of time intervals during which the population is assumed of constant size).

- DMAXPLUS = DMAX+1 (DMAX is the maximal distance between alleles (number of microsatellite motifs) that is used in the estimation algorithm).

- Diagonale (A smoothing parameter to balance the observed covariance structure with a theoretical diagonal variance matrix and avoid numerical instability. Diagonale = 0.5 is a robust choice).

- NumberBatch (number of batch (nbatch) for metrop in MCMC).

- LengthBatch (length of batch (blen) for metrop in MCMC).

- SpaceBatch (space of batch (nspac) for metrop in MCMC).

- Burnin (Length of the burnin period).

- AccRate(Acceptation rate).

3. BATCH FILE

Two files are produced by the model `VarEff`:

-job.Theta provides summary statistics (see below).

-job.Batch which is the core of the estimate.

It reports a list of demographic evolutions described by step functions. Each line includes:

Column 1: the number i of the simulated state (from 1 to Numberbatch).

Column 2: quadratic deviation of data from the i-th simulated state.

Column 3: natural logarithm of the prior probability the i-th state.

Columns 4 to JMAX+4: the JMAX+1 population sizes in the i-th state.

Columns JMAX+5 to 2 JMAX+4: the JMAX times when the size changed in the i-th state.

Columns 2 JMAX + 5: value of the C parameter of the mutation model.

Results are kept in the job.Batch file in reduced scales:

Theta's for population sizes, products of generation numbers times mutation rate for times of size changes.

The additional C parameter is a constant set to 0 for the Single Step Mutation Model, to a positive value for geometrical model or a negative value for the Two Phase Model.

4. OUTPUTS

The VarEff() function produces summary statistics in a file job.Theta.

Firstly, it produces the same as given by the Theta() function:

Global Theta0, Theta1 and Theta2 estimates.

Imbalance indices ln(Theta1/Theta0) and ln(Theta2/Theta0).

Expected range of Ne values, from the minimum and maximum global Theta estimates.

Secondly, the Theta file provides the realized acceptation rate, means, and standard deviations over simulations of the quadratic deviations of data from simulated state and of natural logarithm of the prior probabilities of the simulated states.

To obtain the distributions of effective size at a number of generations in the past, from the time of sampling to an ancestral time, use the function called NatSizeDist().

This function provides 3 files with the results in the Ne scale:

-job.Nstat

-job.Ndist

-job.Nsize

To obtain the distributions of logarithm of effective size at a number of generations in the past, from the time of sampling to an ancestral time, use the function called LogSizeDist().

This function provide 2 files with the results given in the Log(Ne) scale:

-job.Lstat

-job.Ldist

Format of Nstat or Lstat file:

Column 1: Time in generations (if MUTAT is not 0) or the corresponding reduced time.

Columns 2: Arithmetic Mean of Ne or Log(Ne).

Columns 3: Harmonic means of Ne (not provided for Log(Ne)).

Columns 4: Mode of Ne or Log(Ne).

Columns 5: Median of Ne or Log(Ne).

Columns 6: Quantile 5 percent of Ne or Log(Ne).

Columns 7: Quantile 95 percent of Ne or Log(Ne.

Columns 8 : Standard deviation of Ne or Log(Ne).

Format of Ndist or Ldist file:

Posterior densities of Ne or Log(Ne) at past times (fitted distribution using the density R function).

File with (Nbinstants+1) lines and 514 (Ndist) or 1025 (Ldist) columns.

Lines: Instants when the distribution of N(T(i-1)) was calculated.

File Ndist:

Columns in line i:

Column 1: Values of T(i-1).

Columns 2: Size of each of the intervals (=TMAX/511) in the abcsissa (Ne scale).

Columns 3 to 514: Ordinates (densities of Ne at 512 points).

File Ldist:

Columns in line i:

Column 1: Values of T(i-1).

Columns 2 to 513: Abscissa (Log(Ne) values).

Columns 514 to 1025: Ordinates (densities of these Log(Ne)).

Format of Nsize or Lsize file:

A file giving the values (N(0), N(1) ... N(NBT), (or log10(N(0)), log10(N(1)) ... log10(N(NBT)))), generated in the simulated states, at the required NBT+1 times.

It is a matrix with as many lines as simulated states (NumberBatch parameter) and NBT+1 columns.

5. PLOT

- Three functions are provided to plot posterior distributions:

1) NTdist(): Graphical summary of the posterior distribution of estimates of Log(Ne) in the past (a 2D plot), using the job.Batch file and the length of time during which distributions are retrieved.

2) plotNdistrib(): Plots the posterior distributions of the estimates of Ne (or Log(Ne)) at a number of times in the past, using a job.Ndist or job.Ldist file as previously calculated with NatSizeDist() or LogSizeDist().

3) Tmrca(): Plots the posterior distributions of time (or Log(Time)) in gerations at a number ansking by the user, using a job.Batch.

- Plot Harmonic Mean from the .Nstat file. In addition to the proposed functions, we illustrate how to plot specific results from a file produced by NatSizeDist():

```
dat2=read.table("job.Nstat")
x2=dat2[,1]
y2=dat2[,3]
x2=as.numeric(x2)
```

y2=as.numeric(y2)

maxX2=max(x2)

maxY2=max(y2)

plot(x2,y2,type='l', lwd="2",ylim=c(0,maxY2),xlab="Time T in the past (generations)",ylab="Ne(T)")

Note

More details on the model can be found on the website: `https://qgp.jouy.inra.fr` This package needs mcmc package. The example is done with few batches to test the model. The minimum is NumberBatch =10000, LengthBatch=10, SpaceBatch=10.

Author(s)

Natacha Nikolic <documents_57@hotmail.com> and Claude Chevalet <claude.chevalet@toulouse.inra.fr>

References

Nikolic N., Chevalet C. 2014. Detecting past changes of effective population size. Evolutionary Applications, 7(6):663-81.

Chevalet C., Nikolic N. 2010. Distribution of coalescent times and distances between microsatellite alleles with changing effective population size. Theoretical Population Biology, 77(3): 152-163.

See Also

Overview: `VarEff-package`
Exemple: `InputTest`
HelpData: `HelpData`

Examples

```
VarEff(infile=system.file("data/InputTest.txt", package = "VarEff"),
                          parafile = 'job',
                          NBLOC=20,
                          JMAX=3,
                          MODEL = 'S',
                          MUTAT=0.01,
                          NBAR=1000,
                          VARP1=3,
                          RHOCORN=0,
                          GBAR=5000,
                          VARP2=3,
                          DMAXPLUS=12,
                          Diagonale=0.5,
                          NumberBatch = 100,
                          LengthBatch = 10,
                          SpaceBatch = 10,
                          Burnin=10000,
                          AccRate=0.25)
```