

Package ‘Geneclust’

September 10, 2009

Title Simulation and analysis of spatial structure of population genetics data

Version 1.0.0

Author Sophie Ancelet

Description Simulation and analysis of spatial structure of population genetics data

Depends deldir, fields, spatial

Maintainer Sophie Ancelet <Sophie.Ancelet@orange.fr> , Gilles Guillot <gigu@imm.dtu.dk>

URL <http://www.sophie-ancelet.com/>

License GPL

Repository CRAN

Date/Publication 2009-09-10 18:32:59

R topics documented:

Geneclust-package	2
FormatGenotypes	3
Fst	4
geneclust	5
mcmcgeneclust	8
postclassif	14
postfis	15
postpsi	16
setplot	17
simpatch	17
simpottsidr	19
subsample	21
tablecst	22
Index	25

Geneclust-package *Bayesian clustering and MCMC inference in spatial population genetics*

Description

Performs inference of spatial structure using multilocus genotypes and spatial coordinates.

Details

Package: Geneclust
Type: Package
Version: 0.1
Date: 2006-04-11
License: GPL2

Bayesian clustering and computations of individual membership probabilities are performed using a MCMC algorithm similar to STRUCTURE (Pritchard et al, 2000) implemented in the main functions `geneclust` and `mcmcgeneclust`

In addition, the package includes the use of Hidden Markov Random Fields (HMRF) priors enabling the simultaneous analysis of spatial coordinates. So the input data include individual genotypes and spatial coordinates.

Basically the HMRF is used as a model for the spatial continuity of genotypes within a population. It contains a spatial interaction parameter ψ which represents the intensity at which individual genotypes depends from their neighbors. For $\psi = 0$, the program corresponds to another implementation of STRUCTURE. The HMRF assumes an interaction graph for the individuals. In the default implementation the graph is computed as the Dirichlet-Delaunay structure via the package `deldir`. But the program allows modifications or other implementation of graphs.

The model also assumes linkage equilibrium, but tolerates departures from the HW equilibrium using inbreeding coefficients. For $\psi > 0$, the program includes an automatic selection procedure for the actual number of clusters in the population based on Bayesian regularization.

The following functions are provided by the package:

`geneclust`: Performs Bayesian inference of membership coefficients. Users at the expert stage or `Geneland` users may prefer the function `mcmcgeneclust`.

Users at the expert stage or `Geneland` users may prefer the following functions:

`mcmcgeneclust`: Full Bayesian Markov Chain Monte Carlo inference of all parameters.

`postclassif`: Post-processing of MCMC outputs. Calculates posterior membership probabilities and cluster assignments

`postpsi`: Post-processing of MCMC outputs. Calculates posterior distribution of the spatial interaction parameter ψ .

`postfis`: Post-processing of MCMC outputs. Calculates the posterior distribution of the inbreeding coefficients

`tablecst`: Computes a normalization table needed to perform MCMC inference on the spatial interaction parameter ψ

`Fst`: Computes F-statistics.

`simpatch`: Simulates geo-referenced multilocus data from the prior distributions.

`FormatGenotypes`: Formats file of genotype data

`subsample`: Makes a subsample from an initial dataset by reducing the number of loci to consider

`setplot`: Sets graphical parameters so as to have equal ratio for horizontal and vertical axis.

Author(s)

Sophie Ancelet Gilles Guillot

References

On mixture models in population genetics:

- J.K. Pritchard, M. Stephens and P. Donnelly, Inference of population structure using multilocus genotype data, *Genetics*, pp 945-959 vol. 155, 2000

- Falush D., M. Stephens and J.K. Pritchard, Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies, *Genetics*, pp 1567-1587, vol 164, 2003

- G. Guillot, Estoup, A., Mortier, F. Cosson, J.F. A spatial statistical model for landscape genetics. *Genetics*, 170, 1261-1280, 2005.

- G. Guillot, Mortier, F., Estoup, A. Geneland : A program for landscape genetics. *Molecular Ecology Notes*, 5, 712-715, 2005.

On the implementation of MCMC inference on the spatial interaction parameter ψ of a Potts-Dirichlet model:

- P.Green, S.Richardson: Hidden Markov models and disease mapping, *Journal of the American Statistical Association* 97(460): 1055-1070

On the model (and sub-models) implemented in `geneclust`

- O.Francois, S. Ancelet, G. Guillot (2006). papers in preparation.

`FormatGenotypes` *Formatting file of genotype data.*

Description

Takes genotype data as a matrix with one line per individual and two columns per locus, with alleles coded by integers (number of replications for micro-satellites data). Build a new matrix with alleles codes as consecutive integers. If a locus has 7 alleles they will be coded as 1,2,...7. Since version 1.0.1, this function does not have to be called by users.

Usage

`FormatGenotypes` (`genotypes`)

Arguments

`genotypes` A matrix with one line per individual and two columns per locus, with alleles coded by integers

Value

A list with elements:

`genotypes` a matrix with one line per individual and two columns per locus with alleles coded by integers

`allele.numbers` a vector giving the number of possible alleles per locus

Author(s)

Gilles Guillot

Examples

```
# library(Geneclust)
```

Fst

F statistics

Description

Computes F statistics according to Weir and Cockerham's estimators.

Usage

```
Fst(genotypes, allele.numbers, pop.mbrship, npopmax)
```

Arguments

`genotypes` Genotypes of individuals. A matrix with one line per individual and 2 columns per locus

`allele.numbers` A vector of integer that contains the number of alleles at each locus

`pop.mbrship` A vector of integer that contains the posterior cluster labels (assignments)

`npopmax` A likely maximal number of clusters in the population

Value

A list with components

Total.Fit A real number estimating the total Fit
 Pairwise.Fis A matrix of real numbers estimating the pairwise Fis
 Pairwise.Fst A matrix of real numbers estimating the pairwise Fst
 Pairwise.Fit A matrix of real numbers estimating the pairwise Fit

Author(s)

Arnaud Estoup for original code in Turbo Pascal. Translation in Fortran and interface with R by Gilles Guillot

References

Weir, B.S. and C.C. Cockerham, Estimating F-statistics for the analysis of population structure, *Evolution*, 1984, vol. 38, 1358-1370.

Examples

```
## see also the example described in the function mcmcgeneclust

## Not run:
data(bear)

hc <- hgclust(bear, K=4)
plot(hc)

n.all <- c(3,3,5,6,6,6,6,7,7,7,7,7,8,8,8,8,8,10,10,10)

Fst(as.matrix(bear)[,3:40], n.all, hc$labels, npopmax=4)$Pairwise.Fst

## End(Not run)
```

geneclust

Bayesian inference of population structure using multilocus genotypes and spatial coordinates

Description

Main function of the "geneclust" package. Performs inference of population structure using spatial coordinates and multilocus genotypes.

Usage

```
geneclust(project.name = "Data", data, npopmax = 3, psi = 0.5, nit = 1000, burnin =
matngh = NULL, fis = rep(0, npopmax), varpsi = FALSE, varfis = FALSE, otherconfig=N
```

Arguments

`project.name` A path to output files directory.

`data` Object of class "geneclustdata". A dataframe that contains individuals locations and their genotypes.

`npopmax` An initial number of clusters. May be different from the final number of clusters.

`psi` A (numeric) value for the spatial interaction parameter. The `psi = 0` option corresponds to an implementation of the algorithm STRUCTURE. Typical values should be between 0 and 1.

`nit` Number of MCMC cycles. One cycle visits each locus and each individual.

`burnin` Number of cycles corresponding to the MCMC burnin period (Markov chain internal parameter).

`thinning` Number of recorded cycles (Markov chain internal parameter).

`c` A vector containing starting cluster labels for the sample. If `c = NULL`, the program starts with the uniform distribution on $(1, \dots, npopmax)$

`freq` A three dimensional array which contains the initial allele frequencies in each cluster, for each locus and each allele. If `freq = NULL`, the initial allele frequencies are randomly sampled according to the Dirichlet distribution $D(1, 1, \dots, 1)$.

`tabcst` If `varpsi = T` and `tabcst = NULL`, the program computes a partition function table (cf function `tablecst`) which may takes time. A vector with 11 components is returned and used afterwards for the inference of the Potts-Dirichlet spatial interaction parameter `psi`.

`matngh` An binary matrix which defines the neighbourhoods to be used in the Potts prior model. If `matngh[i, j]=1` then `i` and `j` are neighbours. If `matngh = NULL`, the matrix is computed from the Delaunay graph via the package `deldir`.

`fis` A vector with `npopmax` components containing the initial values for inbreeding coefficients. If `fis = NULL`, the initialization is at random.

`varpsi` Logical: if `varpsi = TRUE`, the spatial interaction parameter `psi` is treated as an unknown parameter and varies along the MCMC run. If `varpsi = FALSE`, then `psi` is kept fixed.

`varfis` Logical: if `varfis = TRUE`, the inbreeding coefficients are treated as unknown parameters and vary along the MCMC run. If `varfis = FALSE`, they are kept fixed to the initial value `fis`.

`otherconfig` A spatial configuration of individuals to compare with the posterior spatial configuration reached after the MCMC run. For example, it could be a configuration obtained with a non-Bayesian hierarchical clustering algorithm such as the Ward reconstruction method.

`write` Logical: If `TRUE`, some outputs are written in ascii files in the directory `project.name`

Details

Bayesian clustering and computations of individual membership probabilities are performed using a MCMC algorithm similar to STRUCTURE (Pritchard et al, 2000) implemented in the main functions `geneclust` and `mcmcgeneclust`

In addition, the package includes the use of Hidden Markov Random Fields (HMRF) priors enabling the simultaneous analysis of spatial coordinates. So the input data include individual genotypes and spatial coordinates.

Basically the HMRF is used as a model for the spatial continuity of genotypes within a population. It contains a spatial interaction parameter ψ which represents the intensity at which individual genotypes depends from their neighbors. For $\psi = 0$, the program corresponds to another implementation of STRUCTURE. The HMRF assumes an interaction graph for the individuals. In the default implementation the graph is computed as the Dirichlet-Delaunay structure via the package `deldir`. But the program allows modifications or other implementation of graphs.

The model also assumes linkage equilibrium, but tolerates departures from the HW equilibrium using inbreeding coefficients. For $\psi > 0$, the program includes an automatic selection procedure for the actual number of clusters in the population based on Bayesian regularization.

Basically this function does the same as `mcmcgeneclust` but it proposes a simplified version, and an easier access to data summaries. Users of the `geneland` package or users at the expert stage may prefer the function `mcmcgeneclust` which offers more parameters.

Value

An object of class `geneclust`.

<code>prob</code>	A matrix with indicates the posterior distributions of membership coefficients for each individual
<code>membership</code>	A vector containing the most likely cluster membership
<code>K</code>	Estimated number of clusters (less than the initial number)
<code>postmodepsi</code>	The posterior mode of ψ
<code>postmeanfis</code>	A numerical vector which contains the posterior mean of inbreeding coefficient in each identified cluster
<code>postquantfis</code>	A matrix that stores the posterior distribution quantiles of each inbreeding coefficient. Each line corresponds to one cluster
<code>diffclassif</code>	A rate of misclassification computed if another spatial configuration is given as argument
<code>coord</code>	Individual spatial coordinates
<code>psi</code>	Spatial interaction parameter after <code>nit</code> MCMC cycles
<code>fis</code>	Inbreeding coefficients after <code>nit</code> cycles
<code>path</code>	Path to the MCMC program output data
<code>c</code>	Cluster configuration after <code>nit</code> MCMC cycles
<code>freq</code>	Allele frequencies after <code>nit</code> MCMC cycles
<code>matngh</code>	Neighbourhood matrix
<code>tabcst</code>	Partition function table

Author(s)

Sophie Ancelet

References

On mixture models in population genetics:

- J.K. Pritchard, M. Stephens and P. Donnelly, Inference of population structure using multilocus genotype data, *Genetics*, pp 945-959 vol. 155, 2000

- Falush D., M. Stephens and J.K. Pritchard, Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies, *Genetics*, pp 1567-1587, vol 164, 2003

- G. Guillot, Estoup, A., Mortier, F. Cosson, J.F. A spatial statistical model for landscape genetics. *Genetics*, 170, 1261-1280, 2005.

- G. Guillot, Mortier, F., Estoup, A. Geneland : A program for landscape genetics. *Molecular Ecology Notes*, 5, 712-715, 2005.

On the implementation of MCMC inference on the spatial interaction parameter ψ of a Potts-Dirichlet model:

- P.Green, S.Richardson: Hidden Markov models and disease mapping, *Journal of the American Statistical Association* 97(460): 1055-1070

On the model (and sub-models) implemented in `geneclust`

- O.Francois, S. Ancelet, G. Guillot (2006). papers in preparation.

See Also

Functions `mcmcgeneclust`, `postclassif`, `postpsi`, `postfis`

mcmcgeneclust

MCMC inference of population structure using multilocus genotypes

Description

Runs a Markov Chain Monte-Carlo for Bayesian clustering of multilocus genotypes using a hidden spatial model.

Usage

```
mcmcgeneclust(path.mcmc, genotypes, coordinates, npopmax, nit = 3000, burnin = 200,
```

Arguments

<code>path.mcmc</code>	Path to output files directory. Path should be given in the Unix style even under Windows (use / instead of \). The path must be ended by slash (/) (e.g. <code>path.mcmc="/home/me/Genetics/"</code>)
<code>genotypes</code>	(Codominant) genotypes of individuals. A matrix with one line per individual and 2 columns per locus
<code>coordinates</code>	Spatial coordinates of individuals. A matrix with 2 columns and one line per individual.
<code>npopmax</code>	An initial number of clusters. Maybe different from the final number of clusters
<code>nit</code>	Number of MCMC cycles. One cycle visits every locus and every individuals.
<code>burnin</code>	Number of cycles to throw away. Results are stored in ascii files from <code>burnin</code> and only each <code>thinning</code> cycles.
<code>thinning</code>	Number of cycles between two writings
<code>c</code>	A vector with <code>nindiv</code> components containing the initial population memberships. If <code>c = NULL</code> , the program is started according to the uniform distribution on $(1, \dots, npop)$
<code>psi</code>	A value for the spatial interaction parameter (numeric) to be used when <code>varpsi = F</code> . The <code>psi = 0</code> and <code>varpsi = F</code> options correspond to an implementation of the algorithm STRUCTURE.
<code>fis</code>	A vector with <code>npopmax</code> components containing the initial values for each population inbreeding coefficient. If <code>fis = NULL</code> , initialization will be made according to the Beta distribution with parameters <code>alpha</code> and <code>beta</code>
<code>freq</code>	A three dimensional array (<code>npopmax*number of loci*maximum number of alleles</code>) which contains the initial allele frequencies for each population, each locus and each allele. If <code>freq=NULL</code> , the initial allele frequencies are distributed according to a non-informative Dirichlet distribution $D(1, 1, \dots, 1)$.
<code>tabcst</code>	A vector with 11 components used in the inference of the Potts-Dirichlet spatial interaction parameter (<code>psi</code>). If <code>tabcst=NULL</code> , the program computes the normalization constant table (cf function <code>tablecst</code>).
<code>matngh</code>	A matrix with one line per individual and one column per individual defining the neighbourhood relationships between individuals. <code>matngh[i,j]=1</code> if <code>i</code> and <code>j</code> are neighbouring individuals. If <code>matngh=NULL</code> , the neighbourhood matrix is defined by a Delaunay graph via the function <code>deldir</code> contained in library <code>deldir</code> .
<code>flat.prior.psi</code>	If <code>flat.prior.psi==TRUE</code> , the prior on the spatial interaction parameter <code>psi</code> is a uniform distribution over $(0, 0.1, 0.2, \dots, 1)$. If <code>flat.prior.psi==FALSE</code> , the prior is a beta distribution (discretized) with <code>mean=0.6</code> and <code>sd=0.084</code>
<code>stepval</code>	Step of discretization of the interval $[0, psimax]$ defining the values for which $E(U(c) psi)$ is computed.
<code>nit.table</code>	Number of MCMC iterations to generate Potts-Dirichlet configurations in order to compute $E(U(c) psi)$ where <code>U(c)</code> is the Potts-Dirichlet configuration energy
<code>burnin.table</code>	Number of MCMC updates to throw away to compute $E(U(c) psi)$
<code>stepw.table</code>	Number of MCMC updates between two stored updates to compute $E(U(c) psi)$

alpha	The shape1 parameter of the beta prior on the inbreeding coefficients. alpha must be positive.
beta	The shape2 parameter of the beta prior on the inbreeding coefficients. beta must be positive
varpsi	Logical: if varpsi=TRUE the spatial interaction parameter psi is treated as unknown and will vary along the MCMC inference. If varpsi=FALSE it will be fixed to the initial value psi.
varfis	Logical: if varfis=TRUE the inbreeding coefficients are treated as unknown and will vary along the MCMC inference. If varfis=FALSE they will be fixed to the initial value fis.

Details

For those who want to handle the all obscure parameters of an MCMC run. This is the core function used by `geneclust`. Its syntax is very similar to the one used by the package `geneland`. So MCMC experts and `geneland` users may prefer this function to `geneclust`.

Value

Successive states of all blocks of parameters are written in external files contained in `path.mcmc` and named after the type of parameters that they contain.

c	Cluster configuration of individuals reached after nit MCMC cycles
psi	Spatial interaction parameter psi value reached after nit MCMC cycles
fis	Inbreeding coefficients values reached after nit MCMC cycles
freq	Allele frequencies values reached after nit MCMC cycles
matngh	Neighbourhood matrix
table	Normalization constant table

Storage format

All parameters processed by function `mcmcgeneclust` are written in the directory specified by 'path.mcmc' as follows:

File 'population.numbers.txt' contains values of the estimated number of clusters (nit lines, one line per cycles of the MCMC algorithm)

File 'frequencies.txt' contains allele frequencies of current populations. Column xx contains frequencies of the population labelled xx. In each column, the values of the allele frequencies are stored by increasing allele index and locus index (allele index varying first), and values for successive cycles are pasted. The file has $n_{\text{allmax}} * n_{\text{loc}} * \text{nit} / \text{thinning}$ lines where n_{allmax} is the maximum number of alleles over all loci.

File 'psi.txt' contains values of the spatial interaction parameter psi (nit lines, one line per cycle of the MCMC algorithm)

File 'fis.txt' contains values of the inbreeding coefficients (nit lines, one line per cycle of the MCMC algorithm)

File 'parameters.txt' contains a summary of the main inference parameters

Author(s)

Sophie Ancelet

See AlsoFunction [geneclust](#), [tablecst](#), [postclassif](#), [postpsi](#), [postfis](#)**Examples**

```
# Below is a sequence of R commands using geneclust functions
# The commands are in the same format as the 'Geneland' package by G. Guillot

# library(Geneclust)

## Not run:

# Simulation of a dataset according to the prior distributions. This one
# is made of 2 populations
# 10 loci and 10 alleles per loci
# on a spatial domain enclosed in a rectangle
# (x.coord. in [0,1], y.coord. in [0,1])
# Spatial interaction parameter is 0.4
# We suppose the inbreeding coefficients are the same in each population
# and equal to 0.1

#To define a place for simulation outputs
path <- "./tmpData/"
system("mkdir ./tmpData/")

data<- simpatch (   path=path,
                   nindiv=100,
                   coordinates=NULL,
                   coord.lim=c(0,1,0,1),
                   npop=2,
                   nall=rep(10,10),
                   psi=0.4,
                   fis=c(0.1,0.1),
                   nchain=50000,
                   burnin=40000,
                   stepw=100,
                   seed=123,
                   plot=FALSE,
                   write=FALSE,
                   print=FALSE,
                   file=path)

## go to file path to read simulation outputs

#To run Full Bayesian Markov Chain Monte Carlo inference of all
#parameters
#To define a place for MCMC outputs
```

```
path.mcmc<- path

infer    <- mcmcgeneclust(
  path.mcmc=path.mcmc,
  genotypes=data$genotypes,
  coordinates=data$coordinates,
  npopmax=3,
  nit=3000,
  burnin=200,
  thinning=5,
  c=NULL,
  psi=0,
  fis=rep(0,times=3),
  freq=NULL,
  tabcst=NULL,
  matngh=NULL,
  flat.prior.psi=TRUE,
  stepval=0.02,
  nit.table=20000,
  burnin.table=10000,
  stepw.table=10,
  alpha=4,
  beta=40,
  varpsi=TRUE,
  varfis=TRUE)

# To make a second run of MCMC algorithm from the last MCMC
#configuration:

infer2    <- mcmcgeneclust(
  path.mcmc=path.mcmc,
  genotypes=data$genotypes,
  coordinates=data$coordinates,
  npopmax=3,
  nit=3000,
  burnin=0,
  thinning=5,
  c=infer$c,
  psi=infer$psi,
  fis=infer$fis,
  freq=infer$freq,
  tabcst=infer$table,
  matngh=infer$matngh,
  flat.prior.psi=TRUE,
  stepval=0.02,
  nit.table=20000,
  burnin.table=10000,
  stepw.table=10,
  alpha=4,
  beta=40,
  varpsi=TRUE,
```

```

varfis=TRUE)

#Go to file path.mcmc to read MCMC outputs

#To computes for posterior probabilities and posterior
#modal populations

classif<- postclassif(path.mcmc=path.mcmc,
                      coordinates=data$coordinates,
                      popmbrship=data$popmbrship,
                      plot=TRUE,
                      print=TRUE,
                      file=path.mcmc,
                      write=TRUE)

#Number of inferred populations
classif$K.est

#Rate of misclassifications
classif$errclassif

#To get the main properties of the posterior distribution of the spatial interaction parameter
interparam<- postpsi(path.mcmc=path.mcmc,
                     plot=TRUE,
                     print=TRUE,
                     file=path.mcmc)

#Postmode psi
interparam$postmode.psi

#To get the main properties of the inbreeding coefficients posterior
#distributions

inbreedcoeff<- postfis(path.mcmc=path.mcmc,
                      postmode.indiv= classif$postmode.indiv,
                      probs = c(0.025, 0.25, 0.5, 0.75, 0.975),
                      plot = TRUE,
                      print = TRUE,
                      file=path.mcmc)

#Postmean fis
inbreedcoeff$postmean.fis

#Credibility intervals
inbreedcoeff$quant.fis

#To computes indices which quantify the level of genetic differentiation between inferred populations
#estimators

differindex<- Fst(genotypes=data$genotypes,
                  allele.numbers=rep(10,10),
                  pop.mbrship=classif$postmode.indiv,
                  npopmax=3)

```

```
#Weir and Cockerham's estimator of Fst between inferred populations
differindex$Pairwise.Fst

## End(Not run)
```

postclassif *Posterior membership probabilities and cluster assignment*

Description

Post-processing of MCMC outputs computing individual membership coefficients and the most likely cluster assignment.

Usage

```
postclassif(path.mcmc, coordinates, popmbrship=NULL, plot=TRUE, print=FALSE, file=p
```

Arguments

path.mcmc	Path to the MCMC program output files directory
coordinates	Spatial coordinates of individuals. A matrix with 2 columns and one line per individual
popmbrship	A vector with number of individuals components containing the true population memberships (if they are known!). If popmbrship=NULL, the population of origin of each individuals is unknown.
plot	Logical: if TRUE, maps are plotted
print	Logical: if TRUE, maps are also printed.
file	Character : Path to file where maps should be printed
write	Logical: If TRUE, data are written in ascci files named "postmode.indiv.txt" and "proplab.txt" respectively

Value

A list with components:

postmode.indiv	A vector that contains the cluster assignment of each individual
effpop.est	A vector which contains the number of individuals in each population defined
proplab	Matrix of posterior membership probabilities for each individuals
K.est	The number of populations defined from the cluster assignment of each individual
errclassif	A rate of misclassification computed if the true population memberships are given as argument
K.distrib	The posterior distribution of K (number of populations)
modeK	The mode of the posterior distribution of K (number of populations)

Author(s)

Sophie Ancelet

Examples

```
# library(Geneclust)

## see the example described in the function mcmcgeneclust
```

postfis *Posterior distribution of inbreeding coefficients*

Description

Computes the posterior mean and some posterior quantiles for inbreeding coefficients.

Usage

```
postfis(path.mcmc, postmode.indiv, probs = c(0.025, 0.25, 0.5, 0.75, 0.975), plot =
```

Arguments

path.mcmc	Path to MCMC program output files
postmode.indiv	A vector which contains the population of origin of each individual
probs	Numeric vector of probabilities with values in [0,1]. By default,probs=c(0.025,0.25,0.50,0.75,0.975) in order to compute credibility intervals.
plot	Logical: if TRUE histograms are plotted
print	Logical: if TRUE histograms are also printed
file	Character: Path to file where figures should be printed

Value

postmean.fis	A numerical vector which contains the posterior mean of inbreeding coefficient in each identified cluster
quant.fis	A matrix that stores the posterior distribution quantiles of each inbreeding coefficient. Each line corresponds to one cluster

Author(s)

Sophie Ancelet

Examples

```
# library(Geneclust)

## see the example described in the function mcmcgeneclust
```

postpsi *Posterior distribution of the spatial interaction parameter*

Description

Posterior mode of psi from the MCMC outputs

Usage

```
postpsi(path.mcmc, plot=TRUE, print=FALSE, file=path.mcmc)
```

Arguments

path.mcmc	Path to the MCMC program output files
plot	Logical: if TRUE an histogram is plotted.
print	Logical: if TRUE the histogram is also printed.
file	Character: Path to file where the histogram should be printed

Value

Postmode.psi The posterior mode of psi

Author(s)

Sophie Ancelet

Examples

```
# library(Geneclust)

## see the example described in the function mcmcgeneclust
```

setplot	<i>Set graphical parameters for geographical maps so as to have equal ratio for horizontal and vertical axis. Internal function.</i>
---------	--

Description

Set graphical parameters so as to have equal ratio for horizontal and vertical axis.

Usage

```
setplot(xdata, ydata, pretty.call = TRUE, maxdim, axes = FALSE)
```

Arguments

xdata	A vector of x coordinates
ydata	A vector of y coordinates
pretty.call	~~Describe pretty.call here~~
maxdim	~~Describe maxdim here~~
axes	~~Describe axes here~~

Value

Reset graphical parameters

Author(s)

Gilles Guillot

Examples

```
#library(Geneclust)
```

simpatch	<i>Simulate geo-referenced multilocus data sets</i>
----------	---

Description

Simulates multilocus genotypes and group memberships according to the prior distributions used in the Bayesian algorithm.

Usage

```
simpatch(path, nindiv, coordinates = NULL, coord.lim = c(0,1,0,1),
npop, nall, psi, fis, nchain = 50000, burnin=40000, stepw = 100, seed = NULL, plot
= TRUE, write=FALSE, print=FALSE, file=path)
```

Arguments

<code>path</code>	path to the MCMC program output files
<code>nindiv</code>	Number of individuals
<code>coordinates</code>	Spatial coordinates of individuals. A matrix with 2 columns and one line per individual
<code>coord.lim</code>	Vector: Ranges of the spatial domain to be considered (xmin,xmax,ymin,ymax)
<code>npop</code>	Number of populations
<code>nall</code>	Vector of integers giving the number of alleles at each locus
<code>psi</code>	A nonnegative spatial interaction parameter. If $\psi=0$, populations are not spatially structured. Typical values are between 0 and 1.
<code>fis</code>	Vector of population inbreeding coefficients. If $fis[i]=0$, there is no inbreeding in population i
<code>nchain</code>	Number of MCMC iterations (Gibbs steps) to generate a Potts-Dirichlet configuration
<code>burnin</code>	Number of Gibbs steps to throw away. Results are stored in ascii files from <code>burnin</code> and only each <code>thinning</code> cycles.
<code>stepw</code>	Number of MCMC iterations between two writing steps (if <code>stepw=1</code> , all states are saved whereas if e.g. <code>stepw=10</code> only each 10 iterations is saved)
<code>seed</code>	Logical: Seed to initialize the random number generator
<code>plot</code>	Logical: if TRUE, the map giving the population membership of each individual and the barplots for allele frequencies are plotted
<code>write</code>	Logical: if TRUE, coordinates, allele frequencies, genotypes, population memberships, <code>matngh</code> , number of alleles and other variables involved in the simulation are also written in plain ascci files
<code>print</code>	Logical: if TRUE the map and the barplots for allele frequencies are also printed
<code>file</code>	Character: Path to file where figures should be printed and/or variables involved in the simulation should be written

Value

A list of variables involved in the simulation. The elements of this list are: `coordinates`,`matngh`,`popmbrship`,`genotypes`,`frequencies`

Storage format

All parameters processed by function `simpatch` are written in the directory specified by `'path'` as follows:

File `'simpotts.txt'` contains MCMC updates of each individual population membership

File `'energy.txt'` contains the Potts system energy for each MCMC update of the population memberships

Author(s)

Sophie Ancelet

Examples

```

# library(Geneclust)

## Not run:

# Simulation of a dataset made of 2 populations
# 10 loci and 10 alleles per loci
# on a spatial domain enclosed in a rectangle
# (x.coord. in [0,1], y.coord. in [0,1])
# Spatial interaction parameter is 0.5
# We suppose the inbreeding coefficients are the same in each population
# that is to say 0.1

# define a place for simulation outputs
system("mkdir ./tmpData/")
path <- "./tmpData/"

data<- simpatch(path=path,
                nindiv=100,
                coordinates=NULL,
                coord.lim=c(0,1,0,1),
                npop=2,
                nall=rep(10,10),
                psi=0.5,
                fis=c(0.1,0.1),
                nchain=50000,
                burnin=40000,
                stepw=100,
                seed=123,
                plot=TRUE,
                write=FALSE,
                print=FALSE,
                file=path)

## go to file path to read simulation outputs
## End(Not run)

```

simpottsdir

Simulation of data from the Potts-Dirichlet model on an irregular lattice

Description

Simulates some spatially organised populations according to the Potts-Dirichlet model. Plots a map giving each individual population membership.

Usage

```

simpottsdir(path, coordinates, matngh=NULL, npop, psi, nchain = 50000,
burnin=40000, stepw = 10, plot=TRUE, ploth=FALSE, print=FALSE, file=path)

```

Arguments

<code>path</code>	path to output files directory
<code>coordinates</code>	Spatial coordinates of individuals. A matrix with 2 columns and one line per individual
<code>matngh</code>	A neighbourhood matrix with <code>nindiv</code> lines and <code>nindiv</code> columns. If <code>matngh=NULL</code> , neighbourhood relationships are defined by a graph of Delaunay.
<code>npop</code>	Number of populations
<code>psi</code>	The real value of the hidden markov random field spatial interaction parameter. If <code>psi=0</code> , populations aren't spatially structured
<code>nchain</code>	Number of MCMC iterations (Gibbs steps) to generate a Potts-Dirichlet configuration
<code>burnin</code>	Number of Gibbs steps to throw away. Results are stored in ascii files from <code>burnin</code> and only each <code>thinning</code> cycles.
<code>stepw</code>	Number of MCMC iterations between two writing steps (if <code>stepw=1</code> , all states are saved whereas if e.g. <code>stepw=10</code> only each 10 iterations is saved)
<code>plot</code>	if <code>plot=TRUE</code> , the map giving the population membership of each individual is plotted
<code>ploth</code>	if <code>ploth=TRUE</code> , other charts concerning the MCMC simulation are plotted
<code>print</code>	Logical: if <code>print=TRUE</code> the map is also printed.
<code>file</code>	Character: Path to file where the map should be printed

Value

A list of variables involved in the simulation. The elements of this list are: `matngh`, `popmbrship`.

Storage format

All parameters processed by function `simpottsdir` are written in the directory specified by 'path' as follows:

File 'simpotts.txt' contains MCMC updates of each individual population membership

File 'energy.txt' contains the Potts-Dirichlet system energy for each MCMC update of the population memberships.

Author(s)

Sophie Ancelet

Examples

```
#library(Geneclust)

## Not run:

# Below is a sequence of R commands using geneclust functions
```

```

# Simulation of a dataset according to Potts-Dirichlet model
# The dataset is made of 2 populations
# on a spatial domain enclosed in a rectangle
# (x.coord. in [0,1], y.coord. in [0,1])
# Spatial interaction parameter is 0.4

#To define a place for simulation outputs
system("mkdir ./tmpData/")
path <- "./tmpData/"

#To generate the coordinates of 100 individuals which are supposed
# uniformly distributed in a rectangle

coordinates<- matrix(runif(200,0,1),nrow=100,ncol=2)

data<- simpottsdir(path=path,
                  coordinates=coordinates,
                  matngh=NULL,
                  npop=2,
                  psi=0.4,
                  nchain=50000,
                  burnin=40000,
                  stepw=100,
                  plot=TRUE,
                  ploth=FALSE,
                  print=FALSE,
                  file=path)

## go to file path to see simulation outputs
## End(Not run)

```

subsample

Make a subsample from an initial dataset

Description

This function enables to make a subsample from an initial dataset by reducing the number of loci to consider.

Usage

```
subsample(data, nloci, lst = NULL)
```

Arguments

data	An object of class <code>geneclustdata</code>
nloci	Number of loci to consider to make the new dataset
lst	The indices of loci chosen to make the new dataset. If <code>lst=NULL</code> , these indices are chosen randomly

Value

`new.dat` The new dataset which is again an object of class `geneclustdata`. Genetics information are reduced to "nloci" number of loci.

Author(s)

Olivier Francois

Examples

```
#library(Geneclust)

## Not run:

# Simulation of a dataset according to the prior distributions. This one
# is made of 2 populations
# 10 loci and 10 alleles per loci
# on a spatial domain enclosed in a rectangle
# (x.coord. in [0,1], y.coord. in [0,1])
# Spatial interaction parameter is 0.5
# We suppose the inbreeding coefficients are the same in each population
# and equal to 0.1

#define a place for simulation outputs
system("mkdir ./tmpData/")
path <- "./tmpData/"

data<-simpatch(path=path,nindiv=100,npop=2,nall=rep(10,10),psi=0.5,fis=c(0.1,0.1),nchain=200)

datagc<- as.geneclustdata(data$coordinates[,1],data$coordinates[,2],data$genotypes)

#Subsample with 5 randomly chosen loci
geneclustobjsub<- subsample(data=datagc,nloci=5,lst=NULL)
## End(Not run)
```

tablecst

A normalization constants table to make MCMC inference on the Potts-Dirichlet model spatial interaction parameter

Description

Computes the Potts-Dirichlet model normalization constants table by the method proposed by Sylvia Richardson and Peter J.Green in the article: "Hidden Markov Models and Disease Mapping"(JASA Dec 2002).

Usage

```
tablecst(pathtable, npopmax, coordinates, matngh,
stepval = 0.02, nit.table = 20000, stepw.table = 10, burnin.table =
10000, plot=TRUE, write = FALSE)
```

Arguments

pathtable	Path to output file directory
npopmax	Initial number of populations
coordinates	Spatial coordinates of individuals. A matrix with 2 columns and one line per individual
matngh	The neighbourhood matrix with nindiv lines and nindiv columns. If matngh[i,j]=1 then the individuals i and j are neighbours.
stepval	Step of discretization of the interval [0,1]. The expected Potts system energy will be computed for each psi value of this discretized interval by MCMC simulations of Potts-Dirichlet configurations. By default, stepval=0.02.
nit.table	Number of MCMC iterations to generate Potts-Dirichlet configurations.
stepw.table	Number of MCMC iterations between two writing steps (if stepw.table=1, all states are saved whereas if e.g. stepw.table=10 only each 10 iterations is saved)
burnin.table	Number of MCMC iterations to throw away to compute the expected Potts-Dirichlet system energy for each value of psi
plot	Logical: if plot=TRUE the Potts-Dirichlet model normalization constants are plotted for each value of psi (0,0.1,0.2,...,1)
write	Logical: if write=TRUE the table is written in a plain ascii file named table.txt

Value

A numerical vector (the table) with $(10*1)+1$ components. Each component is an approximation of the Potts-Dirichlet model normalization constant (at log scale) for each value of psi. We supposed that psi takes its values between 0 and 1 with a discretization step of 0.1.

Storage format

All parameters processed by function `tablecst` are written in the directory specified by 'pathtable' as follows:

File 'table.txt' contains a numerical vector with 11 components. Each component is an approximation of the Potts-Dirichlet model normalization constant (at log scale) for each value of psi. We supposed that psi takes its values between 0 and 1 with a discretization step of 0.1.

$(1/\text{stepval})+1$ repertories. Each one contains the outputs of a Gibbs sampler to generate data according to Potts-Dirichlet model for psi from 0 to 1 with a discretization step of 0.02.

Author(s)

Sophie Ancelet

References

Sylvia Richardson, Peter J.Green: "Hidden Markov Models and Disease Mapping"(JASA December 2002)

See Also

Function [simpottsdir](#)

Examples

```
#library(Geneclust)
# Below is a sequence of R commands using Geneclust functions to compute
# the Potts-Dirichlet model normalization constants table when we consider 100 individuals
# organized in 2 populations.

## Not run:

#To define a place for outputs
system("mkdir ./tmpData/")
pathtable <- "./tmpTable/"

#To generate the coordinates of 100 individuals which are supposed
#uniformly distributed in a rectangle
coordinates<- matrix(runif(200,0,1),nrow=100,ncol=2)

#To compute the neighbourhood matrix
del<- deldir(x=coordinates[,1],y=coordinates[,2])
colngh<- del$delsgs[,5:6]
pt<- nrow(colngh)

matngh<- matrix(0,nrow=100, ncol=100)
for(i in 1:pt){
  matngh[colngh[i,1],colngh[i,2]]=1
  matngh[colngh[i,2],colngh[i,1]]=1
}

table<- tablecst (pathtable=pathtable,
                 npopmax=2,
                 coordinates=coordinates,
                 matngh=matngh,
                 stepval=0.02,
                 nit.table=20000,
                 stepw.table=10,
                 burnin.table=10000,
                 plot=TRUE,
                 write=TRUE)

## go to file pathtable to read outputs
## End(Not run)
```

Index

FormatGenotypes, [3](#)
Fst, [2](#), [4](#)

Geneclust (*Geneclust-package*), [1](#)
geneclust, [2](#), [5](#), [6](#), [9](#), [10](#)
Geneclust-package, [1](#)

mcmcgeneclust, [2](#), [6](#), [8](#), [8](#), [10](#)

postclassif, [2](#), [8](#), [10](#), [13](#)
postfis, [2](#), [8](#), [10](#), [14](#)
postpsi, [2](#), [8](#), [10](#), [15](#)

setplot, [16](#)
simpatch, [2](#), [17](#), [18](#)
simpottsdir, [20](#), [23](#)
simpottsdir (*simpottsidr*), [19](#)
simpottsidr, [19](#)
subsample, [21](#)

tablecst, [2](#), [10](#), [22](#), [23](#)