Documentation of BIMr (version 1.1)
Bayesian Inference of Migration rates

Pierre Faubet *
pierre.faubet@e.ujf-grenoble.fr

November 26, 2007

*Génomique des Populations et Biodiversité, Laboratoire d'Ecologie Alpine, CNRS UMR 5553 Université Joseph Fourier 38041 Grenoble France

Contents

1 Introduction

BIMr is a free software that makes inferences of recent gene flows in subdivided populations and can identify the environmental factors that are more likely to explain observed patterns using a generalized linear model. The program implements the Bayesian method described by Faubet and Gaggiotti (200?) which also estimates population specific inbreeding coefficients, F, allele frequencies, p and ˜ p , local population F_{ST} s and performs individual assignments, M; individual immigration rates, ˜ m, can also be reported.

The approach requires multilocus genotype data, X, from codominant markers and assumes that loci are unlinked and individuals sampled at random from their source population S. The

inference model is represented by the DAG of the Figure 1. Posterior estimates are obtained using Markov Chain Monte Carlo (MCMC) and Reversible Jump MCMC methods.

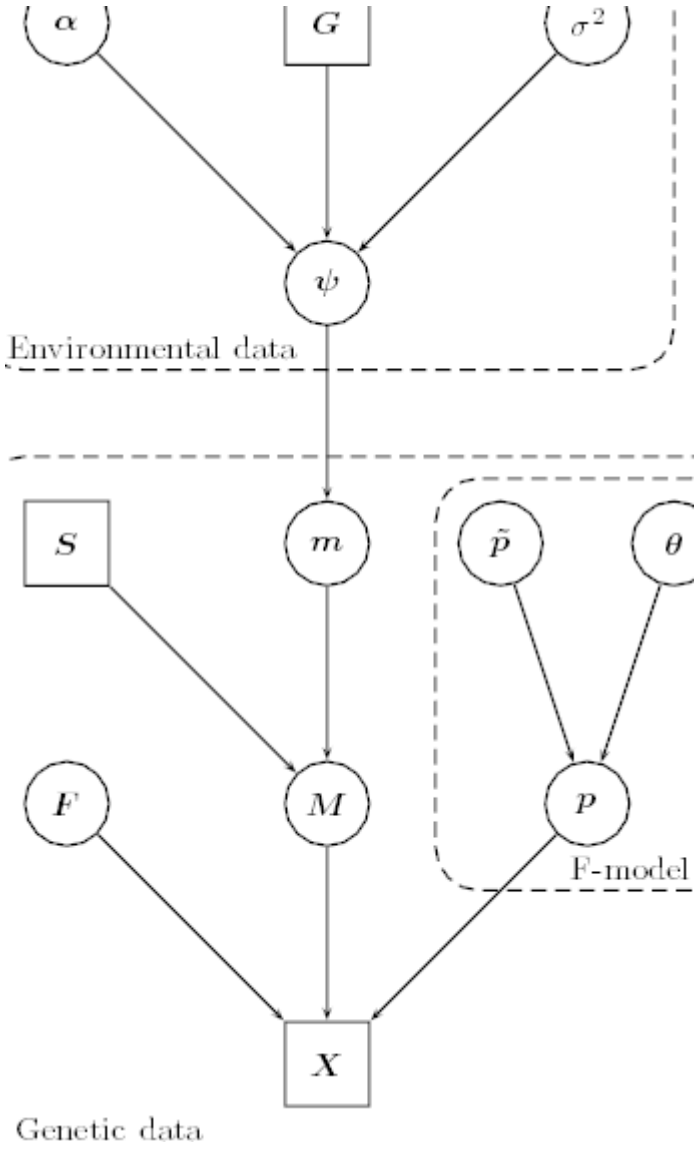

Figure 1: Directed Acyclic Graph (DAG) for the inference model implemented by BIMr. Square nodes denote known quantities (data) and circles represent parameter to be estimated. Lines between nodes represent direct stochastic relationships within the model. The variables within each node correspond to the different model parameters discussed in the text.

1.1 Overview

The software reads data from formatted input files provided by the user. Then parameters for MCMC runs are set and analyses can be performed. Several parallel runs can be performed in order to avoid convergence issues. Advanced tools include plotting and statistical features implemented in order to ease posterior estimation process. Previous results can be loaded.

1.2 Availability

BIMr is available on the three major platforms Windows XP, Mac OS X and Linux with a friendly graphical user interface (GUI). It is distributed under the terms of the GNU General Public License (GPL, version 2.0). The source code is written in C++ and is based on Qt, Qwt and GSL open source libraries.

2 Download and installation

BIMr releases consist of binary versions or can be compiled from source code on your platform [1]. Executables are distributed with approriate platform specific dynamic libraries while compilation requires building them.

2.1 Binary distributions

Binary archives that contain BIMr executable, source code and corresponding dynamic libraries ('*.dll' for Windows XP, '*.dylib' for Mac OS X and '*.so' for Linux). After decompressing the archive, users will find BIMr executable in the 'bin' directory. Desktop shortcuts can be created depending on the platform.

2.2 Source code

Tarballs enclosing BIMr source code are distributed alone and building executable requires appropriate dynamic libraries.

2.2.1 Requirements

As mentionned above the source code makes use of:

- C/C++ compiler (MingW for Windows XP, xcode or gcc for Mac OS X and gcc for linux)
- Qt Open Source Edition (version 4.3.1) [2]
- Qwt - Qt Widgets for Technical Applications (version 5.0.2) [3]
- GSL - GNU Scientific Library (version 1.9) [4]

Thus it is necessary to build and install these softwares and libraries (in the order their appear above). Instructions for this task are enclosed in corresponding distribution documentations. Do not forget to add directories that contain these libraries to your dynamic library path (environment variable PATH for Windows XP, DYLD_LIBRARY_PATH for Mac OS X and LD_LIBRARY_PATH for Linux).

2.2.2 Compilation

Once all these libraries installed, BIMr is ready for compilation after decompressing the source code tarball. First you need to generate a Makefile for the BIMr Qt project using qmake.

Windows XP Start a Windows Shell where Qt4 is initialized. (F.e. with 'Programs->Qt by Trolltech ...->Qt 4.x.x Command Prompt') and move to the root of the uncompressed directory. Then type:

1. qmake in order to create the Makefile
2. then make builds the application
3. and finally complete installation by typing make install

Unix and Mac OS X Open a terminal and type the following commands:
qmake-qt4
make
make install as root

3 Getting started with BIMr

BIMr provides an intuitive and user friendly graphical interface. Before running analyses, data must be appropriately formated. Wizard, dialogs and statistical tools will help users to set up, perform, and interpret MCMC runs or to load previous results.

3.1 File formats

Both genetic and environmental data must be stored in two different files whose formats are described in the following paragraphs. A file converter that supports GENEPOP input format is supplied so that users can easily prepare their genetic data (see 3.6).

Both data files for BIMr must be text files in UNIX format and must end with a new line. Note that sections are defined by character strings between square brackets followed by an equal symbol and are mandatory for both input files (see examples below). Note that the software is case sensitive.

3.1.1 Multilocus genotypes

The suffix of file that contains multilocus genotype data must be '*.gen'. Here is an example:

```
[individuals]= 400

[populations]= 4

[loci]= 10

[alleles]= 11

[genotypes]=
Ind1,   1  0202 0704 0303 0601 0202 0506 0505 0504 0203 0601
Ind2,   1  0202 0407 0403 0110 0602 0505 0505 0205 0201 0504
Ind3,   1  0201 0101 0303 0601 0202 0505 0510 0505 0505 0605
      .  .  .  .  .  .  .  .  .  .  .
h,     Sh loc1 loc2  .  .  .  .  .  .  . locJ
      .  .  .  .  .  .  .  .  .  .  .
Ind398,4  0101 0707 0304 0607 0202 0502 0303 0402 0304 0501
Ind399,4  0203 0507 0403 0606 0202 0505 0303 0503 0503 0605
Ind400,4  0111 0107 0307 0806 0202 0405 1003 0305 0304 0605
```

The first lines of the file must consist of the size of the data (numbers of individuals, populations, loci and alleles) as described above. Therefore, in this dataset, a total of n = 400 individuals sampled from I = 4 populations were scored at L = 10 loci and K = 11 alleles were observed at the most polymorphic locus (excluding missing data).

Then the following n + 1 lines contain the multilocus genotype section. Individuals are in rows and loci in columns. The first column consits of the identifier, h, for each individual and the second one stores the population, Sh, where individual was sampled from. The following J columns contain multilocus genotypes; alleles are two digit coded, from 1 to K, and missing alleles are represented by 00. Then, for each locus, the first two digits encode the first allele and the last two digits the second one. Note that columns are space or tab delimited and are not necessarily aligned. The first column, that contains individuals' identifiers, is not mandatory.

3.1.2 Environmental factors

The file that contains values for environmental factors requires almost the same format. First numbers of populations and factors (without first order interactions) must be given. Then a new section begins for each factor followed by a square matrix with zero diagonal. Each row represents the focal population (in the same order they appear in the multilocus genotype input file). For each pair of distinct populations, values for environmental factors are given.

[populations]= 4

[factors]= 2

[G]= 1
0 0.987519 0.879926 -1.89132
0.987519 0 -0.116062 -0.500205
0.879926 -0.116062 0 0.640139
-1.89132 -0.500205 0.640139 0

[G]= 2
0 0.240328 -1.68737 0.143252
-0.240328 0 1.15934 -0.444753
1.68737 -1.15934 0 -1.23799
-0.143252 0.444753 1.23799 0

Note that the software will normalize these values and will add first order interactions if considered.

3.2 Running BIMr

BIMr executable can be run from a command line or by double-clicking on its icon. This will open the GUI and start the Wizard.

3.3 Wizard and Dialogs

BIMr software implements a wizard and dialogs in order to help users to prepare and perform their analyses that can be saved and used later.

3.3.1 Perform analyses

The main feature of BIMr is its ability to perform MCMC analyses. There are several steps in order to prepare this task, each one consists of giving appropriate information.

Input files First data are needed (see Figure 2), valid path to input files must be given. Multilocus genotype file is always required. If non genetic data check box is filled, then environmental factor file must be supplied. In the last case first order interactions are considered if corresponding check box is filled. Note that filenames can be given by typing the absolute paths to data files or using tool buttons that open file dialogs in order to browse your system.

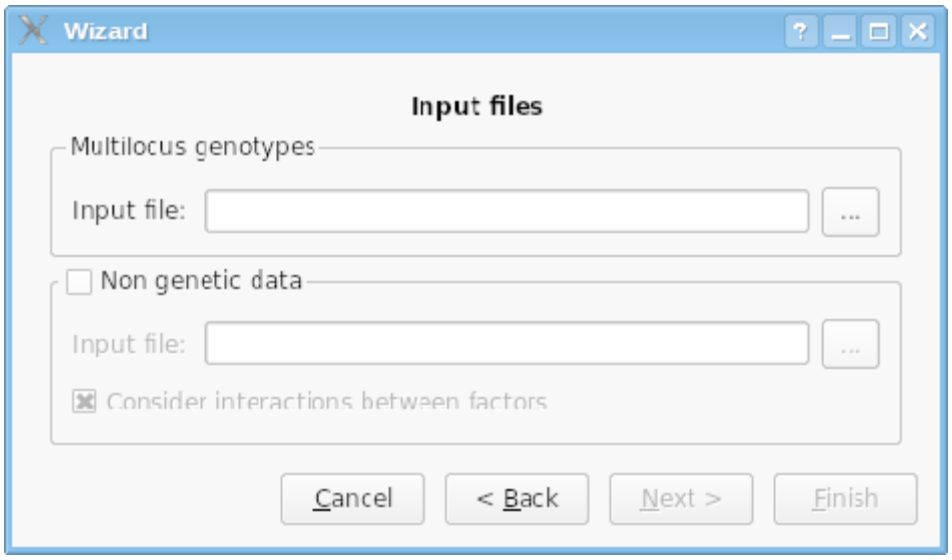

Figure 2: Input file dialog

Once this step is completed, the next one consits of setting MCMC runs.

MCMC Settings Figure 3 presents all MCMC parameters in group boxes. Below are details of the contents of these group boxes.

Run does not require a lot of comments. The burnin is the number of iterations before sampling (i.e. until the Markov chain has converged), the sample size is the number of iterations used for posterior estimation, the thining interval is the number of iterations between two samples and the number of replicates is the number of MCMC runs to perform.

Advanced features consist of using RJMCMC to identify which environmental factors are more likely to explain observed immigration rates (enabled if non genetic data are considered), correlated allele frequency model (see 3.5) and tuning up proposal distributions with short pilot runs (see Incremental values below).

Prior distributions for parameters are shaped according to given values that can be adjusted in order to consider vague or harsh prior densities.

- s2_alpha: variance of the normal distribution for regression parameters.
- a_tau, b_tau: shape and rate parameters of the gamma distribution for the inverse of the deviation from the exact regression.
- psi: shape parameter of the Dirichlet distribution for proportions of migrant genes (only when environmental factors are not considered).
- lambda: shape parameter of the Dirichlet distribution for allele frequencies
- omega, xi: mean and standard deviation of the (log-)normal distribution for local population F_{ST} s.

Incremental values are parameters for proposal distributions within Metropolis-Hastings updates. If pilot runs are used then these values are starting points and will be modified to obtain acceptance rates between 0.25 and 0.45, otherwise they remain fixed.

- e_m: proportions of migrant genes.
- e_nm: proportions of non-migrant genes.
- e_F: population specific inbreeding coefficients.
- e_p: allele frequencies.
- s2_theta: local population F_{ST} s.
- s2_psi: shape parameters of the Dirichlet prior for migration rates.

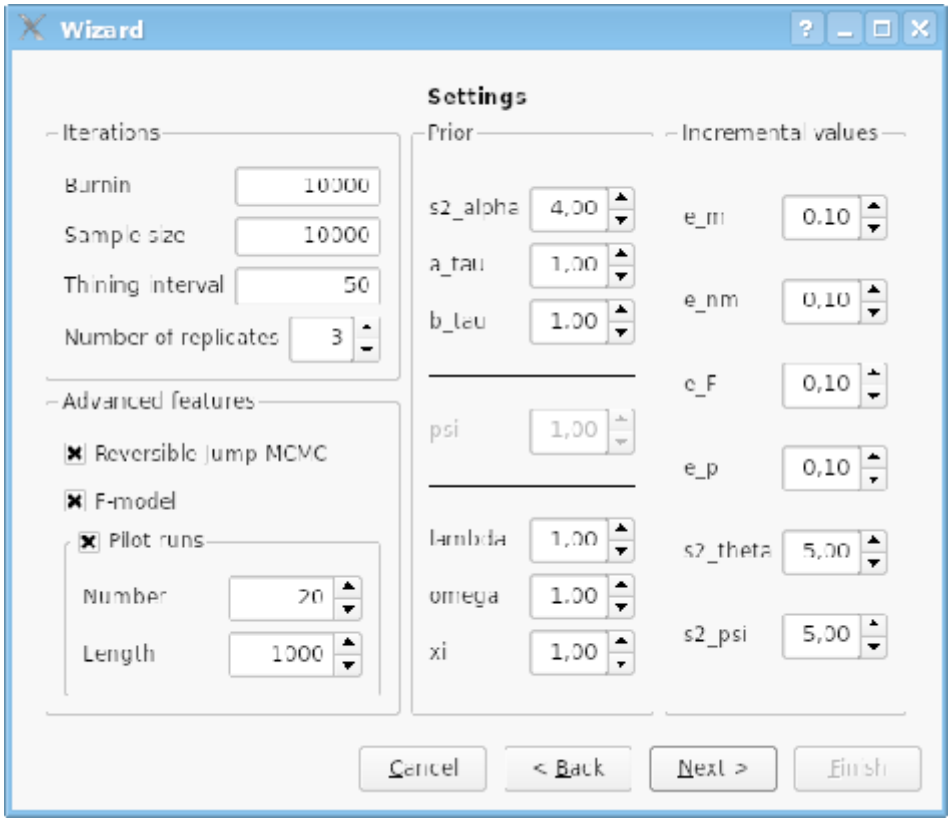

Figure 3: MCMC setting dialog

Then the next step consists of specifying the content of ouputs.

Output options specify parameters and results to print out (see Figure 4). Note that each parameter is saved in a separate file that respect BOA ASCII text file or distruct format specifications (see 4).

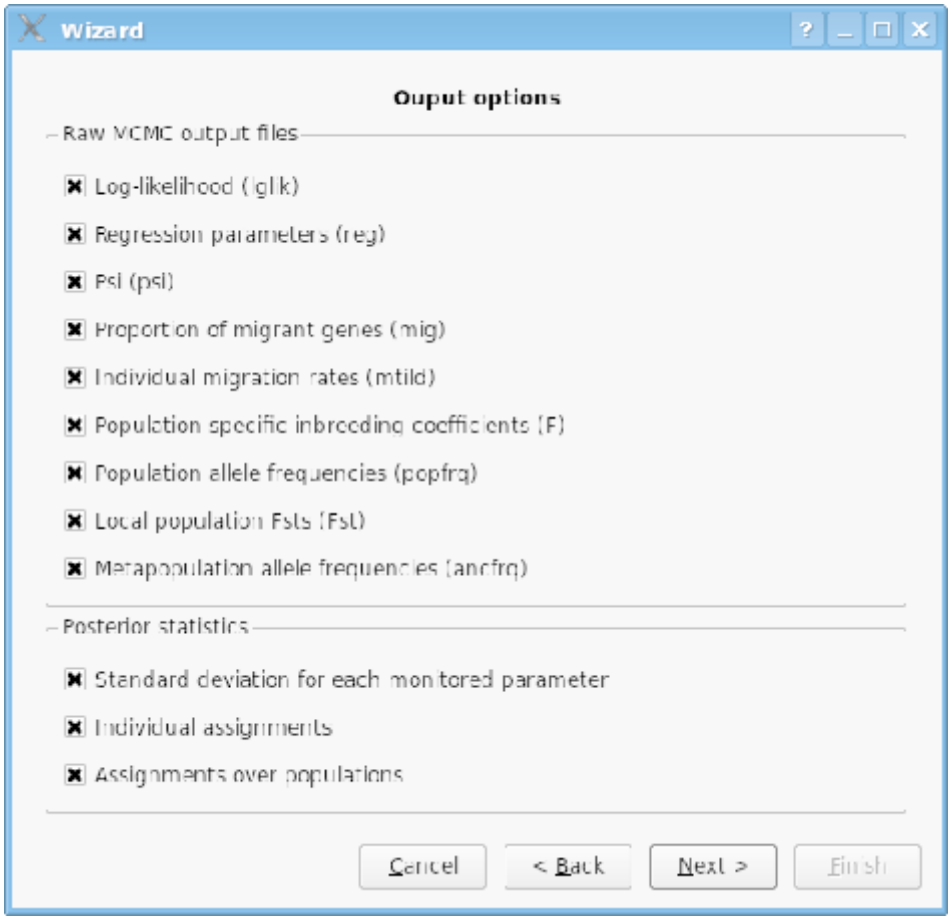

Figure 4: Output dialog

Finally, BIMr asks if analyses can be performed. If it is not the case then users can launch MCMC runs later by clicking the run push button of the GUI. Once there is running job its progress and acceptance rates are displayed in the 'Run' tab of the GUI and you can stop the process at any moment. Then output files remain valid and contain what has been done up to the time you stopped the process but you won't be able to load results. If all the runs terminate normally a message appears and a project file ('*.bpj') is created for future loading.

3.3.2 Load previous results

All results and optionnal output files are available in the directory that contains multilocus genotype data file. They can be imported in BIMr by loading the corresponding project file ('*.bpj') from the wizard. Then posterior estimates are computed using the statistical tools provided by the software (see 3.4).

3.4 Posterior estimation process

BIMr implements a plotting utility and a statistical toolbox that performs kernel density estimation.

3.4.1 Plot

The replicate and the parameter to plot are selected with combo boxes of the 'Plot' tab. Iterations can be discarded by adjusting burin and burnout sliders.

Several types of plot are available by clicking radio buttons:

- Trace displays history for the specified parameter.
- Running mean computes and displays the running mean for the specified parameter.
- Histogram constructs and displays an histogram from the given number of bins for the specified parameter.
- Density estimates and displays the density function from the given kernel and bandwidth for the specified parameter.

Current plot can be saved by entering the 'Posterior' menu and then can be included in Open Office Writer or in LATEXdocuments. Note that several image formats are supported.

3.4.2 Posterior statistics

Several posterior summary statistics for the specified parameter are displayed at the left of the plot. Two posterior estimates are available: the mean of the sample and the mode computed from the estimated density function. The level of HPDI can be set manually by users.

Two additionnal tabs present posterior summary statistics across replicates. The first one, 'Statistics', deals with regular parameters while the second one, 'Model', presents posterior model probabilities and regression parameter estimates for the current model of the combo box (when environmental factors are considered). Summary statistics in these tables consist of posterior mean and mode and 95% HPDI.

Additionnally, BIMr computes the Bayesian deviance for assignments,

$$D_{assign} = -2\overline{\log Pr(M|m)}$$

Following Faubet et al. (2007) we suggest to carry out many MCMC runs, say 10, and select the one with the lowest deviance for obtaining the parameter estimates in order to minimize convergence problems.

Results can be exported to text files that can be imported in spreadsheet programs such as Open Office Calc or in R.

3.5 Correlated allele frequency model

BIMr can estimate parameters for the F-model (advanced feature). It uses local population F_{ST} s and the global allele frequencies as a prior for population allele frequencies.

The F-model was first implemented in order to improve convergence properties for population allele frequencies. Note that this feature becomes obsolete when intial values for these parameters are closed to the real ones.

3.6 File conversion utility

BIMr provides a file converter in order to import GENEPOP formatted files. This feature is supplied in the File menu and will open a dialog (see Figure 5).

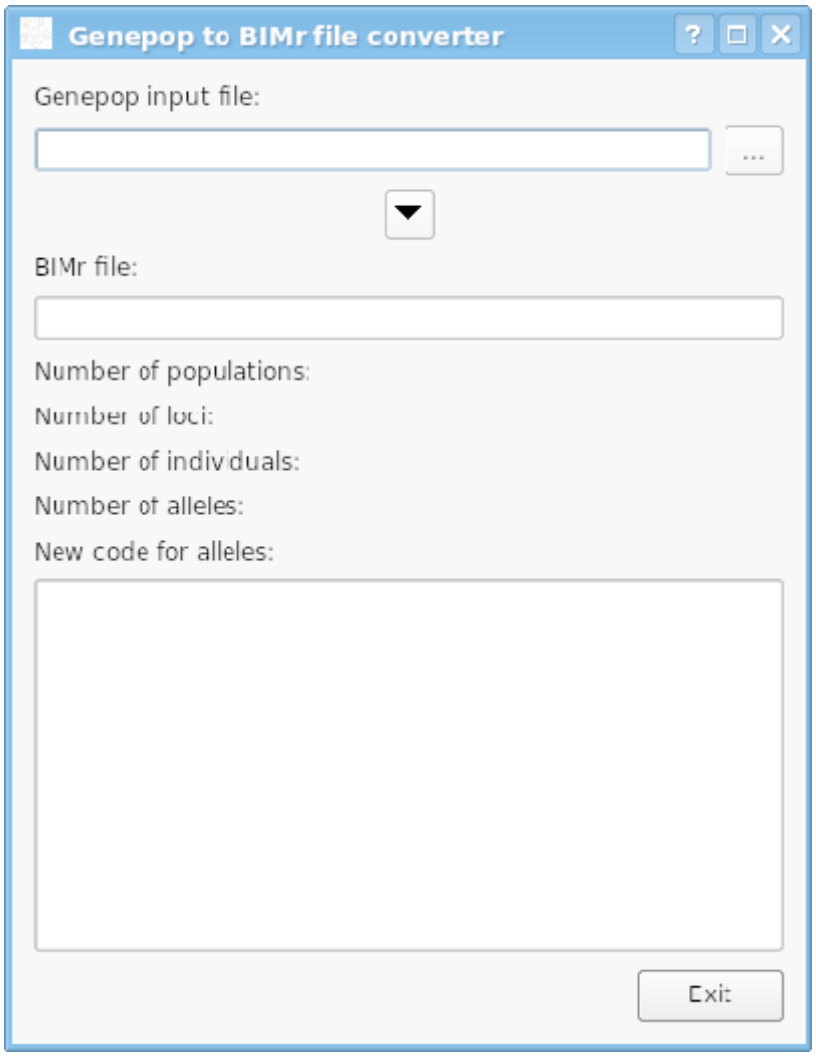

Figure 5: File conversion utility dialog.

The first line must contain the path to the GENEPOP input file which can be supplied manually or by using the tool button on the right and browsing local directories. Click the tool button below (the one with a down arrow) in order to convert your data. If the output filename is not specified manually before conversion, then the user will be asked to give the location for the BIMr formatted file with a file browser. Remember that the suffix for multilocus genotype filename must be '.gen'. Once conversion done, information about the data are displayed and allele codes are given in a table.

4 Other useful softwares

Individual or population assignments can be plotted using distruct sotware provided by Noah Rosenberg's lab [5] .

Additionnal posterior analysis can be performed using the Bayesian Output Analysis (BOA) R package developed by Brian Smith [6] . Raw MCMC runs can be easily imported as they respect the format required by BOA.

5 Credits and License

Many thanks are owed to those who spend many hours of contributions sacrificing their time to develop tools we can all use under the GNU license.

Here is a short list of people who made this project possible.

Contributions from: Oscar Gaggiotti, Jochen Wolf

Many thanks to all those involved!

6 References

Faubet, P., R. S. Waples and O. E. Gaggiotti, 2007 Evaluating the performance of a multilocus Bayesian method for the estimation of migration rates. Molecular Ecology 16: 1149–1166.

Rosenberg, N. A., 2004 DISTRUCT: a program for the graphical display of population structure. Molecular Ecology Notes 4: 137–138.

Raymond, M. and F. Rousset, 1995 GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J. Heredity 86:248–249