Geneland: A computer package for landscape genetics

Gilles Guillot *, Frédéric Mortier †, Arnaud Estoup ‡

February 28, 2005

*Unité de Mathématiques et Informatique Appliqués INRA-INAPG-ENGREF
†CIRAD Département Forêts
‡Centre de Biologie et de Gestion des Populations, INRA
Running Head: A program for landscape genetics

Key Words: landscape genetics, simulation, MCMC inference, software

Corresponding Author:

Gilles Guillot

Unité de Mathématiques et Informatiques Appliquées
INRA-INAPG-ENGREF
Institut National Agronomique
16, rue Claude Bernard
75231 Paris cedex 5, France
(+33) 1 45 49 89 24 (ph.)
(+33) 1 44 08 16 66 (fax)
guillot@inapg.inra.fr

www.inapg.inra.fr/ens_rech/mathinfo/personnel/guillot/welcome.html
Abstract

Geneland is a computer package that allows to make use of geo-referenced individual multi-locus genotypes for the inference of the number of populations and of the spatial location of genetic discontinuities between those populations. The main hypothesis and parameters of the model, as well as the different algorithms to perform inferences are first briefly presented. Major running steps and outputs are then illustrated from the analysis of a simulated dataset, which was also produced by Geneland.
Introduction
Recent developments in molecular markers and statistical tools, combined with powerful computers have led to the emergence of a new scientific field, landscape genetics, which is an amalgamation of population genetics and landscape ecology (Manel et al. 2003). This discipline aims to provide information on how landscape and environmental features influence gene flow, population structure and local adaptation. The two key steps of landscape genetics are the detection and location of genetic discontinuities between populations and the correlation of these discontinuities with landscape and environmental features (e.g. mountains, rivers, roads, gradient of humidity, deforested areas). Ideally, the first step should be based on methods that do not require assumptions of population boundaries beforehand. This implies that the individual is the operational unit of study. Efficient methods to achieve this first step were lacking so far. In a recent paper, Guillot et al. (2005) developed a method which provides an efficient tool for inferring the number of populations at Hardy-Weinberg equilibrium and for locating the genetic discontinuities within a landscape between those populations. Such inferences are carried out from individual geo-referenced multi-locus genotypes, without any a priori knowledge on the populational units and limits. Guillot et al. (2005) have described the spatial statistical model and Markov Chain Monte-Carlo technique used for inferences, and they thoroughly tested, from simulated datasets, the efficiency of the method to estimate the number of populations, assign individuals to populations of origin, map the borders between populations, and detect/map migrants. Their spatial method have shown good behaviors for all these tasks, and it compared favorably well with other methods, i.e. to STRUCTURE (Pritchard et al. 2000; Falush et al. 2003), when comparison was possible (e.g. first and second task). All computations reported in
(Guillot et al. 2005) used a personal computer program written in Fortran language. In order to provide biologists with a user friendly computer package with large functionalities, we have embedded this previous program into a contributed package to R, which is a free statistical software with numerous statistical and graphical functionalities. This paper aims to present this new computer package called Geneland.

**What is an R package ?**

R is a free software for statistical analysis and complex computations available for Unix, Windows and Macintosh (Ihaka and Gentleman 1996; R Development Core Team 2004). Our main motivation to built Geneland as an R package is the very large set of commonly used statistical and graphical functionalities readily available in R and documented in its on-line help. The use of an R package requires a little bit of training. But as a counterpart, it opens the door to the full functionalities of R for data handling and graphics. In addition to the core of Geneland which has been written in Fortran, some extra functions have been written in R in order to facilitate the interpretation of Geneland outputs. Therefore, only very little knowledge of R is necessary to run Geneland and produce one’s owns results.

R is available from [http://www.r-project.org/](http://www.r-project.org/), and instructions for installations are given in the Frequently Asked Questions (FAQ) of this site. Once R is installed, the Geneland package can be installed straightforwardly by typing `install.packages("Geneland")` in the R prompt. This has to be done once only and will install automatically the current version of Geneland from the Comprehensive R Archive Network (CRAN). All the material available from CRAN is under the GNU license. In particular it means that the source code is available and can be freely modified and redistributed provided proper citation is made.
Further details on Geneland (installation, instruction to start and source code) are given on the Geneland homepage:


Main hypotheses and parameters of model

In our spatial statistical model, the whole set of geo-referenced individuals is viewed as belonging to one of several populations at Hardy-Weinberg equilibrium. Individuals within populations are assumed to be randomly located and linkage equilibrium is assumed between loci. The populations are assumed to be spatially organized through the so-called colored Poisson-Voronoi tessellation (Lantuéjoul 2002). Allele frequencies are drawn from Dirichlet distributions which are assumed to be independent (Pritchard et al. 2000) or non-independent among populations (Falush et al. 2003).

Inference is performed via simulation of the posterior distribution of parameters by Markov Chain Monte-Carlo techniques (MCMC). See (Gilks et al. 1996) for an introduction to Bayesian computations and (Clark 2005) for a recent introduction to their use in ecology. Although inference has to be made simultaneously on all the above parameters, one will generally focus on the number of populations and on the posterior probabilities of population membership for any pixel of a given size to locate genetic discontinuities between populations (see section below Illustration from a simulated dataset).

Different algorithms to perform inference

In addition to the data, the user must provide to the program the values of \texttt{delta.coord} and \texttt{npopmax}. The parameter \texttt{delta.coord} is related to the amount of uncertainty attached to spatial coordinates of individuals. If \texttt{delta.coord} = 0, spatial coordinates are considered
as true coordinates. If \texttt{delta.coord} > 0, it is assumed that observed coordinates are true coordinates blurred by an additive noise uniform on a square of side \texttt{delta.coord} centered on 0. \texttt{npopmax} is the upper bound for the number of populations. There is no obvious rule for choosing this parameter. A rule of thumb consists in taking it large enough so that the value given is never reached by the chain along the whole run. The number of populations can also be treated as a known parameter. In that case, the user should provide its value and this variable is not updated in the MCMC scheme. It is also possible to run the model without any reference to spatial coordinates with parameter \texttt{spatial} set to \texttt{FALSE}; the model then becomes equivalent to that implemented in \textsc{structure} without admixture and without linkage disequilibrium (Pritchard et al. 2000). Eventually, the user may choose between two models for allele frequencies: D-model and F-model (cf. previous section). Considering the choice about the number of populations, the way geographical information is handled and the frequency model (and disregarding the choice about \texttt{delta.coord} which is a minor option), we end-up with \(2 \times 2 \times 2 = 8\) different algorithms to perform inference.

It is worth mentioning that we have included in Geneland an R function (\texttt{simFmodel}) that allows simulating datasets of geo-referenced individual multilocus according to the spatial F-model, and hence to test the ability of our MCMC algorithm to retrieve known true parameters. Moreover, a few functions have been developed to help in the convergence diagnostic of MCMC (\texttt{Plotnpop, PlotFreq, Plotntile}), and to post-process the simulated parameters (\texttt{PostProcessChain}).

\textit{Illustration from a simulated dataset}

Although the program can handle large data sets (several hundreds of individuals genotypes at several tens highly polymorphic loci), the use of Geneland is illustrated here with a small
dataset of two populations (total of hundred individuals), and five loci with five alleles per locus. The two populations are separated by a straight line on the unit square, and allele frequencies follow the spatial F-model. This dataset is simulated using the following command under R:

```r
simdata = simFmodel(nindiv=100,coord.lim=c(0,1,0,1),number.nuclei=2,nloc=5,
nall=c(5,5,5,5,5),npop=2,drift=c(.5,.5),seed=12)
```

We get the map given on figure 1. Note the random location of individuals of each population.

![Figure 1 about here.](image)

We now forget the true known parameters and try to make inference of all of them, including the number of populations itself. This is done using both spatial and genetic information (`spatial=TRUE`), the spatial D-model as a prior for allele frequencies (`freq.model="Dirichlet"`), with 10000 iterations (`nit=10000`), saving only one each ten (`thinning=10`), and by using the following command:

```r
mcmcFmodel(simdata$coordinates,simdata$genotypes,simdata$allele.numbers,
            path.mcmc="/tmp/",rate.max=100,
delta.coord=0,npopmin=1,npopinit=1,npopmax=10,
            nb.nuclei.max=200,nit=10000,thinning=10,
            freq.model="Dirichlet",varnpop=TRUE,spatial=TRUE)
```

The path passed to `path.mcmc` should be an existing directory and should end with a "/". The data should be existing R objects typically loaded in R through function `read.table` (see R on-line help).
Note that the D-model was found to perform generally better than the F-model for inferences, even for datasets simulated under the F-model (see Guillot et al. 2005 for details)

The number of populations along the run is displayed on Figure 2 which can be obtained by: 
\[ \text{Plotnpop(path.mcmc}=/tmp/) \]

[Figure 2 about here.]

A clear mode at npop=2 suggests to consider this value as a sensible estimate of the number of populations present in the dataset. Because this first chain contains states with various values for the number of population, it is not recommended to compute averages of other parameters from this chain. Indeed, the population labelled \( k \) may have very few to do with the population labelled \( k \) several thousands of iterations later, when the number of populations has moved across a broad range of values. It is therefore recommended to re-run the function mcmcFmodel with npopmin=npopinit=npopmax=2 and varnpop=FALSE, namely:

\[
\text{mcmcFmodel(simdata$coordinates,simdata$genotypes,simdata$allele.numbers,}
\text{ path.mcmc}="/tmp/",\text{rate.max}=100,\text{delta.coord}=0,\n\text{npopmin}=2,\text{npopinit}=2,\text{npopmax}=2,\n\text{nb.nuclei.max}=200,\text{nit}=10000,\text{thinning}=10,\n\text{freq.model}="\text{Dirichlet}"\text{,varnpop=FALSE,\text{spatial}=TRUE)}\]

This second run (with npop=2) gives new outputs. These outputs of the MCMC algorithm should be post-processed to obtain posterior probabilities of population membership for any pixel of size say 1/50 \times 1/50 by the following command:

\[
\text{PostProcessChain(simdata$coordinates,simdata$genotypes,simdata$allele.numbers,}
\text{...)}\]
Then the map of posterior probabilities of population membership for any pixel can be obtained by using the following command:

```r
PlotTessellation(simdata$coordinates, path.mcmc="/tmp/", printit=TRUE, path="/tmp")
```

The options `printit=TRUE` actually produces files in the directory "/tmp" suitable for inclusion in a manuscript. This map (figure 3) provides the spatial location of genetic discontinuities (i.e. genetic boundary between the two populations). Recent migrants between populations can be usually detected in a map similar to the one of Figure 3 (results not shown). It sometimes happens that some of the $k$ populations inferred in the first run (with a variable number of populations) does not show up on such maps because they appear to be the modal population for no individual of the dataset. In this case, the best estimate for the number of populations is the total number of populations which really show up in the map obtained by `PlotTessellation` and the output file of function `PosteriorMode`.

[Figure 3 about here.]

The computing time for the above sequence of command is of the order of a few tens of seconds on a PC equipped with a 2GHz chip-set. For a larger set of 500 individuals with 20 loci with 20 alleles per locus, the computing time is about 20 minutes for a run of length 100000. Certain datasets may require several runs to assess convergence of the Markov chain, as often recommended in MCMC-based inferential methods (Gilks et al. 1996). Further details on the different functions of Geneland are given in the manual available from [http://cran.r-project.org/doc/packages/Geneland.pdf](http://cran.r-project.org/doc/packages/Geneland.pdf). An on-line help is also avail-
able as explained from the Geneland homepage and examples given in the on-line help can be easily modified and run by cut-and-paste.

**Acknowledgements:** This work has been supported by a grant of Bureau des Ressources Génétiques. We want to acknowledge Annie Bouvier from INRA-MIA Jouy-en-Josas for providing first insight into writing R packages, the active members of the R mailing list for their answers to numerous questions, and Aurélie Coulon and Jean-François Cosson for comments on first releases of the package.

**LITERATURE CITED**


Manel, S., M. Schwartz, G. Luikart, and P. Taberlet, 2003 Landscape genetics:


List of Figures

1 Locations of simulated individuals. True population membership is coded as empty and full circles. Larger circles depict locations of simulated nuclei of the Voronoi tessellation. ................................................................. 14
2 Trace of number of population along the MCMC run with variable $n_{pop}$ and histogram of simulated values. ................................................................. 15
3 Map of posterior probabilities of population membership .......................... 16
Figure 1: Locations of simulated individuals. True population membership is coded as empty and full circles. Larger circles depict locations of simulated nuclei of the Voronoi tessellation.
Figure 2: Trace of number of population along the MCMC run with variable \textit{npop} and histogram of simulated values.
Figure 3: Map of posterior probabilities of population membership