Inferring recent Migration rates from Individual Genotypes

A quick-start guide to the IMIG scripts

Thomas Broquet and Jon Yearsley
March 2009

Contact: thomas.broquet@unil.ch
jon.yearsley@ucd.ie

This document provides guidelines for using the migration inference method presented in Broquet et al. (2009), which we will refer to as "IMIG". IMIG aims at estimating pairwise rates of migration between discrete populations using multilocus genotypes. The inference is based upon changes in the genetic composition of populations during the dispersal phase of a species' life cycle. IMIG requires populations to be sampled twice: the first sample, "sample 1", is the pre-dispersal data, and some time later the second sample, "sample 2", is the post-dispersal data. IMIG will estimate the pairwise rates of dispersal that occurred in the time between these two samples. IMIG can be used when dispersal occurs over an extended period of time, but IMIG will not estimate dispersal that occurred before sample 1, or dispersal that continued after sample 2.

IMIG is based on scripts developed in the R environment for statistical computing (R Development Core Team 2007) and uses the Bayesian computation program OpenBugs (Thomas et al. 2006). After software installation, using IMIG involves four steps:

1. Parameterize the model
2. Format the data
3. Run the analysis
4. Visualize the results.

IMIG may be used in different ways, depending upon the users' objectives (and R skills!): An R script is used to format the data, while the user-friendly windows-based interface of OpenBugs is used to run the model and visualize the results. Alternatively, the model can be run and visualized entirely within the R environment (using package BRugs, Thomas et al. 2006). We provide the scripts for this, which will prove useful for analyzing repetitive data sets (typically produced through simulations) or when a large number of populations must be analyzed (e.g. >10). We recommend using openBugs for most applications of the method because in our experience it is quicker and more robust than the BRugs package.

In addition to this document, important information can be found in the reference paper (Broquet et al. 2009). Further useful implementation details can also be found within the R scripts, which are extensively commented.

0- Software installation

- R can be freely downloaded from http://www.r-project.org/. The additional packages HIERFSTAT (Goudet 2005) and BRugs (Thomas et al. 2006) must be downloaded and installed (this can be done using R menus *Packages*/*install package(s)*).

- A version of openBugs is bundled within the BRugs package. However, we recommend that IMIG is run directly in openBugs (without the BRugs interface) and for this OpenBugs must be downloaded (http://mathstat.helsinki.fi/openbugs/) and unzipped.

- IMIG files can be downloaded from http://www.unil.ch/dee/page6759_en.html. The files must be unzipped in a directory where input files will also be found.

1- Model parameterisation

The parameters for the model are specified in the file *imig_params.r*. The file is in two sections:

a) parameters that must be specified to format the data
b) parameters that must be specified if BRugs is to be run

a) Parameters to format the data

sample1Filename

The name of Fstat input file containing pre-dispersal data

sample2Filename

The name of Fstat input file containing pre-dispersal data

demeGroups

A variable allowing the user to group populations together before running the analysis. *demeGroups* is a vector specifying which group each population should be placed in. A group called 0 will be removed from the analysis. For instance: c(1,2,2,3,0) means the raw data has 5 populations with populations 2 and 3 grouped together and population 5 not included in the analysis. Setting *demeGroups* = 1 will specify that no grouping is required.

mPrior_Mean and *mPrior_CV*

Define prior distribution (mean and coefficient of variation) for probabilities of migration. Default values will generally do it.

nChain

The number of MCMC chains to be run with IMIG. Initial data are generated for each chain. We recommend using several chains, although there are other views (see for example this page about MCMC methodology and diagnostics http://www.stat.umn.edu/~charlie/mcmc/diag.html)

runBRugs

if TRUE then the script *imig_launch.r* formats the data and runs the IMIG model with BRugs.

if FALSE then the script *imig_launch.r* just formats the data ready for use in openBUGS (recommended for most applications).

b) Parameters to run BRugs. If BRugs isn't being used then these parameters (with the exception of *overdispersedInits*) will be set when you compile and run a model in openBUGS.

paramSave:

A list of parameters to be saved for downstream analysis. The parameter that we are ultimately interested in is *m*, but saving other parameters (described in Broquet et al. 2009) is very useful to analyze the behavior of the model (e.g. chain convergence).

outputPrefix

The prefix that will be used to name the files created by IMIG

burninIters

The length of the burnin period (4000 is a minimum, but this will depend upon the data analyzed. A burnin of 20000 was used in Broquet et al. 2009).

datarecordIters

The length of chain used to record the data (after a burnin period). We used 2000 in Broquet et al. (2009).

thinParam

The number of iterations between each recorded value. A *datarecordIters* set to 2000 with a thin of 100 gives a total chain length of 200 000 iterations.

overdisperseInits

If TRUE then the initial data are over dispersed. This is done by running a first-pass model and then selecting initial data from the tails of the posterior distributions of the first-pass model. We recommend leaving this option to TRUE.

2- Formatting the data

We start with two input files containing sample 1 (pre-dispersal) and sample 2 (post-dispersal) genotypic dat. These input files must be in Fstat format (www2.unil.ch/popgen/softwares/fstat.htm, Goudet 1995, Goudet 2001). Alleles can be encoded using 2 or 3 digits. Examples of such files are provided: *sample1_Genotypes.txt* and *sample2_Genotypes.txt*.

These files are placed in your working directory (e.g. D:/Recherche/IMIG/). Then open R and type the following commands (Note: the following command lines (including comments) can be pasted directly in R)

```
#In R, first specify your working directory:
setwd("D:/Recherche/IMIG")
# Run the IMIG model
source("imig_launch.r")
```

These commands will produce several files in the working directory. The files are of two broad types: one data file for each focal population, formatted for use in OpenBugs (e.g. *imigModel_FocalPop1_Data.txt*), and one initialization file for each focal population and each MCMC chain (e.g. *imigModel_FocalPop1_Inits1.txt*). These files are ready to be used in OpenBugs.

3a- Running IMIG using openBugs

- Run winbugs.exe (in the directory where OpenBugs was previously installed). The software has extensive help functions.

Stage 1: Model Initialisation:

- Drag and drop the model file *imig_model.bug* into the window of winbugs.exe. Alternatively the *File/Open…* menu can be used to import the file.
- Open the *Model Specification Tool* by selecting the menu *Model/Specification…*
- Hit the *check model* button (wait until "model is syntactically correct" appears at the bottom left of the winbugs.exe window).
- Drag and drop the data file for the focal population into the window of winbugs.exe (e.g. when population 1 is the focal population this will be the file *imigModel_FocalPop1_Data.txt*)
- Hit the *load data* button in the Specification Tool (wait until "data loaded" appears at the bottom left of the winbugs.exe window).
- Then set the number of chains in the Specification Tool.
- Hit the *compile* button in the Specification Tool (wait for "model compiled" to appear at the bottom left of the winbugs.exe window. For large data sets this may take some minutes)

- Just as for the data file, drag and drop the text file containing initial values for the first chain of population 1 (e.g. *imigModel_FocalPop1_Inits1.txt*) and hit the *load inits* button in the Specification Tool. If several chains are specified then repeat this last operation for the initial value data of each chain.
- The information bar at the bottom of the interface should read "model is initialized". If it does not, some data has not been initialized. You can either generate this initial data by selecting the *gen inits* button of the Specification Tool, or you can go back to the files and find out what is missing.

Stage 2: Running the MCMC chains

- Open the *Model/Update…* menu
- Set the *updates* option to the number of iterations corresponding to the length of the burnin period.
- Set the thin parameter (the total number of iterations is thin*updates. For the burnin we use a thin=1)
- Hit *update* in the Update Tool.

- Open the *Inference/Samples…* menu
- To start recording data on the migration rate parameter of the model, type "m" in the *node* box and hit *set*. Repeat this operation with the other model parameters for which information must be saved (e.g. "phi").
- In the Update Tool menu, set updates to the final number of iterations that must be run (e.g. 2000), and set the thin parameter (we use a thin = 100).
- To generate the final MCMC chains hit the *update* button of the Update Tool… and wait.

- When the updating is complete, the results can be visualized with the Sample Monitor Tool under the *Inference/samples…* menu (see Section 4 below). When the focal population is population 1, m[1] is data used to estimate the proportion of individuals in population 1 originating from population 1 (i.e. philopatry), m[2] the proportion individuals in population 1 originating from population 2, etc. After the visualization step, the MCMC chains can be saved before repeating the whole process with another focal population.

3b- Running IMIG using BRugs

- In the parameter file set *runBRugs=TRUE*
- Now when the command *source("imig_launch.r")* is typed into R (see the section on formatting the data above) the data will be formatted as before, and BRugs will be immediately run for each focal population. This may take a while depending upon the amount of data you have.
- The results of the model (which are the MCMC chains) will be saved in "CODA files" starting with the prefix given in the parameter file *outputPrefix.*
- These files can be analysed with the CODA package in R (Plummer et al. 2006).

4- Visualizing the results

- In the *Sample Monitor Tool* menu, select node "m" (or any other parameter that must be analyzed) and hit the *stats* button.
- The median of the posterior distribution is usually a good estimator of migration rates (m[1] gives information on $m_{1,1}$ (philopatry) while m[2], m[3],…m[n] gives information on $m_{1,2}$, $m_{1,3}$…$m_{1,n}$). Highest posterior density intervals (HPDI) are given by the 2.5 and 97.5 percentile values (see Broquet et al. 2009 for comments on HPDI).

BEWARE OF MCMC SAMPLING!

The reliability of the results depends upon the behavior of the MCMC sampling. Chain convergence must be carefully checked. In particular, we recommend using the *trace* tool of the *Sample Monitor Tool* menu to visualize the behavior of parameter "phi". Independent mcmc chains are represented with different colors so that **obvious** convergence issues can be easily seen. Keep in mind the scale for each parameter. A value of phi=-10 and a value of phi=-20 correspond to migration rates of roughly 10^{-7} and 10^{-11}, respectively, which for most populations will be effectively the same (i.e. no migration). The Gelman-Rubin convergence statistic can also be used (button *bgr diag* of the *Sample Monitor Tool* menu). **Read** carefully the *Checking convergence* paragraph of WinBugs' help for more information (see also *bgr diag* in the *inference menu* help section).

Citation

When this method is used, it may be cited as:
Broquet T, Yearsley J, Hirzel AH, Goudet J, Perrin N. 2009. Inferring recent migration rates from individual genotypes. *Molecular Ecology* 18: 1048-60

References

Broquet T, Yearsley J, Hirzel AH, Goudet J, Perrin N. 2009. Inferring recent migration rates from individual genotypes. *Molecular Ecology* 18: 1048-60

Goudet J. 1995. F-STAT (vers. 1.2): a computer program to calculate F-statistics. *Journal of Heredity* 86: 485-86

Goudet J. 2001. FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3), updated from Goudet 1995. Available from http://www.unil.ch/izea/softwares/fstat.html.

Goudet J. 2005. HIERFSTAT, a package for R to compute and test hierarchical *F*-statistics. *Molecular Ecology Notes* 5: 184-86

Plummer M, Best N, Cowles K, Vines K. 2006. coda: Output analysis and diagnostics for MCMC.

R Development Core Team. 2007. R: A language and environment for statistical computing. Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org: R Foundation for Statistical Computing.

Thomas A, O Hara B, Ligges U, Sturtz S. 2006. Making BUGS Open. *R News* 6: 12-17