DIYABC

version 2.0

A user-friendly software
for inferring population history through
Approximate Bayesian Computations

J.M. Cornuet, P. Pudlo, J. Veyssier,

A. Dehne-Garcia, A. Estoup and V. Ravigné
Centre de Biologie et de Gestion des Populations

Contents

1. Preface

In less than 10 years, Approximate Bayesian Computations (ABC) have developed in the Population Genetics community as a new tool for inference on the past history of populations and species. Compared to other approaches based on the computation of the likelihood which are still restrained to a very narrow range of evolutionary scenarios and mutation models, the ABC approach has demonstrated its ability to stick to biological situations that are much more complex and hence realistic. However, this approach still requires numerous computations to be performed so that it has been used mostly by specialists (i.e. statisticians and programmers). This has almost certainly restrained the possible impact of ABC in population genetic studies. We believe that this situation must be improved and therefore we have developed a computer program for the large community of experimental biologists. We therefore designed DIYABC as a user-friendly program allowing non specialist biologists to achieve their own analysis.

The first version ($DIY ABC v0.x$) had been written especially for microsatellite data. There were at least two reasons for that. The first one is that we have been among the first to develop and use this class of markers in population genetic studies (e.g. Estoup *et al.*, 1993). Since then, we have developed microsatellites in numerous species as well as we have published theoretical studies and reviews on these markers (e.g. Estoup *et al.*, 2002). The second reason is that microsatellites have been and still are very popular markers in the population geneticist community and there is now a large quantity of data that might benefit of an ABC approach.

The second version of our software ($DIY ABC v1.x$) has been designed to make use of DNA sequence data.This has several immediate consequences. For instance, the standard Genepop data file format has been extended to incorporate sequence data. This has been done in collaboration with the authors of *Genepop* and explained in subsection 4.1.1. In this version, sequence loci are considered in the same way as microsatellite loci, i.e. they are considered as genetically independent and intra-locus recombination is not (yet) available. Concerning mutation models for DNA sequences, we used the same philosophy as for microsatellites, i.e. the program considers only simple and widely used models, keeping in mind that a higher-dimensional parameter space will be less well explored than a lower-dimensional space. Note that none of these mutation models includes insertion-deletions. Also five categories of loci (either microsatellites or DNA sequences) were considered in this second version : autosomal diploid, autosomal haploid, X-linked, Y-linked and mitochondrial. Note that X-linked loci can be used for an haplo-diploid species in which both sexes have been sampled. If non-autosomal loci have been typed in population samples, the sex-ratio of the species will have to be provided (see subsection 4.1.1).

Other improvements over version 0 included :

1. the use of multithread technology in order to exploit multicore/multiprocessor computers. This is especially useful when building the reference table and for several other intensive computation steps, such as the multinomial logistic regression,

2. a new option which helps the detection of "bad" prior modelisation of the data,

3. another new option which helps evaluate the goodness of fit of a given model-parameter posterior combination (i.e. Model checking),

4. many new screens implemented not only to treat sequence data, but also to cope with the new options described above, as well as to offer useful complementary information on the current run.

The third version of $DIY ABC$ ($DIY ABC v2.x$) has been entirely recoded in order to be used under the usual three OS (Windows, Mac and Linux). Also the code for computations has been separated from that of the graphic user interface (GUI). The former has been rewritten in C++ and the latter is a mixture of Python and Qt (PyQt). The user can then launch computations with or without using the GUI. The GUI 's uses are :

1. the management of projects

2. the input of the historical and genetical models

3. the parameterization of analyses

4. the launch of computations of the reference table and of the various required analyses

5. the visualization of results

Also, as DNA sequences have been added in the second version, a new category of markers has been added to the third version : Single Nucleotide Polymorphisms (SNPs). Instead of extending once more the Genepop format, a new data simple format has been designed for these markers.

This version includes all improvements of version 1.x and a few new improvements such as :

- loci of the same type (i.e. microsatellites on one hand or DNA sequences on the other hand) can be associated in one or more groups. This allows for instance to define different mutation models for microsatellites with motifs of different lengths.

- the model checking option is now presented as a direct option (not a suboption of the ABC estimation of parameters) which largely simplifies its use.

- the logistic regression can be performed on factorial discriminant analysis components instead of all summary statistics. This reduces the number of dependent variables, thus allowing to run large "confidence in scenario choice" analyses including many summary statistics and scenarios.

- ascertainment bias in the design of SNPs can be tentatively corrected by considering "reference" samples in which the loci need to be polymorphic in order to belong to the SNP set. These reference samples are not necessarily included of the actual samples.

1.1 Acknowledgements

1.2 References to cite

- **version 0 :** Cornuet J.M., F. Santos, M.A. Beaumont, C.P. Robert, J.M. Marin, D.J. Balding, T. Guillemaud and A. Estoup. Inferring population history with DIYABC: a user-friendly approach to Approximate Bayesian Computations (2008) *Bioinformatics*, **24** (23), 2713-2719.

- **version 1 :** Cornuet J.M., V. Ravigné and A. Estoup, 2010. Inference on population history and model checking using DNA sequence and microsatellite data with the sofware DIYABC (v1.0) (2010) *BMC Bioinformatics*, **11**, 401.

- **version 2 :** Cornuet J.M., Pudlo P., Veyssier J., Dehne-Garcia A., V. Ravigné and A. Estoup. Improved inference on population history using SNP's and DIYABC (v2). *submitted*

1.3 Web site

http://www.montpellier.inra.fr/CBGP/diyabc

2. Methodology

2.1 Basic notions on ABC

Approximate Bayesian Computation or ABC is a bayesian approach in which the posterior distributions of the model parameters are determined by replacing the computation of the likelihood (probability of observed data given the values of the model parameters) by a measure of similarity between observed and simulated data. The posterior distributions are estimated from parameter values providing simulated data that are the most similar to observed data. Historically, different ways of estimating this similarity have been proposed, but all have been based on statistics summarizing information conveyed by the data set. In population genetics, data most often relate to individuals that have been genotyped at a given set of loci, these individuals being representative of the studied populations. The summary statistics are for instance the mean number of alleles per population or genetic distances between pairs of populations. It is much easier to measure the similarity between small sets of summary statistics than between large sets of multilocus genotype data. When the number of summary statistics is low, it is possible to select simulated data for which *all* the summary statistics are close to those of the observed data (Pritchard *et al.*, 1999; Estoup *et al.*, 2001; Estoup and Clegg, 2003). However, for more complex scenarios necessitating a larger number of summary statistics, it becomes almost impossible to find such simulated data sets. Beaumont *et al.* (2002) have hence proposed to measure similarity through the Euclidian distance between observed and simulated summary statistics, after normalization by standard deviations of simulated statistics. In addition, these authors introduced a step of local linear regression aimed at favoring simulated data sets that are closer to the observed one.

In practice, the ABC approach can be summarized in three successive steps (Excoffier *et al.*, 2005) : i) generating simulated data sets, ii) selecting simulated data sets closest to observed data set and iii) estimating posterior distributions of parameters through a local linear regression procedure.

In addition, this approach provides a way of comparing different scenarios that can explain observed data. Two measures of posterior probabilities of scenarios are proposed. The first measure is simply the relative proportion of each scenario in the simulated data sets closest to observed data sets (Miller *et al.*, 2005; Pascual *et al.*, 2007). The second measure is obtained by a logistic regression of each scenario probability on the deviations between simulated and observed summary statistics (Fagundes *et al.*, 2007; Beaumont, 2008).

In order to simulate data, one has first to define one (or possibly several) model(s). Each model includes a historical model describing how the sampled populations are connected to their common ancestor and a mutational model describing how allelic states of the studied genes are changing along their genealogical trees.

2.2 Historical model parameterization

The evolutionary scenario, which is quantified by the historical model, can be described as a succession in time of "events" and "inter event periods". The events considered in the program are a restricted set of possible events affecting populations evolution. In the current version of the program, we consider only 4 categories of events : population divergence, discrete change of effective population size, admixture and sampling (the last one has been added to allow considering samples taken at different times). Between two successive events affecting a population, we assume that populations evolve independently (e.g. without migration) and with a fixed effective size. The usual parameters of the historical model are the times of occurrence of the various events (counted in generations), the effective sizes of populations and the admixture rates. When writing the scenario, events are provided sequentially backward in time. Although this choice may not be natural at first sight, it is coherent with coalescence theory on which are based all data simulations in the program. For that reason, the keywords for a divergence or an admixture event are `merge` and `split`, respectively. Two other keywords, `varNe` and `sample`, correspond to a discrete change in effective population size and a gene sampling, respectively. Eventually, only for SNPs, we have added the keyword `refsample` to control the ascertainment bias (see below).

A scenario takes the form of a succession of lines (one line per event), each line starting with the time of the event, then the nature of the event, and ending with several other data depending on the nature of the event. Following is the syntax used for each category of event :

population sample : $\langle time \rangle$ `sample` $\langle pop \rangle$ $[nmales\ nfemales]$

$\langle time \rangle$ is the time (always counted in number of generations) at which the sample was taken and

⟨*pop*⟩ is the population number from which is taken the sample. It is worth stressing here that **samples are considered in the same order as they appear in the data file**.

[*nmales nfemales*] is only used for SNP loci to indicate the number of males (respectively females) from the sample that have been used to detect SNPs. Thes males and females appear in the corresponding sample of the data file.

"reference" sample : ⟨*time*⟩ refsample ⟨*pop*⟩ *nmales nfemales* **ONLY FOR SNP loci**

⟨*time*⟩ is the time (always counted in number of generations) at which the "reference" sample was taken and

⟨*pop*⟩ is the population number from which is taken the sample.

nmales nfemales indicate the number of males (respectively females) that have been used to detect SNPs in the reference sample. These males and females **do not appear** in the data file.

population size variation : ⟨*time*⟩ varNe ⟨*pop*⟩ ⟨*Ne*⟩

From time ⟨*time*⟩, looking backward in time, population ⟨*pop*⟩ will have an effective size ⟨*Ne*⟩.

population divergence : ⟨*time*⟩ merge ⟨*pop0*⟩ ⟨*pop1*⟩

At time ⟨*time*⟩, looking backward in time, population ⟨*pop1*⟩ "merges" with population ⟨*pop0*⟩. Hereafter, only ⟨*pop0*⟩ "survives".

population admixture : ⟨*time*⟩ split ⟨*pop0*⟩ ⟨*pop1*⟩ ⟨*pop2*⟩ ⟨*rate*⟩

At time ⟨*time*⟩, looking backward in time, population ⟨*pop0*⟩ "splits" between populations ⟨*pop1*⟩ and ⟨*pop2*⟩. A gene lineage from population ⟨*pop0*⟩ joins population ⟨*pop1*⟩ (respectively ⟨*pop2*⟩) with probability ⟨*rate*⟩ (respectively 1-⟨*rate*⟩).

A scenario is a succession of lines as described above. However, in order to cope with special situations (see explanations in Note 9 below), we added a first line giving the effective sizes of sampled populations before the first event described, looking backward in time. Expressions between arrows, other than population numbers, can be either a numeric value (e.g. 25) or a character string (e.g. t0). In the latter case, it is considered as a parameter of the model. So the only possible parameters of the historical model are times of events, effective population sizes and admixture rates.

The program offers the possibility to add or remove scenarios, by just clicking on the corresponding buttons. The usual shortcuts (CTRL+C, CTRL+V and CTRL+X) can be used to edit the different scenarios. Some or all parameters can be in common among scenarios.

Notes

1. There are two ways of giving a fixed value to effective population sizes, times and admixture rates. Either the fixed value appears as a numeric value in the scenario windows or it is given as a string value like any parameter. In the latter case, one gives this parameter a fixed value by choosing a Unifom distribution and setting the minimum and maximum to that value in the prior setting (see section 2.4).

2. All expressions must be separated by at least one space.

3. All expressions relative to parameters can include sums or differences. For instance, it is possible to write :
   t0 merge 2 3
   t0+t1 merge 1 2
   This means that t1 is the time elapsed between the two events. By imposing t1>0 (as explained in section **prior and posterior distributions**), this implies that the divergence of populations 1 and 2 is always more ancient than the divergence of populations 2 and 3. However, one cannot mix a parameter and a numeric value (e.g. t1+150 will result in an error). This can be done by writing t1+t2 and fixing t2 by choosing a uniform distribution with lower and upper bounds both equal to 150.

4. Time is always given in generations. Since we look backward, time increases towards past.

5. Negative times are allowed (e.g. the example given in section 3), but not recommended.

6. Population numbers must be consecutive natural integers starting at 1. The number of population can exceed the number of samples and vice versa : in other words, unsampled populations can be considered in the scenario on one hand, and the same population can be sampled more than once on the other hand.

7. Multi-furcating population trees can be considered, by writing several divergence events occurring at the same time. However, one has to be careful to the order of the `merge` events. For instance, the following piece of scenario will fail :
   ```
   100 merge 1 2
   100 merge 2 3
   ```
   This is because, after the first line, population 2, which has merged with population 1, does not "exist" anymore (the surviving population is population 1). So, it cannot receive lineages of population 3 as it should as a result of the second line. The correct ways are either to put line 2 before line 1, or to change line 2 to :
   ```
   100 merge 1 3.
   ```

8. Since times of events can be parameters, the order of events can change according to the values taken by the time parameters. In any case, before simulating a data set, the program sorts out events by increasing times [1]. If two or more events occur at the same time, the order is that of the scenario as it is written by the the user.

9. Most scenarios begin with sampling events. We then need to know the effective size of the populations to perform the simulation of coalescences until the next event concerning each population. One way would have been to provide the population size on the same line of the scenario description. However, in some scenarios with varying population sizes, it can not be determined what is the effective size at the sampling time before the set of time parameter values is generated. For that reason, we decided to provide the effective size and the sampling description on two distinct lines.

**Examples** Below are some usual scenarios with increasing complexity. Each scenario is coded on the left side and a graphic representation given by *DIY ABC* is printed on the right side

1. One population from which several samples have been taken at various generations : 0, 3 and 10. The only unknown parameter of the scenario[2] is the effective population size.



```
Ne
0 sample 1
3 sample 1
10 sample 1
```
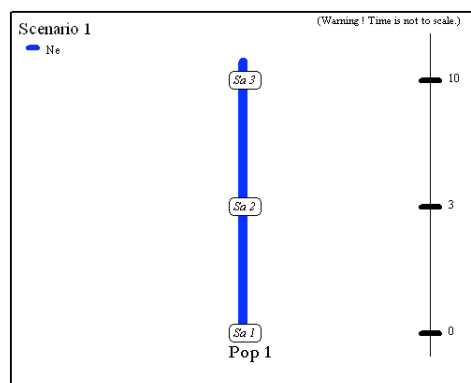
2. Two populations of size `N1` and `N2` have diverged `t` generations in the past from an ancestral population of size `N1+N2`.


[1] Sorting events by increasing times can only be done when all time values are known, i.e. when simulating datasets. When checking scenarios, all time values are not yet defined, so that when visualizing a scenario, events are represented in the same order as they appear in the window used to define the scenario.

[2] Of course, there are also one or more parameter(s) for the mutation model.

```
N1 N2
0 sample 1
0 sample 2
t merge 1 2
t varNe 1 N1+N2
```
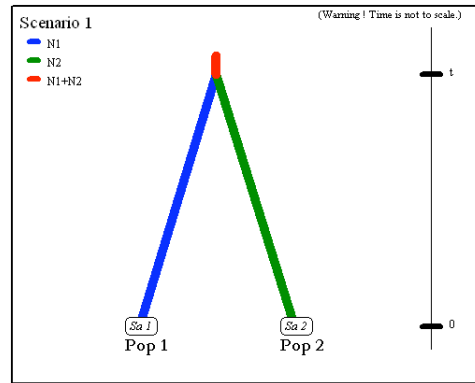
3. Two parental populations (1 and 2) with constant effective populations sizes `N1` and `N2` have diverged at time `td` from an ancestral population of size `NA`. At time `ta`, there has been an admixture event between the two populations giving birth to an admixed population (3) with effective size `N3` and with an admixture rate `ra` relative to population 1.


```
N1 N2 N3
0 sample 1
0 sample 2
0 sample 3
ta split 3 1 2 ra
td merge 1 2
td varNe 1 NA
```
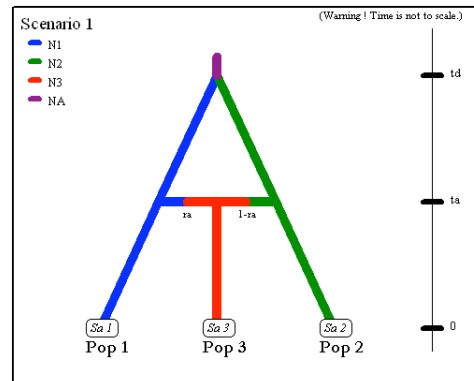
4. The next scenario is slightly more complicated. It includes four population samples and two admixture events. For simplicity sake, all populations are assumed to have identical effective sizes (`Ne`).


```
Ne Ne Ne Ne Ne
0 sample 1
0 sample 2
0 sample 3
0 sample 4
t1 split 3 1 2 r1
t2 split 2 5 4 r2
```
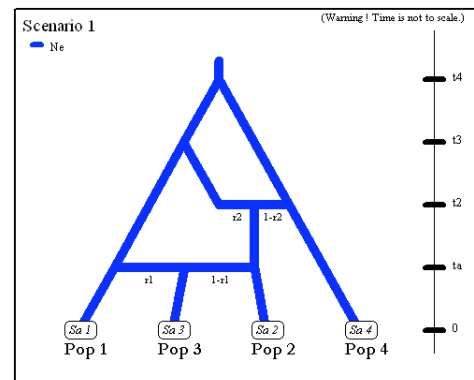
Note that although there are only four samples, the scenario includes a fifth population. This population which diverged from population 1 at time `t3` was a parent in the admixture event occurring at time `t2`. Note also that the first line must include the effective sizes of the *five* populations.

5. The following three scenarii correspond to a classic invasion history from an ancestral population (population 1). In scenario 1, population 3 is derived from population 2, itself derived from population 1. In scenario 2, population 2 derived from population 3, itself derived from population 1. In scenario 3, both populations 2 and 3 derived independently from population 1. The same trio of scenarii will be taken later in a fully described example. Note that when a new population is created

from its ancestral population, there is an initial size reduction (noted here N2b for population 2 and N3b for population 3) since the invasive population generally starts with a few immigrants.

Scenario 1

```
N1 N2 N3
0 sample 1
0 sample 2
0 sample 3
t1-db VarNe 3 N3b
t1 merge 2 3
t2-db VarNe 2 N2b
t2 merge 1 2
```
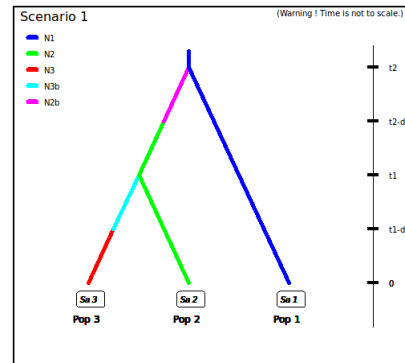
Scenario 2

```
N1 N2 N3
0 sample 1
0 sample 2
0 sample 3
t1-db VarNe 2 N2b
t1 merge 3 2
t2-db VarNe 3 N3b
t2 merge 1 3
```
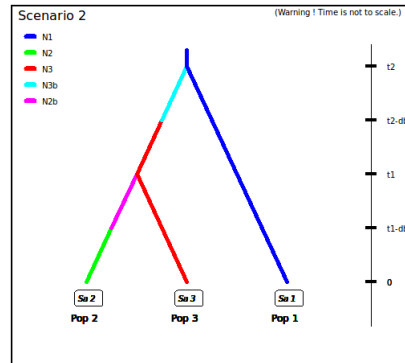
Scenario 3

```
N1 N2 N3
0 sample 1
0 sample 2
0 sample 3
t1-db VarNe 2 N2b
t1 merge 1 2
t2-db VarNe 3 N3b
t2 merge 1 3
```
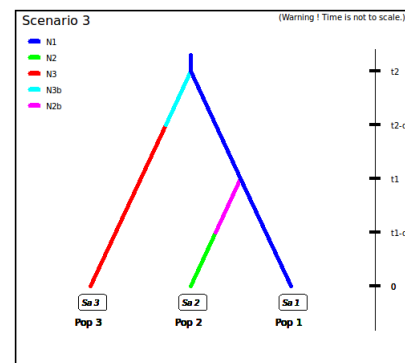
2.3   Mutation model parameterization (microsatellite and DNA sequence loci)

The program can analyse microsatellite data and DNA sequence data altogether as well as separately. In the current version, there are still two restrictions. First, all loci in an analysis must be genetically independent. Second, for DNA sequence loci, intralocus recombination is not considered.

Loci are grouped by the user according to its needs (this an improvement of the current version which imposed all loci of a given category to follow the same mutation model). A different mutation model can be defined for each group. For instance, one group can include all microsatellites with motifs that are 2 bp long and another group those with a 4 bp long motif. Also, with DNA sequence loci, nuclear loci can be grouped together and a mitochondrial locus form a separate group.

The parameterization of the two categories of markers is now described below.

2.3.1   Microsatellite loci

Although a variety of mutation models have been proposed for microsatellite loci (Whittaker *et al.*, 2003), it is usually sufficient to consider only the simplest models (Cornuet *et al.*, 2006). This has the non-negligible advantage of reducing the number of parameters, which can be a real issue when complex scenarios are considered. This is why we chose the Generalized Stepwise Mutation model (Estoup *et al.*, 2002). Under this model, a mutation increases or decreases the length of the microsatellite by a number of repeated motifs following a geometric distribution. This model necessitates only two parameters : the mutation rate (μ) and the parameter of the geometric distribution (P). The same mutation model is imposed to all loci of a given group. However, each locus has its own parameters (μ_i and P_i) and, following a hierarchical scheme, each locus parameter is drawn from a gamma distribution with mean the mean parameter. Note also that :

1. individual loci parameters (μ_i and P_i) are considered as nuisance parameters and hence are never recorded. Only mean parameters are recorded.

2. The variance or shape parameter of the gamma distributions are set by the user and are NOT considered as parameters.

3. The SMM or Stepwise Mutation Model is a special case of the GSM in which the number of repeats involved in a mutation is always one. Such a model can be easily achieved by setting the maximum value of mean P ($\bar{P}$) to 0. In this case, all loci have their P_i set equal to 0 whatever the shape of the gamma distribution.

4. All loci can be given the same value of a parameter by setting the shape of the corresponding gamma distribution to 0 (this is NOT a limiting case of the gamma, but only a way of telling the program).

Eventually, to give more flexibility to the mutation model, the program offers the possibility to consider mutations that insert or delete a single nucleotide to the microsatellite sequence. In the previous version, this option was considered as marginal, and was not treated in the same way as the motif size stepwise mutational process, i.e. there was no associated parameter that could be adjusted to the data. This has been changed in this version : it is now possible to use a mean parameter (named $\mu_{(SNI)}$) with a prior to be defined and individual loci having either values identical to the mean parameter or drawn from a Gamma distribution.

2.3.2   DNA sequence loci

Note first that this version of the program does not consider insertion-deletion mutations, mainly because there does not seem to be much consensus on this topic. Concerning substitutions, only the simplest models are considered. We chose the Jukes-Cantor (1969) one parameter model, the Kimura (1980) two parameter model, the Hasegawa-Kishino-Yano (1985) and the Tamura-Nei (1993) models. The last two models include the ratios of each nucleotide as parameters. However, in order to reduce the number of parameters, these ratios have been fixed to the values calculated from the observed data set for each DNA sequence locus. Consequently, this leaves two and three parameters for the Hasegawa-Kishino-Yano (HKY) and Tamura-Nei (TN), respectively. Also, two adjustments are possible : one can fix the fraction of constant sites (those that cannot mutate) on the one hand and the shape of the Gamma distribution

of mutations among sites on the other hand.

As for microsatellites, all sequence loci of the same group are given the same mutation model with mean parameter(s) drawn from priors and each locus has its own parameter(s) drawn from a Gamma distribution (same hierarchical scheme). Notes 1, 2 and 4 of previous subsection (2.3.1) apply also for sequence loci.

2.4   SNPs do not require mutation model parameterization

SNPs have two characteristics that allow to get rid of mutation models : they are polymorphic and they present only two allelic (ancestral and derived) states. In order to be sure that all analyzed SNP loci have the two characteristics, non polymorphic loci are disgarded right from the beginning of analyses. Consequently, no matter *how* it occurred, we can assume that there occured one and only one mutation in the coalescence tree of sampled genes. We will see below that this largely simplifies (and speeds up) SNP data simulation. Also, this advantageously reduces the dimension of the parameter space (no mutation parameters which are often considered as nuisance parameters). There is however a potential drawback which is the absence of any calibration generally brought by priors on mutation parameters : only (time/effective size) ratios will be informative.

2.5   Prior distributions

The Bayesian aspect of the ABC approach implies that parameter estimations are based on prior knowledge about these parameters. This translates into prior distributions of parameters. The program offers a choice among usual probability distributions, i.e. Uniform, Log-Uniform, Normal or Log-Normal for historical parameters and Uniform, Log-Uniform or Gamma for mutation parameters. Extremum values and other parameters (e. g. mean and standard deviation) must be filled in by the user.

In addition, one can impose some simple conditions on historical parameters. For instance, there can be two times parameters with overlapping prior distributions. However, we want that the first one, say `t1`, to always be larger than the second one , say `t2`. For that, we just need to set `t1 > t2` in the corresponding edit-windows. Such a condition needs to be between two parameters (not a parameter and a number, though this can be set up by giving a minimum and a maximum to the prior distribution) and more precisely between two parameters of the same category (i.e. two effective sizes, two times or two admixture rates). The limit to the number of conditions is imposed by the logics, not by the program. The only binary relationships accepted here are $>, <, >=$ *and* $<=$.

2.6   Algorithms for data simulation : main features

Data simulation is based on the Wright-Fisher model. It consists in generating the genealogy of all sampled genes until their most recent common ancestor using coalescence theory.

This begins by randomly drawing a complete set of parameters from their own prior distributions and that satisfy all imposed conditions. Then, once events have been ordered by increasing times, a sequence of *actions* is constructed. If there are more than one locus, the same sequence of actions is used for all successive loci. Possible *actions* fall into four categories :

adding a sample to a population :
   Add as many gene lineages to the population as there are genes in the sample.

merge two populations :
   Move the lineages of the second population into the first population.

split between two populations :
   Distribute the lineages of the admixed populations between the two parental population according to the admixture rate.

coalesce and mutate lineages within a population :
   There are two possibilities here, depending on whether the population is *terminal* or not. We call *terminal* the population including the most recent common ancestor of the whole genealogy. In a terminal population, coalescences and mutations stop when the MRCA is reached whereas in a non terminal population, coalescence and mutations stop when the upper (most ancient) limit is reached. In the latter case, coalescences can stop before the upper limit is reached because there remains a single lineage, but this single remaining lineage can still mutate.

Two different algorithms are implemented : a generation by generation simulation or a continuous time simulation. The choice, automatically performed by the program, is based on an empirical criterion which ensures that the (approximate[3]) continuous time algorithm is chosen whenever it is faster than the (exact[3]) generation by generation while keeping the relative error on the coalescence rate below 5% (see Cornuet *et al.* (2008) for a description of this criterion).

In any case, a coalescent tree is generated over all sampled genes.

Then the simulation process diverges depending on the type of markers : for microsatellite or DNA sequence loci, mutations are distributed over the branches according to a Poisson process whereas for SNP loci, one mutation is applied to a single branch of the coalescent tree, this branch being drawn at random with probability proportional to its length.

Eventually, starting from an ancestral allelic state (established as explained below), all allelic states of the genealogy are deduced forward in time according to the mutation process. For microsatellite loci, the ancestral allelic state is taken at random in the stationary distribution of the mutation model (not considering potential single nucleotide indel mutations). For DNA sequence loci, the procedure is slightly more complicated. First, the total number of mutations over the entire tree is evaluated. Then according to the proportion of constant sites and the gamma distribution of individual site mutation rates, the number and position of mutated sites are generated. Finally, these mutated sites are given 'A', 'T', 'G' or 'C' states according to the selected mutation model. For SNP loci, the ancestral allelic state is arbitrarily set to 0 and it becomes equal to 1 after le the mutation.

Each category of loci has its own coalescence rate deduced from male and female effective population sizes . In order to combine different categories (e.g. autosomal and mitochondrial), we have to take into account the relationships among the corresponding effective population sizes. This can be achieved by linking the different effective population sizes to the effective number of males (N_M) and females (N_F) through the sum $N_T = N_F + N_M$ and the ratio $r = N_M/(N_F + N_M)$. We use the following formulae for the probability of coalescence of two lineages within this population :

autosomal diploid loci : $p = \frac{1}{8r(1-r)N_T}$

autosomal haploid loci : $p = \frac{1}{4r(1-r)N_T}$

X-linked loci / haplo-diploid loci : $p = \frac{1+r}{9r(1-r)N_T}$

Y-linked loci : $p = \frac{1}{rN_T}$

Mitochondrial loci : $p = \frac{1}{(1-r)N_T}$

Users have to provide a (total) effective size N_T (on which inferences will be made) and a sex-ratio r. If no sex ratio is provided, the default value of r is taken as 0.5.

2.7 Summary statistics

For each category (microsatellite, DNA sequences or SNP) of loci, the program proposes a series of summary statistics among those used by population geneticists. These summary statistics are mean values or variances over loci of the same group and characterize a single, a pair or a trio of population samples. These are :

2.7.1 for microsatellite loci

Single sample statistics :

1. mean number of alleles across loci

2. mean gene diversity across loci (Nei, 1987)

3. mean allele size variance across loci

4. mean M index across loci (Garza and Williamson, 2001; Excoffier *et al.*, 2005)

Two sample statistics :

1. F_{ST} between two samples (Weir and Cockerham, 1984)

---

[3]The terms *approximate* and *exact* are relative to the basic assumptions of the Wright-Fisher model, not to the biological reality of the process.

2. mean index of classification (two samples) (Rannala and Moutain, 1997; Pascual *et al.*, 2007)

3. $(\delta\mu)^2$ distance between two samples (Golstein *et al.*, 1995)

4. mean number of alleles across loci (two samples)

5. mean gene diversity across loci (two samples)

6. mean allele size variance across loci (two samples)

7. shared allele distance between two samples (Chakraborty and Jin, 1993)

Three sample statistics :

1. Maximum likelihood coefficient of admixture (Choisy *et al.*, 2004)

2.7.2 for DNA sequence loci

Single sample statistics :

1. number of distinct haplotypes

2. number of segregating sites

3. mean pairwise difference

4. variance of the number of pairwise differences

5. Tajima's D statistics (Tajima, 1989)

6. Number of private segregating sites (=number of segregating sites if there is only one sample)

7. Mean of the numbers of the rarest nucleotide at segregating sites[4]

8. Variance of the numbers of the rarest nucleotide at segregating sites

Two sample statistics :

1. number of distinct haplotypes in the pooled sample

2. number of segregating sites in the pooled sample

3. mean of within sample pairwise differences

4. mean of between sample pairwise differences

5. F_{ST} between two samples (Hudson *et al.*, 1992)

Three sample statistics :

1. Maximum likelihood coefficient of admixture (adapted from Choisy *et al.*, 2004)

2.7.3 for SNP loci

Single sample statistics :

1. proportion of loci with null gene diverty (= proportion of monomorphic loci)

2. mean gene diversity across polymorphic loci (Nei, 1987)

3. variance of gene diversity across polymorphic loci

4. mean gene diversity across all loci (Garza and Williamson, 2001; Excoffier *et al.*, 2005)

Two sample statistics :

1. proportion of loci with null Nei's distance between the two samples (Nei, 1972)

2. mean across loci of non null Nei's distances between the two samples

3. variance across loci of non null Nei's distances between the two samples

4. mean across loci of Nei's distances between the two samples

---

[4]This statistics can provide information in case of recent demographic variation : a recent expansion increases the number of singletons (nucleotides occuring just once at a segregating site) resulting in a low value of this statistics, whereas a recent decline will produce an opposite result.

5. proportion of loci with null F_{ST} distance between the two samples (Weir and Cockerham, 1984)

6. mean across loci of non null F_{ST} distances between the two samples

7. variance across loci of non null F_{ST} distances between the two samples

8. mean across loci of F_{ST} distances between the two samples

Three sample statistics :

1. Maximum likelihood coefficient of admixture (Choisy *et al.*, 2004)

2.8   Pre-evaluation of scenarios and prior distributions

This option is proposed to users since last version. The purpose is to check that at least one combination of scenarios and priors can produce simulated data sets that are close enough to the observed data set. This is performed through two kinds of analyses. In the first analysis, a principal component analysis is performed in the space of summary statistics on at most 100,000 simulated data set and the observed data is added on each plane of the analysis in order to evaluate how the latter is surrounded by the simulated data sets. In addition to this global approach, there is a second one in which each summary statistic of the observed data set is ranked against those of the simulated data set. This second analysis helps finding which aspects of the model (including prior) have been mistated. For instance, a grossly overestimated genetic distance (in simulated data sets compared to the observed one) may suggest a mispecification of the prior distribution of the time of divergence of the two involved populations or of the mean mutation rate of the markers. Using this new option before running a full ABC treatment is a convenient way to reveal mispecification of models (scenarios) and/or prior distributions of parameters (see Cornuet *et al.*, 2010, for an illustration)

2.9   Estimation of posterior distributions of parameters

Several steps are necessary to get posterior distributions of parameters. First, the normalized Euclidian distance between the observed data set and each simulated data set is computed as the sum of squared differences of summary statistics weighted by the inverse of their variance in the entire set of simulated data. For the *i*-th data set, the distance is :

$$d_i = \sqrt{\sum_{j=1}^{nstat} \frac{(s_{ij} - s_j^{obs})^2}{V_j}} \qquad (1)$$

in which s_{ij} is the *j*-th summary statistics from the *i*-th data set, s_j^{obs} is the *j*-th summary statistics from the *obs*erved data set and V_j is the variance of the the *j*-th summary statistics across all simulated data sets. Only the closest data sets are selected for further treatments. The latter includes a local linear regression step aimed at improving the posterior distributions of the parameters (Beaumont *et al.*, 2002). Basically, a multiple linear regression is performed in which summary statistics are the independent variables and parameters the dependent variables. But this regression is also *local* in the sense that more weight in the regression is given to data sets that are closest to the observed data set. This is performed by using a kernel function (the Epanechnikov kernel following Beaumont *et al.* (2002) :

$$K_\delta(d) = \begin{cases} (1.5/\delta)(1 - (d/\delta)^2), & t \le \delta \\ 0, & t > \delta \end{cases} \qquad (2)$$

Eventually, parameters are adjusted through this process as :

$$\phi_{ik}^* = \phi_{ik} - (\mathbf{s}_i - \mathbf{s}^{obs})\boldsymbol{\beta}_k \qquad (3)$$

in which ϕ_{ik} is the *k*-th parameter of the *i*-th selected data set, ϕ_{ik}^* is the adjusted corresponding parameter, $\mathbf{s}_i$ is the row vector of summary statistics of the *i*-th selected data set, $\mathbf{s}^{obs}$ is the row vector of summary statistics of the observed data set and $\boldsymbol{\beta}_k$ is the transposed *k*-th row vector of the regression coefficient matrix.

The adjusted ϕ_{ik}^* of the selected data sets are an approximate sample of the posterior distribution of parameters (Beaumont *et al.*, 2002).

2.10   Model checking

Checking the model is crucial to statistical analysis (p161 in Gelman *et al.*, 1995). Model checking (i.e. the assessment of the goodness-of-fit of a model parameter posterior combination) is a facet of ABC analysis that has been so far neglected (but see Ingvarsson, 2008). Following Gelman et al. (1995; pp 159-163), we already implemented this option in *DIYABCv*1.0, to measure the discrepancy between a model parameter posterior combination and a real data set by considering various sets of test quantities. These test quantities can be chosen among the large set of ABC summary statistics proposed in the program. This option is based on the same kinds of analysis as section 2.7. The main difference is the set of simulated data. Whereas in section 2.7, prior distributions of parameters have been used to simulate data sets, here we use posterior distributions of the same parameters, hence simulating data from the *posterior predictive distribution.*

The first analysis is a principal component analysis in the space of summary statistics using data sets simulated with the **prior** distributions of parameters (exactly as in section 2.7) and the observed data as well as **data sets from the posterior predictive distribution** are represented on each plane of the PCA. If the model fits well the data, one should see on each PCA plane a wide cloud of data sets simulated from the prior, with the observed data set in the middle of a small cluster of datasets from the posterior predictive distribution.

In the second analysis, each summary statistics of the observed data set is ranked against the distribution of the corresponding summary statistics from the posterior predictive distribution. Summary statistics play here the role of *test statistics* (p169 in Gelman *et al.*, 1995).

Since summary statistics are generally not sufficient, it is advised to use different sets of summary statistics to compute the posterior distribution of parameters on one hand and to check the model on the other hand (see Cornuet *et al.*, 2010). This has been implemented in *DIYABC*.

2.11   Measures of performances

As stressed in previous studies (e.g. Excoffier *et al.*, 2005), the ABC appproach provides an efficient way of assessing its own performances for estimating posterior distributions of parameters. The reference table, the building of which represents generally 95 to 99% of the computing time, can be reused with pseudo-observed (test) data sets which in fact have been obtained through simulation with known values of parameters. It is then rather quick and easy to evaluate the performance of the method for parameter estimation by computing statistics such as estimation biases or mean square errors.

These measures of performance have been fully integrated into *DIYABC*. The performance measures computed by *DIYABC* are :

the average relative bias : the difference between the point estimate (e) and the true value (v)divided by the true value, $\frac{1}{n} \sum_{i=1}^{n} \frac{e_i - v_i}{v_i}$, averaged over the n test data sets,

the square Root of the Relative Mean Square Error (RRMSE) : the square root of the average square difference between the point estimate and the true value, divided by the true value, $\sqrt{\frac{1}{n} \sum_{i=1}^{n} (\frac{e_i - v_i}{v_i})^2}$

the square Root of the Relative Mean Integrated Square Error (RRMISE) : the square root of the average (over test data sets) of the integrated square error (measured on each test data set) divided by the true value, $\sqrt{\frac{1}{n} \sum_{i=1}^{n} (\frac{\sum_{j=1}^{m_i} (x_{ij} - v_i)^2}{m_i v_i^2})}$, x_{ij} and m_i being the sampled values and the sample size of the posterior distribution in the i-th test data set, respectively.

the Relative Mean Absolute Deviation (RMAD) : the average (over test data sets) of the mean absolute deviation (measured on each data set), divided by the true value, $\frac{1}{n} \sum_{i=1}^{n} (\frac{\sum_{j=1}^{m_i} |x_{ij} - v_i|}{m_i |v_i|}$

the 50% and 95% coverages : the proportion of test data sets for which the 50% and 95% credibility intervals respectively include the true value.

the factor 2 :the proportion of test data sets for which the point estimate is at least half and at most twice the true value.

the Relative Median Bias (RMB) : the 50% quantile of the bias (measured on each data set) divided by the true value. The bias is computed respectively for each point estimate

the Relative Median Absolute Deviation (RMedAD) : the 50% quantile (over test data sets) of the median (over each data set) of the absolute difference between each value of the posterior distribution sample and the true value divided by the true value.

the Relative Median of the Absolute Error (RMAE) : the 50% quantile (over test data sets) of the absolute value of the difference between the point estimate (in each data set) and the true value divided by the true value.

DIY ABC considers the following three point estimates : mean, median and mode of the ϕ_{ik}^* (sample of the posterior distribution of each parameter), as defined in subsection 1.7.

Concerning the true value (v) appearing in the above formulae, DIYABC offers two possibilities :

1. All values v are fixed by the user. If any one of these values is outside the limits given to the prior for the corresponding parameter, a warning message is issued but the analysis can proceed if needed.

2. All values v are drawn from distributions. These distributions can be different from those of priors. They may even not be overlapping (no warning message is issued whatever the user's choice).

If you want to fix some parameter values and draw the other from distributions, choose the second option and give the same desired values as minimum and maximum for those fixed parameter values.

2.12   Comparison of scenarios

The ABC approach can also be used to compare possible scenarios for the same data file through the computation of the posterior probabilities of each scenario and this option is naturally implemented in DIYABC.

2.12.1   Reference table

First, the reference table can include as many scenarios as desired. By default, the prior probability of each scenario is uniform, that is each scenario will have approximately the same number of simulated data sets. But, if for any reason, one wants a different prior probability for each scenario, there is the possibility to do so.

Scenarios are drawn according to their own prior probability and then only parameters that are defined for the drawn scenario are generated from their respective prior distribution. Scenarios may or may not share parameters.

When conditions apply to some parameters (see subsection 2.4), the program provides the possibility of choosing between two options :

1. parameter sets are drawn in their respective prior distributions until all conditions are fulfilled.

2. a single parameter set is drawn and only if all condition are fulfilled, the simulation is performed and the data set is recorded in the reference table.

When there is only one scenario, both options are equivalent, although in the latter option, there might be less simulated data sets that are recorded than one asked. When there is more than one scenario, the second option can be viewed as a way to set prior probabilities on scenario that result from imposed conditions on parameters (see Miller *et al.* (2005) for an example).

2.12.2 Posterior probability of scenarios

The program *DIY ABC* provides two estimates of the posterior probability of each scenario :

a *direct* estimate : This is simply the number of times that a given scenario is found in the first n_δ simulated data sets once the latter, produced under several scenarios, have been sorted by ascending distances to the observed data set.

a logistic regression estimate : Following M.A. Beaumont's suggestion (Fagundes *et al.*, 2007; Beaumont, 2008), a polychotomic weighted logistic regression is performed on the first n_δ data sets with the proportion of the scenario as the dependent variable and the differences between observed and simulated data set summary statistics as the independent variables. The intercept of the regression (corresponding to an identity between simulated and observed summary statistics) is taken as the point estimate. In addition, 95% confidence intervals are computed (Cornuet *et al.*, 2008).

Since both estimates are dependent upon the chosen threshold (δ), the program provides a range of 100 estimates for the direct approach (for each one 100-th of n_δ between 0 and n_δ) and up to 10 estimates for the logistic regression estimates (e.g. one estimate for $kn_\delta/10$ with $k \in [1, 2, ...10]$ when the number of analyses is set to 10). These estimates are represented in two graphs, one for each kind of estimate. These two graphs can be printed and/or saved (in *svg*, *jpg* or *png* format). Values can also be output as a text file.

2.12.3 Confidence in scenario choice

The program DIYABC offers a last option that allows one to evaluate the confidence in a scenario choice. Suppose that we compare 3 scenarios for a given data set and that e.g. scenario 2 had maximum posterior probability. By using this option, we can estimate type I and type II errors when choosing scenario 2 as the true scenario. To do so, we simulate a given number of data sets according to scenario 1, 2 and 3. Then we count the proportion of times that scenario 2 has not the highest posterior probability among the three competing scenarios when it is the true scenario (type I error, estimated from test data sets simulated under scenario 2) or the proportion of times that scenario 2 has highest posterior probability when it *not* the true scenario (type II error, estimated from test data sets simulated under scenarios 1 and 3).

In *DIY ABCv*2.0, a new possibility is offered to the user that may be useful when dealing with many summary statistics and many scenarios. In this particular case, the logistic regression has to deal with large matrices and the amount of needed memory on one hand and the computation time on the other hand can become problematically large. An approximate solution is to replace summary statistics by the components of a factorial discriminant analysis which reduces the number of independent variables to the smallest of number of summary statistics and scenarios. Although the result is only approximate, it can be a useful guide in some specific cases. The gain in time can be large. For instance, the time can be reduced by a 30X factor.

As for the bias/precision analysis, parameter values can be fixed to given values or drawn from given distributions (not necessarily the same as those used as priors for the reference table).

Bibliography

Beaumont, M. A., W. Zhang and D. J. Balding, 2002. Approximate Bayesian Computation in Population Genetics. *Genetics* **162**, 2025-2035.

Beaumont, M.A., 2008. Joint determination of topology, divergence time, and immigration in population trees. In Simulation, Genetics, and Human Prehistory, eds. S. Matsumura, P. Forster, C. Renfrew. McDonald Institute Press, University of Cambridge (*in press*).

Begg, C.B. and R. Gray, 1984. Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika*, **71**, 11-18.

Belkhir K., Borsa P., Chikhi L., Raufaste N. and F. Bonhomme, 1996-2004 GENETIX 4.05, logiciel sous Windows TM pour la gntique des populations. Laboratoire Gnome, Populations, Interactions, CNRS UMR 5171, Universit de Montpellier II, Montpellier (France).

Bertorelle, G. and L. Excoffier, 1998. Inferring admixture proportion from molecular data. *Mol. Biol. Evol.* **15**, 1298-1311.

Choisy, M., P. Franck and J.M. Cornuet, 2004. Estimating admixture proportions with microsatellites : comparison of methods based on simulated data. *Mol. Ecol.* **13**, 955-968.

Chakraborty R and L Jin, 1993. A unified approach to study hypervariable polymorphisms: statistical considerations of determining relatedness and population distances. EXS. *67*, 153175.

Cornuet, J. M., M. A. Beaumont, A. Estoup and M. Solignac, 2006. Inference on microsatellite mutation processes in the invasive mite, *Varroa destructor*, using reversible jump Markov chain Monte Carlo. *Theoret. Pop. Biol.* **69**, 129-144.

Cornuet J.M., V. Ravigné and A. Estoup, 2010. Inference on population history and model checking using DNA sequence and microsatellite data with the sofware DIYABC (v1.0). *submitted*.

Cornuet J.M., F. Santos, M.A. Beaumont, C.P. Robert, J.M. Marin, D.J. Balding, T. Guillemaud and A. Estoup, 2008. Infering population history with DIYABC: a user-friendly approach to Approximate Bayesian Computations. *Bioinformatics*, **24** (23), 2713-2719.

Estoup, A., M. Solignac, M. Harry and J.M. Cornuet, 1993. Characterization of $(GT)_n$ and $(CT)_n$ microsatellites in two insect species: *Apis mellifera* and *Bombus terrestris*. *Nucl. Ac. Res.*, **21**, 1427-1431.

Estoup, A., I. J. Wilson, C. Sullivan, J. M. Cornuet and C. Moritz, 2001 Inferring population history from microsatellite et enzyme data in serially introduced cane toads, *Bufo marinus. Genetics*, **159**, 1671-1687.

Estoup, A., P. Jarne and J.M. Cornuet, 2002. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Mol. Ecol.*, **11**, 1591-1604.

Estoup, A. and S. M. Clegg, 2003. Bayesian inferences on the recent islet colonization history by the bird Zosterops lateralis lateralis. *Mol. Ecol.* **12**: 657-674.

Estoup, A., M.A. Beaumont, F. Sennedot, C. Moritz and J.M. Cornuet, 2004. Genetic analysis of complex demographic scenarios : spatially expanding populations of the cane toad, *Bufo marinus. Evolution*,**58**, 2021-2036.

Excoffier, L., A. Estoup and J.M. Cornuet, 2005. Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics* **169**, 1727-1738.

FAGUNDES, N.J.R., N. RAY, M.A. BEAUMONT, S. NEUENSCHWANDER, F. SALZANO, S.L. BONATTO AND L. EXCOFFIER, 2007. Statistical evaluation of alternative models of human evolution. *Proc. Natl. Acad. Sc.*, **104** : 17614-17619.

Fu, Y.X. and Chakraborty, R., 1998. Simultaneous estimation of all the parameters of a stepwise mutation model. *Genetics*, **150**, 487-497.

Garza JC and E Williamson, 2001. Detection of reduction in population size using data from microsatellite DNA. *Mol. Ecol.* **10**,305-318.

Gelman, A., J.B. Carlin, H.S. Stern and D.B. Rubin, 1995. *Bayesian Data Analysis*. Chapman et Hall, London, 526p.

Goldstein DB, Linares AR, Cavalli-Sforza LL, and Feldman MW, 1995. An evaluation of genetic distances for use with microsatellite loci. *Genetics* **139**, 463-471.

Goudet, J. ,1995. FSTAT (Version 1.2): A computer program to calculate F- statistics. *J. Hered.* **86**, 485-486.

Griffiths, R.C. and S. Tavaré, 1994. Simulating probability distributions in the coalescent. *Theor. Pop. Biol.* **46**, 131-159.

Guillemaud T., M.A. Beaumont, M. Ciosi, J.M. Cornuet and A. Estoup, 2010. Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data. *Heredity*, **104**, 88-99.

Haag-Liautard C., N. Coffey, D. Houle, M. Lynch, B. Charlesworth and P.D. Keightley, 2008. Direct estimation of the mitochondrial DNA mutation rate in Drosophila melanogaster. *Plos Biol*, **6**, e204.

Hamilton, G., M. Stoneking and L. Excoffier, 2005. Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. *Proc. Natl. Acad. Sci. USA*, **102**, 7476-7480.

Hasegawa, M., Kishino, H and Yano, T., 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22:160-174.

Hudson,R. R., M. Slatkin and W.P. Maddison, 1992. Estimation of levels of gene flow fom DNA sequence data. *Genetics*, 132, 583-589.

Ihaka R. and R. Gentleman, 1996. *R*: a language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5**, 299-314

Ingvarsson P.K., 2008. Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula. Genetics*, 180: 329-340.

Jukes, TH and Cantor, CR., 1969. Evolution of protein molecules. Pp. 21-123 in H. N. Munro, ed. *Mammalian protein metabolism.* Academic Press, New York.

Kimura, M., 1980. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16:111-120.

Lombaert E., T. Guillemaud, J.M. Cornuet, T. Malausa, B. Facon and A. Estoup, 2010. Bridgehead effect in the worldwide invasion of the biocontrol harlequin ladybird. *PLoS ONE*, http://dx.plos.org/10.1371/journal.pone.0009743.

Miller N, A. Estoup, S. Toepfer, D Bourguet, L. Lapchin, S. Derridj, K.S. Kim, P Reynaud, F. Furlan and T. Guillemaud, 2005. Multiple Transatlantic Introductions of the Western Corn Rootworm. *Science*, **310**, p. 992

Nei M., 1972. Genetic distance between populations. *Am. Nat.* 106:283-292

Nei M., 1987. *Molecular Evolutionary Genetics.* Columbia University Press, New York, 512 pp.

5 Ohta, T. and Kimura, M., 1973. A model of mutation appropriate to estimate the number of elec-
6   trophoretically detectable alleles in a finite population.

7 Pascual, M., M.P. Chapuis, F. Mestres, J. Balanyá, R.B. Huey, G.W. Gilchrist, L. Serra and A. Estoup,
8   2007. Introduction history of *Drosophila subobscura* in the New World : a microsatellite based survey
9   using ABC methods. *Mol. Ecol.*, **16**, 3069-3083.

10 Pritchard, J., M. Seielstad, A. Perez-Lezaun and M. Feldman, 1999. Population growth of human Y
11   chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.* **16**, 1791-1798.

12 Rannala, B., and J. L. Mountain, 1997. Detecting immigration by using multilocus genotypes. *Pro. Nat.*
13   *Acad. Sci. USA* **94**, 9197-9201.

14 Raymond M., and F. Rousset, 1995. Genepop (version 1.2), population genetics software for exact tests
15   and ecumenicism. *J. Hered.*, **86**, 248-249

16 Stephens, M. and P. Donnelly, 2000, Inference in molecular population genetics (with discussion). J. R.
17   Stat. Soc. B **62**, 605-655.

18 Tajima, F., 1989. Statistical method for testing the neutral mutationhypothesis by DNA polymorphism.
19   *Genetics* 123: 585-595

20 Tamura, K., and M. Nei., 1993. Estimation of the number of nucleotide substitutions in the control region
21   of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* 10:512-526.

741 Weir BS and CC Cockerham , 1984. Estimating F-statistics for the analysis of population structure.
742   *Evolution* **38**: 1358-1370.

743 Whittaker, J.C., Harbord, R.M., Boxall, N., Mackay, I., Dawson, G. and Sibly, R.M. 2003. Likelihood-
744   based estimation of microsatellite mutation rates, *Genetics*, **164**, 781-787.