BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 23 2008, pages 2713–2719
doi:10.1093/bioinformatics/btn514

Genetics and population analysis

Inferring population history with *DIY ABC*: a user-friendly approach to approximate Bayesian computation

Jean-Marie Cornuet[1,2,*], Filipe Santos[2], Mark A. Beaumont[3], Christian P. Robert[4], Jean-Michel Marin[5], David J. Balding[1], Thomas Guillemaud[6] and Arnaud Estoup[2]

[1]Department of Epidemiology and Public Health, Imperial College, St Mary's Campus, Norfolk Place, London W2 1PG, UK, [2]Centre de Biologie et de Gestion des Populations, INRA, Campus International de Baillarguet, CS 30016 34988 Montferrier-sur-Lez, France, [3]School of Biological Sciences, Lyle Building, The University of Reading Whiteknights, Reading RG6 6AS, UK, [4]CEREMADE, Université Paris-Dauphine, Place Delattre de Tassigny, 75775 Paris cedex 16, [5]INRIA Saclay, Projet select, Université Paris-Sud, Laboratoire de Mathématiques (Bât. 425), 91400 Orsay and [6]UMR 1301 I.B.S.V. INRA-UNSA-CNRS, 400 Route des Chappes, BP 167 - 06903 Sophia Antipolis cedex. France

Received on June 17, 2008; revised on September 5, 2008; accepted on October 2, 2008

Advance Access publication October 7, 2008

Associate Editor: Martin Bishop

ABSTRACT

Summary: Genetic data obtained on population samples convey information about their evolutionary history. Inference methods can extract part of this information but they require sophisticated statistical techniques that have been made available to the biologist community (through computer programs) only for simple and standard situations typically involving a small number of samples. We propose here a computer program (*DIY ABC*) for inference based on approximate Bayesian computation (ABC), in which scenarios can be customized by the user to fit many complex situations involving any number of populations and samples. Such scenarios involve any combination of population divergences, admixtures and population size changes. *DIY ABC* can be used to compare competing scenarios, estimate parameters for one or more scenarios and compute bias and precision measures for a given scenario and known values of parameters (the current version applies to unlinked microsatellite data). This article describes key methods used in the program and provides its main features. The analysis of one simulated and one real dataset, both with complex evolutionary scenarios, illustrates the main possibilities of *DIY ABC*.

Availability: The software *DIY ABC* is freely available at http://www.montpellier.inra.fr/CBGP/diyabc.

Contact: j.cornuet@imperial.ac.uk

Supplementary information: Supplementary data are also available at http://www.montpellier.inra.fr/CBGP/diyabc

1 INTRODUCTION

Until now, most literature and software about inference in population genetics concern simple standard evolutionary scenarios: a single population (Beaumont, 1999; Griffiths and Tavaré, 1994; Stephens and Donnelly, 2000), two populations exchanging genes (De Iorio and Griffiths, 2004; Hey and Nielsen, 2004) or not (Hickerson *et al.*, 2007) or three populations in the classic admixture scheme (Excoffier *et al.*, 2005; Wang, 2003). The main exception to our knowledge is the computer program *BATWING* (Wilson *et al.*, 2003) which considers a whole family of scenarios in which an ancestral population splits into as many subpopulations as needed. However, in practice, population geneticists collect and analyse samples which rarely correspond to one of these standard scenarios. If they want to apply methods developed in the literature and for which computer programs are available, they have to select subsets of samples (to fit these standard situations), at the price of lowering the power of the analysis. The other solution is to develop their own software, which requires specific skills or the right collaborators. Rare examples of inference in non-standard scenarios can be found in O'Ryan *et al.* (1998) including three populations and two successive divergences, or Estoup *et al.* (2004) (10 populations that sequentially diverged with initial bottlenecks and exchanging migrants with neighbouring populations).

Inference in complex evolutionary scenarios can be performed in various ways, but all are based on the genealogical tree of sampled genes and coalescence theory. A first approach used in programs such as *IM* (Hey and Nielsen, 2004) or *BATWING* consists of starting from a gene genealogy compatible with the observed data and exploring the parameter and genealogy space through MCMC algorithms. One difficulty with this approach is to be sure that the MCMC has converged, because of the huge dimension of the parameter space. With a complex scenario, the difficulty is increased. Also, although not impossible, it seems quite challenging to write a program that would deal with very different scenarios. A second approach pioneered by Beaumont (2003) consists in combining MCMC exploration of the scenario parameter space with an importance sampling (IS)-based estimation of the likelihood. The strength of this approach is that the low number of parameters ensures a (relatively) fast convergence of the MCMC. Its weakness is that the likelihood is only approximated through IS, sometimes resulting in poor acceptance rates.

When dealing with complex situations, the two previous approaches raise difficulties which mainly stem in the computation

[*]To whom correspondence should be addressed.

of the likelihood. Consequently, a line of research including the works of Tavaré *et al.* (1997), Weiss and von Haeseler (1998), Pritchard *et al.* (1999) and Marjoram *et al.* (2003) developed a new approach termed approximate Bayesian computation (or ABC) by Beaumont *et al.* (2002). In this approach, the likelihood criterion is replaced by a similarity criterion between simulated and observed datasets, similarity usually measured by a distance between summary statistics computed on both datasets. Among examples of inference in complex scenarios given above, all but one (the simplest) have used this approach, showing that it can indeed solve complex problems.

The ABC approach presents two additional features that can be of interest for experimental biologist. One characteristic, already noted by Excoffier *et al.* (2005), is the possibility to assess the bias and precision of estimates for simulated datasets produced with known values of parameters with little extra computational cost. To get the same information with likelihood-based methods would require a huge amount of additional computation whereas, with ABC, the largest proportion of computation used for estimating parameters can be recycled in a bias/precision analysis. The second feature is the simple way by which the posterior probability of different scenarios applied to the same dataset can be estimated (e.g. Miller *et al.*, 2005; Pascual *et al.*, 2007).

In its current state, the ABC approach remains inaccessible to most biologists because there is not yet a simple software solution. Therefore, we developed the program *DIYABC* that performs ABC analyses on complex scenarios, i.e. which include any number of populations and samples (samples possibly taken at different times), with populations related by divergence and/or admixture events and possibly experiencing changes of population size. The current version is restricted to unlinked microsatellite data. In this article, we describe the rationale for some methods involved in the program. Then we give the main features of *DIYABC* and we provide two complete example analyses performed with this program to illustrate its possibilities.

2 KEY METHODS INVOLVED IN *DIY ABC*

Inference about the posterior distribution of parameters in an ABC analysis is usually performed in three steps (see Figure S1 in Supplementary Material). The first one is a simulation step in which a very large table (the *reference table*) is produced and recorded. Each row corresponds to a simulated dataset and contains the parameter values used to simulate the dataset and summary statistics computed on the simulated dataset. Parameter values are drawn from prior distributions. Using these parameter values, genetic data are simulated as explained in the next section. The summary statistics correspond to those traditionally used by population geneticists to characterize the genetic diversity within and among samples (e.g. number of alleles, genic diversity and genetic distances). The idea is to extract maximum genetic information from the data, admitting that exhaustivity or sufficiency are generally out of reach. The simulation step is generally the most time-consuming step, since the number of simulated datasets can reach several millions. The second step is a rejection step. Euclidian distances between each simulated and the observed dataset in the space of summary statistics are computed and only the simulated data sets closest to the observed dataset are retained. The parameter values used to simulate these selected datasets provide a sample of parameter values approximately distributed according to their own posterior distribution. Beaumont *et al.* (2002) have shown that a local linear regression (third step = estimation step) provides a better approximation of the posterior distribution.

This synoptic of ABC is well established and we now concentrate on more specific issues that are implemented in *DIYABC*.

2.1 Simulating genetic data in complex scenarios

Thanks to coalescence theory, it has become easy to simulate datasets by a two-steps procedure. The first step consists of building a genealogical tree of sampled genes according to rather simple rules provided by coalescence theory (see below). The second step consists of attributing allelic states to all the nodes of the genealogy, starting from the common ancestor and simulating mutations according to the mutation model of the genetic markers. In a complex scenario, only the first step needs special attention and we will concentrate on it now.

In a single isolated population of constant effective size, the genealogical tree of a sample of genes is simulated backward in time: starting from the time of sampling, the gene lineages are merged (coalesced) at times that are drawn from an exponential distribution with rate $j(j-1)/4N_e$, when there are j distinct lineages and the (diploid) effective population size is N_e. The genealogical tree is completed when there remains a single lineage.

Consider now two isolated populations (effective population sizes N_1 and N_2, respectively) that diverged t_d generations before their common sampling time. Since the two populations do not exchange genes, lineages within each population will coalesce independently. Coalescence simulation will stop either when there remains a single lineage or when the simulated time is beyond the divergence (looking back in time). In the latter case, the coalescence event is simply discarded. At generation t_d, the remaining lineages are simply pooled and will coalesce in the ancestral population. Because of the memoryless property of the exponential distribution, the time to the first coalescence in the ancestral population is independent of the times of the last coalescence in each daughter population and can be simulated as in the single isolated population above. Again, the genealogical tree is completed when there remains a single lineage in the ancestral population. Note that the two populations need not be sampled at the same generation since this has no bearing on the simulation process.

Consider now the classic admixture scenario with one admixed and two parental populations, as in Figure 1 in Excoffier *et al.* (2005). Simulating the complete genealogical tree can be achieved with the following steps: (i) coalesce gene lineages in each population independently until reaching admixture time, (ii) distribute remaining lineages of the admixed population among the two parental populations, each with a Bernoulli draw with probability equal to the admixture rate, (iii) coalesce gene lineages in the two parental populations until reaching their divergence time, (iv) pool the remaining gene lineages of the two parental populations and place them into the ancestral population and (v) coalesce gene lineages in the ancestral population.

We first note the modular form of this algorithm which involves only three modules:

(1) a module that performs coalescences in an isolated constant size population between two given times,

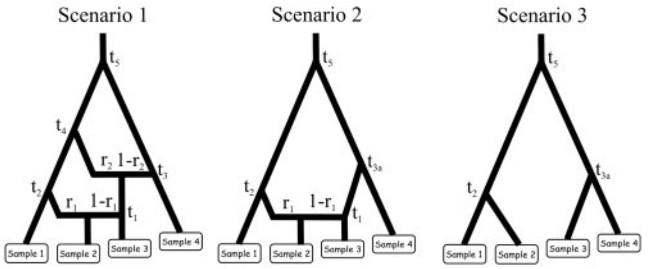

Fig. 1. First example: the three evolutionary scenarios. The dataset used as an example has been simulated according to scenario 1 (left). The parameter values were the following: all populations had an effective (diploid) size of 1000, the times of successive events (backward in time) were $t_1 = 10$, $t_2 = 500$, $t_3 = 10\,000$, $t_4 = 20\,000$ and $t_5 = 200\,000$, the two admixture rates were $r_1 = 0.6$ and $r_2 = 0.4$. Scenario 1 includes six populations, the four that have been sampled and two *left* parental populations in the admixture events. Scenario 2 and 3 include 5 and 4 populations, respectively. Samples 3 and 4 have been collected 2 and 4 generations earlier than the first two samples, hence their slightly upward locations on the graphs. Time is not at scale.

(2) a module that pools gene lineages from two populations (for divergence),

(3) a module that splits gene lineages from the admixed population between two parental populations (for admixture).

We also note that the last two modules are quite simple and that the first one might be extended to include population size variations.

We have introduced a fourth module that proves useful in many instances. It performs the (simple) task of adding a gene sample to a population at a given generation. The interest of this module is to allow for multiple samples of the same population taken at different generations. By combining the aforementioned four modules, it is possible to simulate genetic data involving any number of populations according to a scenario that can include divergence, admixture events as well as population size variations. In addition, populations can be sampled more than once at different times. Compared with our previous definition of complex scenarios, the only restriction so far concerns the absence of migrations among populations. If migrations have to be taken into account, coalescences in two (or more) populations exchanging migrants are no longer independent and should be treated in the same module. Such a module would require to consider simultaneously two kind of events, coalescences of lineages within population and migrations of gene lineages from one population to another. In the current stage of *DIYABC*, this has not yet been achieved.

2.2 Two ways of simulating coalescence events

Simulating coalescences can be performed in two ways. The most traditional way is based on the usually fulfilled assumption that the effective population size is large enough so that the probability of coalescence is small and hence that the probability that two or more coalescences occur at the same generation is low enough so that it can be neglected (e.g. Nordborg, 2007). Time is then considered as a continuous variable in computations. The corresponding algorithm, called here the *continuous time* (CT) algorithm, consists in drawing first times between two successive coalescence events and then drawing two lineages at random at each coalescence event.

However, in practice, population size can be so small (e.g. during a bottleneck) that multiple coalescences at the same generation become common, including with the same parental gene (producing multifurcating trees). Simulating gene genealogies with multiple coalescences is possible, (e.g. Laval and Excoffier, 2004). In effect, lineages are reconstructed one generation at a time: lineages existing at generation g are given a random number drawn in $U[1, 2N_e]$ and lineages with the same number coalesce together. The latter is termed here the *generation by generation* (GbG) algorithm.

The CT algorithm is much faster in most cases and is used in most softwares, but in some circumstances, the approximation becomes unacceptable. The solution taken in *DIYABC* is to swap between the two algorithms according to a criterion based on the effective population size, the time during which the effective size keeps its value, and the number of lineages at the start of the module. The criterion is such that the generation per generation (GbG) algorithm is taken whenever it is faster (this occurs when the effective size is very small) or when the CT algorithm overestimates by more than 5% on average the number of lineages at the end of the module.

A specific comparison study has been performed to establish this criterion. For different time periods counted in number of generations (g), effective population sizes (N_e) and number of entering lineages (n_{el}), coalescences have been simulated according to each algorithm 10 000 times and the average number of remaining lineages at the end of the period have been recorded as well as the average computation duration of each algorithm. Our results (Figure S2) show that the following rules optimize computation speed, while keeping the relative bias in coalescence rates under the 5% threshold:

if $(1 < g \leq 30)$ do CT if $n_{el}/N_e < 0.0031g^2 - 0.053g + 0.7197$ else do GbG

if $(30 < g \leq 100)$ do CT if $n_{el}/N_e < 0.033g + 1.7$ else do GbG

if $(100 < g)$ do CT if $n_{el}/N_e < 5$ else do GbG

2.3 Comparing scenarios

Using ABC to compare different scenarios and infer their posterior probability has been performed in two ways in the literature. Starting with a reference table containing parameters and summary statistics obtained with the different scenarios to be compared (or pooling reference tables, each obtained with a given scenario), datasets are ordered by increasing distance to the observed dataset. A first method (termed hereafter the *direct* approach) is to take as an estimate of the posterior probability of a scenario the proportion of datasets obtained with this scenario in the n_δ closest datasets (Miller *et al.*, 2005; Pascual *et al.*, 2007). The value of n_δ is arbitrary and unless the results are quite clear cut, the estimated posterior probability may vary with n_δ.

Following the same rationale that introduced the local linear regression in the estimation of posterior distributions for parameters (Beaumont *et al.*, 2002), we perform a weighted polychotomous logistic regression to estimate the posterior probability of scenarios, termed hereafter the *logistic* approach (see also Beaumont, 2008; Fagundes *et al.*, 2007). In the estimation of parameters, a linear regression is performed with dependent variable the parameter and predictors the differences between the observed and simulated statistics. This linear regression is local at the point (in the predictor space) corresponding to the observed dataset, using an

Epanechnikov kernel based on the distance between observed and simulated summary statistics [see formula (5) in Beaumont *et al.*, 2002]. Parameters values are then replaced by their estimates at that point in the regression.

Keeping the differences between observed and simulated statistics as the predictor variables in the regression, we consider now the posterior probability of scenarios as the dependent variable. Because of the nature of the 'parameter', an indicator of the scenario number, a *logit* link function is applied to the regression. The local aspect of the regression is obtained by taking the same weights as in the linear adjustment of parameter values as described in Beaumont *et al.* (2002). Confidence intervals for the posterior probabilities of scenarios are computed through the limiting distribution of the maximum likelihood estimators. See Annex 1 in Supplementary Material for a detailed explanation.

2.4 Quantifying confidence in parameter estimations on simulated test datasets

In order to measure bias and precision, we need to simulate datasets (i.e. test datasets) with known values of parameters and compare estimates with their true values. In the ABC estimation procedure, the most time-consuming task is to produce a large enough reference table. However, when such a reference table has been produced, e.g. for the analysis of a real dataset, it can also be used to quantify bias and precision on test datasets as well.

Measuring bias is straightforward, but precision can be assessed with different measures. In *DIYABC*, the latter include the relative square root of the mean square error, the relative square root of the mean integrated square error, the relative mean absolute deviation, the 95% and 50% coverages and the factor 2. See Annex 2 in Supplementary Material for more details.

3 *DIY ABC*: A COMPUTER PROGRAM FOR POPULATION BIOLOGISTS

3.1 Main features

DIYABC is a program that performs ABC inference on population genetic data. In its current state, the data are genotypes at microsatellite loci of samples of diploid individuals (missing data are allowed). The inference bears on the evolutionary history of the sampled populations by quantifying the relative support of data to possible scenarios and by estimating posterior densities of associated parameters. *DIYABC* is a program written in Delphi running under a 32-bit Windows operating system (e.g. Windows XP) and it has a user-friendly graphical interface.

The program accepts complex evolutionary scenarios involving any number of populations and samples. Scenarios can include any number of the following timed events: stepwise change of effective population size, population divergence and admixture. They can also include unsampled as well as serially sampled populations as in Beaumont (2003). The main restriction regarding scenario complexity is the absence of migrations between populations.

Since the program has been written for microsatellite data, it proposes two mutation models, namely the stepwise mutation model (SMM) and the generalized stepwise mutation (GSM) model (Estoup *et al.*, 2002). Note that the same mutation model has to be applied to all microsatellite loci, but these may have different values of mutation parameters.

The historico-demographic parameters of scenarios may be of three types: effective sizes, times of events (in generations) and admixture rates. Marker parameters are mutation rates and the coefficient of the geometric distribution (under the GSM only). The program can also estimate composite parameters, such as $\theta = 4N_e\mu$ and $\tau = t\mu$, with N_e being the diploid effective population size, t the time of an event and μ the mean mutation rate. Prior distributions are defined for original parameters and those for composite parameters are obtained via an assumption of independence of their component prior densities. Priors for historico-demographic parameters can be chosen among four common distributions: Uniform, Log-uniform, Normal and Log-normal. Users can set minimum and maximum (for all distributions) and mean and SD (for Normal and Log-normal). In addition, priors can be modified by setting binary conditions ($>$, $<$, $\geq$ and $\leq$) on pairs of parameters of the same category (two effectives sizes or two times of event). This is especially useful to control the relative times of events when these are parameters of the scenario. For priors of mutation parameters, only the Uniform and the Gamma distributions are considered, but hierarchical schemes are possible, with a mean mutation rate or coefficient P (of the geometric distribution in the GSM) drawn from a given prior and individual loci parameter values drawn from a gamma distribution around the mean.

Available summary statistics are usual population genetic statistics averaged over loci: e.g. mean number of alleles, mean genic diversity, Fst, $(\delta\mu)^2$, admixture rates, etc.

Regarding ABC computations, the program can (i) create a reference table or append values to an existing table, (ii) compute the posterior probability of different scenarios, (iii) estimate the posterior distributions of original and/or composite parameters for one or more scenarios and (iv) compute bias and precision for a given scenario and given values of parameters . Finally, the program can be used simply to simulate datasets in the popular *Genepop* format (Raymond and Rousset, 1995).

3.2 Two examples of analysis with *DIY ABC*

3.2.1 Illustration on a simulated dataset In order to illustrate the capabilities of *DIYABC*, we take first an example based on a dataset simulated according to a complex scenario including three splits and two admixture events (scenario 1 in Figure 1). The scenario includes six populations: two of them have been sampled at time 0, the third one at time 2 and the fourth one at time 4, the last two have not been sampled. Each population sample includes 30 diploid individuals and data are simulated at 10 microsatellite loci. This scenario is not purely theoretical as it could be applied for instance to European populations of honeybees in which the Italian populations (*Apis mellifera ligustica*) result from an ancient admixture between two evolutionary branches (Franck *et al.*, 2000) that would correspond here to population samples 1 and 4. Furthermore, in the last 50 years, Italian bees have been widely exported and sample 2 could well correspond to a population of a parental branch that has been recently introgressed by Italian queens. This example also stresses the ability of *DIYABC* to distinguish two events that are confounded in the usual admixture scheme: the admixture event itself and the time at which the real parental populations in the admixture diverged from the population taken as 'parental'.

Our ABC analysis will address the following questions: (i) Suppose that we are not sure that the scenario having produced our example dataset does include a double admixture and that we want to challenge this double admixture scenario with two simpler scenarios, one with a single admixture (scenario 2 in Figure 1) and the other with no admixture at all (scenario 3). The questions addressed are: (i) What is the posterior probability of these three scenarios, given identical prior probabilities ? (ii) What are the posterior distributions of parameters, given that the right scenario is known ? and (iii) What confidence can we have in the posterior probabilities of scenarios and posterior distributions of parameters?

First, a reference table is built up. Using different screens of *DIYABC*, (i) the three scenarios are coded and prior distributions of parameters are defined (Figure S3), (ii) based on previous studies (e.g. Dib *et al.*, 1996), the GSM model is selected and prior distributions of mutation parameters are defined (Figure S4), (iii) motif sizes and allele ranges of loci are set (Figure S5) and (iv) summary statistics are selected (Figure S6). After some hours, a reference table with 6 million simulated datasets (i.e. 2 million per scenario) is produced.

To answer the first question, the $n_\delta = 60\,000$ (1%) simulated datasets closest to the pseudo-observed dataset are selected for the logistic regression and $n_\delta = 600$ (0.01%) for the direct approach. The answer appears in two graphs (upper row in Figure S7). Both approaches are congruent and show that scenario 1 is significantly better supported by data than any other scenarios.

To answer the second question, scenario 1 is chosen and posterior distributions of parameters are estimated taking the 20 000 (1%) closest simulated datasets, after applying a *logit* transformation of parameter values. Here again, the output is mostly graphical. Each graph provides the prior and posterior distributions of the corresponding parameter (Figure S8). Below each graph are given the mean, median and mode as well as four quantiles (0.025, 0.05, 0.95 and 0.975) of the posterior distribution (Table S1 in Supplementary Material). Since the true values are known, we can remark that some parameters are rather well estimated with peaked posteriors such as the common effective population size and the two admixture rates, whilst other including all time parameters suggest that data are not very informative for them. Very similar results (data not shown) have been obtained with 5000 and 40 000 simulated datasets selected for the local linear regression, as well as when using a smaller reference table (1 million datasets).

To evaluate the confidence that can be put into the posterior probability of scenarios, 500 test datasets were simulated with each scenario and known parameter values (i.e. the same values as those used to produce the original dataset). Posterior probabilities of the three scenarios were estimated as above and used to compute type I and II errors in the choice of scenario. Results show that scenario 3 is always rightly chosen or excluded. Consequently type I error for scenario 1 is identical to type II error for scenario 2 and vice versa. For scenario 1, type I errors amount to 0.414 and 0.3 for the direct approach and the logistic regression, respectively, whereas type II errors amount to 0.014 and 0.020 (cf. Fig S9, S10 and S11 for detailed distributions of scenario probabilities). The 500 test datasets simulated with scenario 1 have also been used to estimate posterior distributions of parameters, taking the same proportion (1%) of closest simulated datasets as above. Relative biases and dispersion measures are given in Table S2 (upper part). It is clear that several parameters are biased and/or dispersed, the worst case being that of parameter t_1. The bias is undoubtedly related to the lack of information in the data, so that point estimates are drawn towards the mean values of prior distributions. The effect of prior distribution is also illustrated in the lower part of Table S2 that provides the same measures, but obtained with different prior distributions for effective size and time of event parameters.

3.2.2 Illustration on a real dataset Our second example concerns populations of the Silvereye, *Zosterops lateralis lateralis* (Estoup and Clegg, 2003). During the 19th and 20th century, this bird colonized Southwest Pacific islands from Tasmania. The importance of serial founder events in the microevolution of this species has been questioned in a study based on a six microsatellite loci dataset (Clegg *et al*, 2002).

Our analysis with *DIYABC* differs by at least four aspects from the initial ABC analysis processed from the same dataset by Estoup and Clegg (2003). First, all island populations are treated here in the same analysis whereas, for tractability reasons, independent analyses were made using all pairs of populations. Second, the initial treatment was based on the algorithm of Pritchard *et al.* (1999), whereas *DIYABC* uses the local linear regression method of Beaumont *et al.* (2002), which allows a larger number of statistics (see below) and hence makes a better use of data. Third, we have chosen here non-informative flat priors for all demographic parameters. Fourth, because *DIYABC* is able to treat samples collected at different times, we did not have to pool samples collected at different years from the same island and average sample year collection over islands. We hence end up with a colonization scenario involving five populations and seven samples, two samples having been collected at different times in two different islands (Fig. 2). The sequence and dates of colonization by silvereyes to New Zealand (South and North Island) and Chatham and Norfolk Islands have been historically documented. This allows the times for the putative population size fluctuation events in the coalescent gene trees to be fixed, thus limiting the number of parameters. Our scenario was specified by six unknown demographic parameters: the stable effective population size (N_S) and the duration of the initial bottleneck (D_B), both assumed to be the same in all Islands and potentially different effective number of founders in Norfolk, Chatham and South and North Island of New Zealand (N_{F1}, N_{F2}, N_{F3} and N_{F4}, respectively). As in Estoup and Clegg (2003), we also assumed that all populations evolved as totally isolated demes after the date of colonization.

We chose uniform priors $U[300, 30000]$ for N_S, $U[2, 500]$ for all N_{Fi} and $U[1, 5]$ for D_B. Prior information regarding the mutation rate and model for microsatellites was the same as in the previous example. Summary statistics included the mean number of alleles, the mean genic diversity (Nei, 1987), the mean coefficient M (Garza and Williamson, 2001), Fst between pairs of population samples (Weir and Cockerham, 1984), and the *mean classification index*, also called *mean individual assignment likelihood* (Pascual *et al.*, 2007). We produced a reference table with 1 million simulated datasets and estimated parameter posterior distributions taking the 10 000 (1%) simulated datasets closest to the observed dataset for the local linear regression, after applying a *logit* transformation to parameter values. Similar results were obtained when taking the 2000 to 20 000 closest simulated datasets and when using a log or log-tangent transformation of parameters as proposed in Estoup *et al.* (2004) and Hamilton *et al.* (2005) (options available in *DIYABC*).

Results for the main demographic parameters are presented in Table 1. They indicate the colonization by a small number of

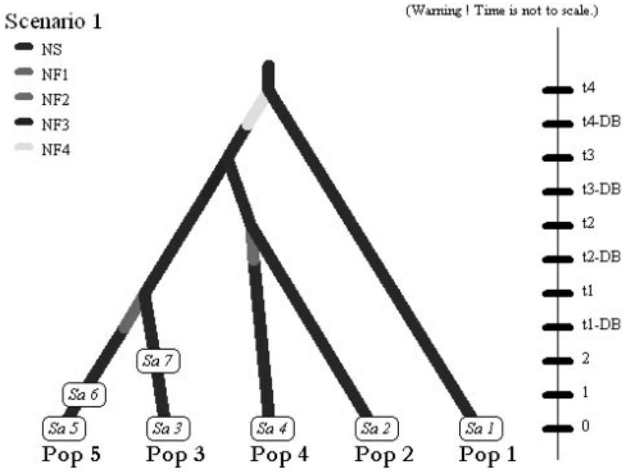

Fig. 2. Second example: screenshot of the scenario used in the analysis of the *Z. lateralis lateralis* dataset. In 1830, *Z. l. lateralis* colonized the South Island of New Zealand (Pop 2) from Tasmania (Pop 1). In the following years, the population began expanding and dispersing, and reached the North Island by 1856 (Pop 3). Chatham Island (Pop 4) was colonized in 1856 from the South Island, and Norfolk Island (Pop 5) was colonized in 1904 from the North Island (historical information reviewed in Estoup and Clegg 2003). Sample collection times are 1997 for Tasmania (*Sa* 1), South and North Island of New Zealand (*Sa* 2 and *Sa* 3, respectively), Chatham Island (*Sa* 4) and Norfolk Island (*Sa* 5), 1994 for the second sample from Norfolk (*Sa* 6), and 1992 for the second sample from the North Island of New Zealand (*Sa* 7). Splitting events and sampling dates in years were translated in number of generations since the most recent sampling date by assuming a generation time of 3 years (Estoup and Clegg, 2003). We hence fixed t1, t2, t3 and t4 to 31, 47, 47 and 56 generations, respectively.

Table 1. Second example: mean, median, mode, quantiles and SD of posterior distribution samples for effective population sizes (original parameters) for the *Z.lateralis lateralis* dataset.

Parameter	mean	median	mode	$Q_{0.050}$	$Q_{0.950}$	SD
N_S	9399	7446	4107	2706	23 007	6273
N_{F1}	19	18	16	9	33	8.7
N_{F2}	202	173	108	55	435	118
N_{F3}	197	168	112	55	430	116
N_{F4}	293	288	278	129	470	105

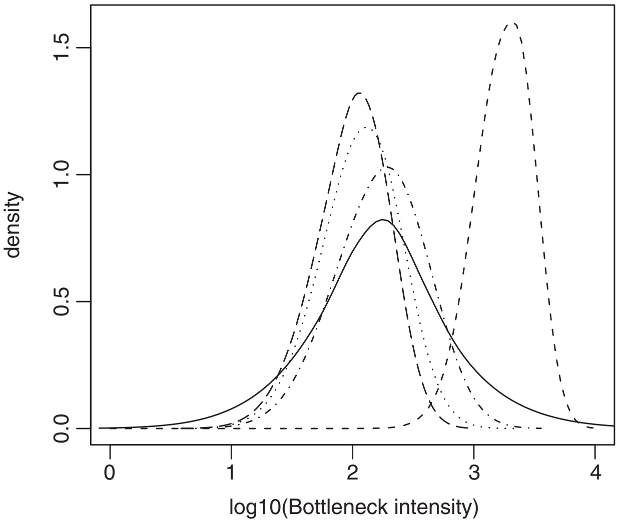

Fig. 3. Second example: posterior distributions of the bottleneck severity (see definition in text) for the invasions of four Pacific Islands by *Z.lateralis lateralis*. The four discontinuous lines with small dashes, dots, dash-dots and long dashes correspond to Norfolk, Chatham, North Island and South Island of New Zealand, respectively. The continuous line corresponds to the prior distribution, which is identical for each island. This graph has been made with the *locfit* function of the R statistical package (Ihaka and Gentleman, 1996), using an option of *DIYABC* which saves the sample of the parameter values adjusted by the local linear regression (Beaumont *et al.*, 2002).

founders and/or a slow demographic recovery after foundation for Norfolk island only (median N_{F1} value of 18 individuals). Other island populations appear to have been founded by silvereye flocks of larger size and/or have recovered quickly after foundation. In agreement with this, the bottleneck severity (computed as $BS_i = D_B \times N_S/N_{Fi}$) was more than one order of magnitude larger for the population from Norfolk than for other island populations (Fig. 3). These results are in the same vein as those obtained by Estoup and Clegg (2003) and agree with their main conclusions. Discrepancies in parameter estimation are observed however (e.g. larger N_S values and more precise inferences for N_{F2}, N_{F3} and N_{F4} in the present treatment). This was expected due to the differences in the methodological design underlined above. With the possibility of treating all population samples together, *DIYABC* allows a more elaborate and satisfactory treatment compared with previous analyses (Estoup and Clegg, 2003; Miller *et al.*, 2005).

4 CONCLUSIONS

So far, the ABC approach has remained inaccessible to most biologists because of the complex computations involved. With *DIYABC*, non-specialists can now perform ABC-based inference on various and complex population evolutionary scenarios, without reducing them to simple standard situations, and hence making a better use of their data. In addition, this programs also allows them to compare competing scenarios and quantify their relative support by the data. Eventually, it provides a way to evaluate the amount of confidence that can be put into the various estimations. The main limitations of the current version of *DIYABC* are the assumed absence of migration among populations after they have diverged and the mutation models which mostly refer to microsatellite loci. Next developments will aim at progressively removing these limitations.

ACKNOWLEDGEMENTS

We are most grateful to Sonya Clegg for authorizing the use of Silvereye microsatellite data in our second example.

Funding: The development of *DIYABC* has been supported by a grant from the French Research National Agency (project *MISGEPOP*) and an EU grant awarded to JMC as an EIF Marie-Curie Fellowship (project *StatInfPopGen*) that allowed him to spend two years in D.J.B.'s Epidemiology and Public Health department at Imperial College (London, UK) where he wrote a major part of this program.

Conflict of Interest: none declared.

REFERENCES

Beaumont,M.A. (1999) Detecting population expansion and decline using microsatellites. *Genetics*, **153**, 2013–2029.

Beaumont,M.A. (2003) Estimation of population growth or decline in genetically monitored populations. *Genetics*, **164**, 1139–1160.

Beaumont, M.A. (2008) Joint determination of topology, divergence time and immigration in population trees. In Matsumura,S., Forster,P. and Renfrew,C. (eds) *Simulations, Genetics and Human Prehistory*, (McDonald Institute Monographs), McDonald Institute for Archaeological Research, Cambridge, pp 134–154.

Beaumont,M.A. *et al.* (2002) Approximate Bayesian computation in Population Genetics. *Genetics*, **162**, 2025–2035.

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

Clegg,S.M. *et al.* (2002) Genetic consequences of sequential founder events by an island colonising bird. *Proc. Natl Acad. Sci.*, **99**, 8127–8132.

De Iorio,M. and Griffiths,R.C. (2004) Importance sampling on coalescence histories. ii: subdivided population models. *Adv. Appl. Probab.*, **36**, 434–454.

Dib,C. *et al.* (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature*, **380**, 152–154

Estoup,A. *et al.* (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Mol. Ecol.*, **11**, 1591–1604.

Estoup,A. and Clegg,S.M. (2003) Bayesian inferences on the recent island colonization history by the bird *Zosterops lateralis lateralis*. *Mol. Ecol.*, **12**, 657–674.

Estoup,A. *et al.* (2004) Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, *Bufo marinus*. *Evolution*, **58**, 2021–2036.

Franck,P. *et al.* (2000) Hybrid origins of honeybees from Italy (*Apis mellifera ligustica*) and Sicily (*A. m. sicula*). *Mol. Ecol.*, **9**, 907–992.

Excoffier,L. *et al.* (2005) Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics*, **169**, 1727–1738.

Fagundes,N.J.R. *et al.* (2007). Statistical evaluation of alternative models of human evolution. *Proc. Natl Acad. Sci.*, **104**, 17614–17619.

Garza,J.C. and Williamson,E. (2001) Detection of reduction in population size using data from microsatellite DNA. *Mol. Ecol.*, **10**, 305–318.

Griffiths,R.C. and Tavaré,S. (1994) Simulating probability distributions in the coalescent. *Theor. Popul. Biol.*, **46**, 131–159.

Hamilton,G. *et al.* (2005) Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. *Proc. Natl Acad. Sci. USA*, **102**, 7476–7480.

Hey,J. and Nielsen,R. (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, **167**, 747–760.

Hickerson,M.J. *et al.* (2007) msBayes: Pipeline for testing comparative phylogeographic histories using hierarchical approximate Bayesian computation. *BMC Bioinformatics*, **8**, 268–274.

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

Laval,G. and Excoffier,L. (2004) SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics*, **20**, 2485–2487.

Marjoram,P. *et al.* (2003) Markov chain Monte Carlo without likelihood. *Proc. Natl Acad. Sci.*, **100**, 15324–15328.

Miller,N. *et al.* (2005) Multiple transatlantic introductions of the western corn rootworm. *Science*, **310**, 992.

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

Nordborg,M. (2007) Coalescent theory. In *Handbook of Statistical Genetics* Balding,D.J. *et al.* (eds) 3rd edn. Wiley & Sons, Chichester, UK, pp. 843–877.

O'Ryan,C. *et al.* (1998) Genetics of fragmented populations of African buffalo (*Syncerus caffer*) in South Africa. *Anim. Conserv.*, **1**, 85–94.

Pascual,M. *et al.* (2007) Introduction history of *Drosophila subobscura* in the New World: a microsatellite based survey using ABC methods. *Mol. Ecol.*, **16**, 3069–3083.

Pritchard,J. *et al.* (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.*, **16**, 1791–1798.

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

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

Tavaré,S. *et al.* (1997) Inferring coalescence times from DNA sequences. *Genetics*, **145**, 505–518.

Wang,J. (2003) Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics*, **164**, 747–765.

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

Weis,G. and von Haeseler,A. (1998) Inference of population history using a likelihood approach. *Genetics*, **149**, 1539–1546.

Wilson,I.J. *et al.* (2003) Inferences from DNA data: population histories,evolutionary processes, and forensic match probabilities. *J. R. Stat. Soc. A*, **166**, 155–187.