Infering population history with \textit{DIY ABC}: a user-friendly approach to Approximate Bayesian Computation

Jean-Marie Cornuet\textsuperscript{a,b}, Filipe Santos\textsuperscript{b}, Mark A. Beaumont\textsuperscript{c}, Christian P. Robert\textsuperscript{d}, Jean-Michel Marin\textsuperscript{e}, David J. Balding\textsuperscript{a}, Thomas Guillemaud\textsuperscript{f} and Arnaud Estoup\textsuperscript{b}

April 11, 2008

\textsuperscript{a}Department of Epidemiology and Public Health, Imperial College, St Mary’s Campus Norfolk Place, London W2 1PG, U.K.
\textsuperscript{b}Centre de Biologie et de Gestion des Populations,INRA, Campus International de Baillarguet, CS 30016 Montferrier-sur-Lez, 34988 Saint-Gély-du-Fesc Cedex, France
\textsuperscript{c}School of Biological Sciences, Lyle Building, The University of Reading Whiteknights, Reading RG6 6AS, UK
\textsuperscript{d}CEREMADE, Université Paris-Dauphine, Place Delattre de Tassigny, 75775 Paris cedex 16, France
\textsuperscript{e}INRIA Saclay, Projet select, Université Paris-Sud, Laboratoire de Mathématiques (Bât. 425), 91400 Orsay, France
\textsuperscript{f}UMR 1301 I.B.S.V. INRA-UNSA-CNRS. 400 Route des Chappes. BP 167 - 06903 Sophia Antipolis cedex. France
Running head: ABC inference of population history

Corresponding author: Jean-Marie Cornuet
Department of Epidemiology and Public Health, Imperial College
St Mary’s Campus, Norfolk Place, London W2 1PG, U.K.
Tel :+44 2075943420
e-mail :j.cornuet@imperial.ac.uk

Keywords: ABC, Microsatellite, Coalescent, Software, Statistical inference

Abstract: Genetic data obtained on population samples convey information about their evolutionary history. Inference methods can extract this information (at least partially) 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 (DIYABC) 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 stepwise population size changes. DIYABC 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 data set, both with complex evolutionary scenarios, illustrates the main possibilities of DIYABC.
Until now, most literature and software about inference in population genetics concern simple standard evolutionary scenarios: a single population (Griffiths and Tavaré, 1994; Stephens and Donnelly, 2000; Beaumont, 1999), two populations exchanging genes (Hey and Nielsen, 2004; De Iorio and Griffiths, 2004) or not (Hickerson et al., 2007) or three populations in the classic admixture scheme (Wang, 2003; Excoffier et al., 2005). 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 that 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 3 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 in starting from a gene genealogy compatible with the observed data and exploring the parameter space through MCMC algorithms. In this approach, the tree topology and the branch lengths are parameters in the analysis in the same way as are the scenario parameters. 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 developed 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.

With complex scenarios, both previous approaches raise difficulties that mainly stem in the computation/estimation of the likelihood. This prompted a new 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) which developed in a new likelihood-free 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 data sets, similarity usually measured by a distance between summary statistics computed on both data sets. 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 parameter estimates for simulated data sets produced with known values of parameters with little extra computational cost. To get the same important information with likelihood-based methods would require a huge amount of additional computations whereas, with ABC, the largest proportion of computations 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 data set can be estimated (e.g. Miller et al., 2005; Pascual et al., 2007; Fagundes et al., 2007).

However, in its current state, the ABC approach remains inaccessible to most biologists because there is not yet a simple software solution. In order to fill this gap, we developed the program DIYABC that performs ABC analyses on complex scenarios. By complex scenarios, we mean scenarios 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.
KEY METHODS INVOLVED IN DIY ABC

Inference about the posterior distribution of parameters in an ABC analysis is usually performed in three steps (Figure 1). 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 data set and contains the parameter values used to simulate the data set and summary statistics computed on the simulated data set. 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 data sets generally lies between $10^5$ and $10^7$. The second step is a rejection step. Euclidian distances between each simulated and the observed data set in the space of summary statistics are computed and only the simulated data sets closest to the observed data set are retained. The parameter values used to simulate these selected data sets 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 now well established and we now concentrate on more specific issues that are implemented in DIY ABC.

Simulating genetic data in complex scenarios

Thanks to coalescence theory, it has become easy to simulate data sets 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 \( j(j - 1)/4N \), when there are \( j \) distinct lineages and the (diploid) effective population size is \( N \). 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. Depending on whether the last (not discarded) coalescence or the divergence time comes first (looking backward in time), there will remain one or more gene lineages in each population at generation \( t_d \). In any case, the remaining lineages of both populations 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 other than giving the two populations different number of generations for simulating coalescences before divergence.

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/generations,
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 discrete changes of population size can be simulated with the first module just varying the effective population size parameter value when the time of change is reached. It might also be extended to include continuous (linear or exponential) 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 to 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.

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 coalescences occur at the same generation is low enough so that it can be neglected. Time is then considered as a continuous variable in computations (see NORDborg (2007) for a detailed explanation). The corresponding algorithm, called here the *continuous time* (CT) algorithm, consists in drawing first times between two successive coalescence events (these times are exponentially distributed with rates equal to \((k/2)\) when there are \(k\) remaining lineages) and then drawing 2 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 \(U[1, 2N]\) and lineages with the same number coalesce at generation \(g+1\). This second algorithm is termed here the *generation by generation* (GbG) algorithm. It corresponds exactly to the Wright-Fisher model definition.

The CT algorithm is much faster in most cases and is generally preferred in many softwares, but in some circumstances, the approximation becomes unacceptable. The solution taken in DIY ABC is to swap between the two methods 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 can occur when the effective size is very small) or when the continuous time (CT) algorithm produces on average a number of lineages at the end of the module that is more than 5% larger than that obtained through the GbG algorithm.

A specific comparison study has been performed to establish this criterion. For different time periods (counted in generations), 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. Three types of output were distinguished:
1. the CT algorithms produced an average number of coalescences that differed from that of GbG by at least 5%;

2. the CT algorithm produced an average number of coalescences within 5% of that produced by the GbG, but the GbG was faster;

3. the CT algorithm produced an average number of coalescences within 5% of that produced by the GbG, and it was faster.

Only in the situation (3) is the CT advantageous over the GbG. Figure 2 shows that the boundary separating situation (3) from the other two is quasi-linear with zero intercept and a slope that varies with the duration ($g$ counted in generations) of the coalescence module. We found that a piecewise approximation of the slope as a function of $g$ provided a good fit of the limit of the area in which the CT algorithm could advantageously replace the GbG algorithm. DIYABC hence uses the following decision rules:

- 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

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), data sets are ordered by increasing distance to the observed data set. A first way (termed hereafter the direct approach) is to take as an estimate of the posterior probability of a scenario the proportion of data sets obtained with this scenario in the $n_\delta$ closest data sets (Miller et al., 2005; Pascual et al., 2007). The value of $n_\delta$ is arbitrary and unless the results are quite clear cut, the estimated posterior probability may vary with $n_\delta$.

Following the same rationale that introduced the local linear regression in the estimation of posterior distributions for parameters (Beaumont et al., 2002), we (Beaumont, 2008; Fagundes
et al., 2007) suggested to perform a weighted polychotomous logistic regression to estimate the posterior probability of scenarios (termed hereafter the logistic approach). 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 data set, using an Epanechnikov kernel based on the distance between observed and simulated summary statistics. Parameters values are then replaced by their estimates at that point in the regression. Since that point is the origin of the predictor space (null coordinates), the estimate is the sum of the constant term of the regression (intercept) and the residual 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 Epanechnikov 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 Appendix 1 for more details.

Quantifying confidence in parameter estimations on simulated test data sets

In order to be confident in estimates, these should be unbiased and precise enough. In order to measure bias and precision, we need to simulate data sets (i.e. test data sets) 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 data set, it can also be used to quantify bias and precision on test data sets 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 50 and 95%
coverages and the factor 2. See Appendix 2 for precise definitions of these indices.

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

**Main features**

*DIY ABC* is a program that performs ABC inference on population genetics data. In its current state, the data are genotypes at microsatellite loci of samples of diploid individuals representative of populations. 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. *DIY ABC* 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 including any number of populations and samples. Scenarios can include unsampled as well as multi-sampled (at different times) populations. Missing data are allowed. Scenarios can include any number of the following timed events: stepwise change of effective population size, population divergence and admixture. The main restriction regarding scenario complexity is the absence of migrations between populations after they have diverged.

Since the program has been written for microsatellite data, it proposes two popular mutation models, namely the stepwise mutation model and the generalized stepwise mutation model (*ESTOUP et al.*, 2002). In addition, it includes single nucleotide mutations to simulate uneven insertion/deletion events commonly observed in real microsatellite data sets (*PASCUAL et al.*, 2007; see also example in section 2.2.2). Note that the same mutation model has to be applied to all microsatellite loci, but these may have different values of mutation parameters except for the probability of a single nucleotide mutation which is common to all loci.

The historico-demographic parameters of the studied evolutionary 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 $\mu$ 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), mean and standard deviation (for Normal and Log-normal) and a step value that can be used to round values. In addition, priors can be modified by setting binary conditions on pairs of parameters of the same category (two effectives sizes or two times of event). These binary conditions are $>$, $<$, $\geq$ and $\leq$. 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 distribution are considered, the latter being defined by the mean and the shape. The choice of the mean is meant to facilitate the implementation of a hierarchical scheme, with a mean mutation rate or coefficient $P$ 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, $F_{ST}$, $(\delta\mu)^2$ or admixture rates.

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 on the same data set, 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 data sets in a format compatible with the program Genepop (Raymond and Rousset, 1995).

**Two examples of analysis with DIYABC**

**Illustration on a simulated data set:** In order to illustrate the capabilities of DIYABC, we take first an example based on a simulated data set. To produce this data set, we used the corresponding option of the program. The data set has been simulated according to a complex scenario including
three splits and two admixture events¹ (scenario 1 in Figure 3). The scenario also 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. Each locus has a possible range of 40 contiguous allelic states and is assumed to follow the GSM with $\mu$ drawn from a Gamma(mean=0.0005 and shape=2) and $P$ drawn from a Gamma(mean=0.22 and shape=2). Our ABC analysis will address the following questions:

1. Suppose that we are not sure that the scenario having produced our example data set 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 3) and the other with no admixture at all (scenario 3 in Figure 3). What is the posterior probability of these three scenarios, giving them identical prior probabilities?

2. What are the posterior distributions of parameters, given that the right scenario is known?

3. What confidence can we have into the parameter estimation?

We need first to build up the reference table. Using different screens of DIYABC, i) we code the three scenarios and define the prior distribution of parameters (Figure S1 provided in Supplementary data), ii) we select the GSM and define prior distribution for the mutation parameters (Figure S2), iii) we set the motif size and the allele range of each locus (Figure S3) and iv) we select a set of summary statistics (Figure S4). Eventually, we start running simulations and after some hours (on a laptop), we get a reference table with 6 million simulated data sets (i.e. 2 million per scenario).

To answer the first question, we run the option ”compute posterior probabilities of scenario”, taking the $n_\delta=60,000$ (1%) simulated data sets closest to the pseudo-observed data set for the logistic regression and $n_\delta=600$ (0.01%) for the direct approach. The answer appears in two graphs

¹This 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 in other countries 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".
(Figure 4). Clearly both approaches are congruent and show that scenario 1 is significantly better supported by data than any other scenarios.

To answer the second question, we run the option “estimate parameters” choosing the first scenario. Here again, we take the 20,000 (1%) closest simulated data sets. Output include graphs providing the prior and posterior distributions of the corresponding parameter, as well as mean, median, modes and four quantiles (Figure S5). Under each graph on the DIYABC screen 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 1). Since we know the true values, we can remark that some parameters are rather well estimated with peaked posteriors such as the common effective population size and the admixture rate $r_1$, whilst data are not very informative for other parameters.

To evaluate quantitatively the amount of confidence that can be put into the parameter estimations, we run the option “compute bias and precision” on 100 test data sets with known parameter values, using the same proportion (1%) of closest simulated data sets as above. This requires a few hours at the end of which we have a complete table (see extract in Table 2) showing relative biases and dispersion measures. 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.

**Illustration on a real data set:** As a second example to illustrate analysis with DIYABC, we have chosen a real microsatellite data set obtained from several Pacific island populations of the bird *Zosterops lateralis lateralis*, commonly named silvereye (Estoup and Clegg, 2003). During the 19th and 20th century, this bird colonised Southwest Pacific islands from Tasmania. The importance of single founder events in the microevolution of this species has been questioned in a study based on a six microsatellite loci data set (Clegg et al, 2002), where single founder events were found to have relatively low impact on shaping the neutral genetic variation of at least some island populations of silvereyes.

Our analysis with DIYABC differs by at least four aspects from the initial ABC analysis processed from the same data set by Estoup and Clegg (2003). First all island populations
are treated here in the same analysis whereas, for tractability reasons, the populations have been treated independently by pair with Tasmania as source population in each pair and different number of founder events by ESTOUPE AND CLEGG (2003). Second, because parameter estimation in DIYABC is based on the local linear regression method of BEAUMONT et al. (2002), we could use a much larger number of statistics (see below), and hence better summarize information from the data, than in the initial ABC treatment that was based on the algorithm of PRITCHARD et al. (1999). Third, we have chosen 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 (Figure 5). The sequence and dates of colonisation by silvereyes to New Zealand (South and North Island) and outlying islands (Chatham and Norfolk islands) have been historically documented. This allows fixing the times for the putative population size fluctuation events in the coalescent gene trees, and hence to limit the number of parameters in our sequential colonization with founder event scenario. Our scenario was specified by six variable demographic parameters: the stable effective population size ($N_S$) which was assumed to be the same in all islands, potentially different effective number of founding individuals in Norfolk island, Chatham island and the South and North island of New Zealand ($N_{F1}$, $N_{F2}$, $N_{F3}$ and $N_{F4}$, respectively), and the duration of the bottleneck associated with the foundation of new populations ($D_B$), which was assumed to be the same for all colonization events. As in ESTOUPE AND CLEGG (2003), we also assumed that all populations evolved as totally isolated demes after the date of colonisation.

We chose uniform prior distributions bounded between 300 and 30,000 for $N_S$, between 2 and 500 for all $N_{Fi}$ and between 1 and 5 for $D_B$. It is worth noting that the durations of the bottleneck, the effective numbers of individuals during the bottlenecks and the stable effective population size are barely individually identifiable in the likelihood. Therefore, we also computed for each colonized population a combined parameter called the bottleneck severity index computed as $BS_i = D_B \times N_S / N_{Fi}$ (see Pascual et al. 2007 for a similar index). Such combined parameters
and their posterior distributions were computed from the original parameters values obtained after the regression step using the R package (IHAKA AND GENTLEMAN, 1996). Prior information regarding the mutation rate and model for dinucleotide repeats was the same as in the previous example, except that observed allele size differences indicate a larger possible range of contiguous allelic states (set to 70 for all loci). Regarding uneven insertion/deletion events that could be suspected for most loci from observed allele sizes, we used a Gamma prior distribution with mean=$2.5 \times 10^{-8}$ and shape=2 for drawing single locus insertion-deletion mutation rates (PASCUAL et al., 2007). Genetic variation within and between the seven population samples was summarized with the same summary statistics as in MILLER et al. (2005) and PASCUAL et al. (2007), i.e. the mean number of alleles ($A$), the mean genic diversity ($H$; NEI, 1987), the mean ratio of the number of alleles over the range of allelic sizes per population expressed in base pairs ($M$; GARZA AND WILLIAMSON, 2001), $F_{ST}$ between pairs of population samples (WEIR AND COCKERHAM, 1984), and the mean classification index for individuals collected in the population sample $i$ and assigned to the population $j$ (also called mean individual assignment likelihood; cf. formula 9 in Rannala and Mountain 1997). We hence had a total of 84 summary statistics. We produced a reference table with 1 million simulated data sets and estimate parameter posterior distributions taking the 10,000 (1%) simulated data sets closest to the observed data set for the local linear regression, after applying a logit transformation to parameter values. Similar results were obtained when taking the 2,000 to 20,000 closest simulated data sets and when using a log or log–tangent transformation of parameters as proposed in ESTOUPEt al. (2004) and HAMILTON et al. (2005) (options available in DIYABC).

Results for the main demographic parameters are presented in Table 3. They indicate the colonization by a small number of 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 was more than one order of magnitude larger for the population from Norfolk than for other island populations (Figure 6). These results are in the same vein as those obtained by ESTOUN 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). Such discrepancies are 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 to previous analyses (Estoup and Clegg, 2003; Miller et al., 2005).

**Conclusion**

So far, the ABC approach 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 remove these limitations.

**Acknowledgements**

The development of DIYABC has been supported by a grant from the French Research National Agency (project MISGEPOP) and a grant from the European Union awarded to JM Cornuet as an EIF Marie-Curie Fellowship (project StatInfPopGen) that allowed him to spend two years in DJ Balding’s Epidemiology and Public Health department at Imperial College (London, UK) where he wrote the major part of this program.

We are most grateful to Sonia Clegg for authorizing the use of Silverseye microsatellite data in our second example.
References


OHTA, T. AND M. KIMURA, 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population.


Figure legends

Figure 1: The three steps of an ABC analysis. The two boxes with a double line correspond to the case where there are more than one scenario.

Figure 2: Which algorithm for the coalescence module? Graphs indicate in light grey the area of the plane for which the generation by generation (GbG) algorithm is faster than the continuous time (CT) algorithm, in middle grey the area for which the CT algorithm produces significantly (5%) less coalescences than the GbG algorithm and in dark grey the area for which the CT algorithm produces the same number of coalescences than the GbG algorithm (with tolerance=5%) and is faster. Limits between areas are almost linear. The black line (intercept=0) has a slope taken as $0.0031g^2 - 0.053g + 0.7197$ for $g \leq 30$, $0.033g + 1.7$ for $30 < g \leq 100$ and 5 when $100 < g$, $g$ being the duration of the coalescence module in number of generations.

Figure 3: First example: the three evolutionary scenarios. The data set 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 1,000, 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 6 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. Time is not at scale. 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.

Figure 4: First example: posterior probability of scenarios. The x-axis correspond to the different $n_\delta$ values used in the computations and the y-axis gives the posterior probability of the three scenarios through the direct (left) and the logistic (right) approaches as output by DIYABC. In both graphs, the upper curve corresponds to scenario 1.

Figure 5: Second example: scenario used in the analysis of the Zosterops lateralis lateralis data
set. On the left is the scenario codification for DIYABC and on the right is the corresponding graphic representation produced by the program. Note the keyword varNe to inform of a change in effective size of populations. This keyword is followed by the population number and the new population size. In 1830, Z. l. lateralis colonised the South Island of New Zealand (Pop2) from Tasmania (Pop1). In the following years, the population began expanding and dispersing, and reached the North Island by 1856 (Pop3). Chatham Island (Pop4) was colonised in 1856 from the South Island, and Norfolk Island (Pop5) was colonised in 1904 from the North Island (historical information reviewed in Estoup and Clegg 2003). Sample collection times are 1997 for Tasmania (Sa1), South and North island of New Zealand (Sa2 and Sa3, respectively), Chatham island (Sa4) and Norfolk island (Sa5), 1994 for the second sample from Norfolk (Sa6), and 1992 for the second sample from the North island of New Zealand (Sa7). 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 three years (Estoup and Clegg 2003). We hence fixed t1, t2, t3 and t4 to 31, 47, 47 and 56 generations, respectively.

**Figure 6**: Second example: posterior distributions of the bottleneck severity for the successive invasions of four Pacific islands by Zosterops 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 R *locfit* function, using an option of DIYABC which allows to save the sample of the parameter values adjusted by the local linear regression (Beaumont et al., 2002).
Table 1: **First example: statistics of the posterior distribution of parameters under scenario 1**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>mean</th>
<th>median</th>
<th>mode</th>
<th>$Q_{0.025}$</th>
<th>$Q_{0.050}$</th>
<th>$Q_{0.950}$</th>
<th>$Q_{0.975}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$ (1,000)</td>
<td>1.684</td>
<td>1.425</td>
<td>1.050</td>
<td>659</td>
<td>729</td>
<td>3,422</td>
<td>4,320</td>
</tr>
<tr>
<td>$r_1$ (0.6)</td>
<td>0.577</td>
<td>0.577</td>
<td>0.577</td>
<td>0.495</td>
<td>0.512</td>
<td>0.642</td>
<td>0.658</td>
</tr>
<tr>
<td>$r_2$ (0.4)</td>
<td>0.499</td>
<td>0.500</td>
<td>0.547</td>
<td>0.107</td>
<td>0.162</td>
<td>0.828</td>
<td>0.880</td>
</tr>
<tr>
<td>$t_1$ (10)</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>4</td>
<td>7</td>
<td>95</td>
<td>97</td>
</tr>
<tr>
<td>$t_2$ (500)</td>
<td>742</td>
<td>797</td>
<td>979</td>
<td>227</td>
<td>303</td>
<td>984</td>
<td>992</td>
</tr>
<tr>
<td>$t_3$ (10,000)</td>
<td>12,324</td>
<td>10,126</td>
<td>6,538</td>
<td>5,263</td>
<td>5,526</td>
<td>27,040</td>
<td>32,334</td>
</tr>
<tr>
<td>$t_4$ (20,000)</td>
<td>33,068</td>
<td>34,186</td>
<td>43,836</td>
<td>11,414</td>
<td>13,867</td>
<td>48,402</td>
<td>49,244</td>
</tr>
<tr>
<td>$t_5$ (200,000)</td>
<td>315,213</td>
<td>323,636</td>
<td>335,525</td>
<td>95,697</td>
<td>120,911</td>
<td>481,411</td>
<td>490,450</td>
</tr>
<tr>
<td>$\bar{\mu}$ (0.0005)</td>
<td>0.00056</td>
<td>0.00055</td>
<td>0.00053</td>
<td>0.00018</td>
<td>0.00022</td>
<td>0.00091</td>
<td>0.00095</td>
</tr>
<tr>
<td>$P$ (0.22)</td>
<td>0.24</td>
<td>0.25</td>
<td>0.28</td>
<td>0.14</td>
<td>0.15</td>
<td>0.30</td>
<td>0.30</td>
</tr>
<tr>
<td>$\theta$ (=4N$\bar{\mu}$)</td>
<td>3.21</td>
<td>2.99</td>
<td>2.83</td>
<td>1.73</td>
<td>1.91</td>
<td>5.17</td>
<td>5.93</td>
</tr>
<tr>
<td>$\tau_1$ (=t_1$\bar{\mu}$)</td>
<td>0.028</td>
<td>0.025</td>
<td>0.015</td>
<td>0.002</td>
<td>0.003</td>
<td>0.065</td>
<td>0.072</td>
</tr>
<tr>
<td>$\tau_2$ (=t_2$\bar{\mu}$)</td>
<td>0.40</td>
<td>0.39</td>
<td>0.41</td>
<td>0.10</td>
<td>0.14</td>
<td>0.71</td>
<td>0.77</td>
</tr>
<tr>
<td>$\tau_3$ (=t_3$\bar{\mu}$)</td>
<td>6.57</td>
<td>5.42</td>
<td>4.41</td>
<td>1.90</td>
<td>2.30</td>
<td>14.82</td>
<td>18.45</td>
</tr>
<tr>
<td>$\tau_4$ (=t_4$\bar{\mu}$)</td>
<td>18.2</td>
<td>16.7</td>
<td>11.3</td>
<td>4.7</td>
<td>5.9</td>
<td>35.4</td>
<td>39.0</td>
</tr>
<tr>
<td>$\tau_5$ (=t_5$\bar{\mu}$)</td>
<td>173</td>
<td>156</td>
<td>111</td>
<td>44</td>
<td>54</td>
<td>343</td>
<td>380</td>
</tr>
</tbody>
</table>

Mean, median, mode and quantiles of the posterior distribution sample for original and composite parameters of the simulated data set. True values of original parameters (used to simulate the data set) are given between parentheses in the first column.
Table 2: First example: bias and precision of parameter estimation under scenario 1

<table>
<thead>
<tr>
<th>Parameter</th>
<th>true value</th>
<th>ARB</th>
<th>RRMISE</th>
<th>RMAD</th>
<th>RRMSE</th>
<th>95% cov.</th>
<th>fact2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>1,000</td>
<td>0.595</td>
<td>1.42</td>
<td>0.911</td>
<td>0.745</td>
<td>0.95</td>
<td>0.87</td>
</tr>
<tr>
<td>$r_1$</td>
<td>0.6</td>
<td>-0.015</td>
<td>0.105</td>
<td>0.081</td>
<td>0.068</td>
<td>0.98</td>
<td>1.00</td>
</tr>
<tr>
<td>$r_2$</td>
<td>0.4</td>
<td>0.135</td>
<td>0.664</td>
<td>0.557</td>
<td>0.453</td>
<td>0.99</td>
<td>0.92</td>
</tr>
<tr>
<td>$t_1$</td>
<td>10</td>
<td>3.714</td>
<td>4.764</td>
<td>3.913</td>
<td>3.825</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$t_2$</td>
<td>500</td>
<td>0.240</td>
<td>0.558</td>
<td>0.482</td>
<td>0.402</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>$t_3$</td>
<td>10,000</td>
<td>0.862</td>
<td>1.436</td>
<td>1.093</td>
<td>0.942</td>
<td>1.00</td>
<td>0.64</td>
</tr>
<tr>
<td>$t_4$</td>
<td>20,000</td>
<td>0.856</td>
<td>0.934</td>
<td>0.825</td>
<td>0.869</td>
<td>0.98</td>
<td>0.83</td>
</tr>
<tr>
<td>$t_5$</td>
<td>200,000</td>
<td>0.371</td>
<td>0.750</td>
<td>0.622</td>
<td>0.494</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>$\bar{\mu}$</td>
<td>0.0005</td>
<td>-0.055</td>
<td>0.447</td>
<td>0.377</td>
<td>0.211</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>$\bar{P}$</td>
<td>0.22</td>
<td>0.145</td>
<td>0.227</td>
<td>0.197</td>
<td>0.163</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>$\theta$  (=$4N\bar{\mu}$)</td>
<td>2</td>
<td>0.396</td>
<td>0.785</td>
<td>0.548</td>
<td>0.443</td>
<td>0.96</td>
<td>1.00</td>
</tr>
<tr>
<td>$\tau_1$  (=$t_1\bar{\mu}$)</td>
<td>0.005</td>
<td>3.017</td>
<td>5.197</td>
<td>3.839</td>
<td>3.256</td>
<td>1.00</td>
<td>0.01</td>
</tr>
<tr>
<td>$\tau_2$  (=$t_2\bar{\mu}$)</td>
<td>0.25</td>
<td>0.068</td>
<td>0.714</td>
<td>0.546</td>
<td>0.389</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td>$\tau_3$  (=$t_3\bar{\mu}$)</td>
<td>5</td>
<td>0.622</td>
<td>1.612</td>
<td>1.083</td>
<td>0.769</td>
<td>1.00</td>
<td>0.80</td>
</tr>
<tr>
<td>$\tau_4$  (=$t_4\bar{\mu}$)</td>
<td>10</td>
<td>0.592</td>
<td>1.200</td>
<td>0.889</td>
<td>0.699</td>
<td>1.00</td>
<td>0.85</td>
</tr>
<tr>
<td>$\tau_5$  (=$t_5\bar{\mu}$)</td>
<td>100</td>
<td>0.163</td>
<td>0.976</td>
<td>0.705</td>
<td>0.407</td>
<td>1.00</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Extract of the output of DIY ABC/option “Compute bias and mean square error” with scenario 1 and parameter values identical to those used to simulate the exemple data set. ARB is the average relative bias, RRMISE is the relative square root of the mean integrated square error, RMAD is the relative mean absolute deviation and RRMSE is the relative square root of the mean square error (see Appendix 2). The 95% cov. column contains the proportion of times where the true value is within the 95% credibility interval and the fact2 column contains the proportion of times where the estimate is between the half and the double of the true value. For the RMSE, the chosen estimate has been the median of the posterior distribution sample. Values have been obtained with 100 test data sets simulated with parameter values as shown in the second column.
Table 3: *Second example: statistics of the posterior distribution sample*

<table>
<thead>
<tr>
<th>Parameter</th>
<th>mean</th>
<th>median</th>
<th>mode</th>
<th>$Q_{0.050}$</th>
<th>$Q_{0.950}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_S$</td>
<td>9,399</td>
<td>7,446</td>
<td>4,107</td>
<td>2,706</td>
<td>23,007</td>
</tr>
<tr>
<td>$N_{F1}$</td>
<td>19</td>
<td>18</td>
<td>16</td>
<td>9</td>
<td>33</td>
</tr>
<tr>
<td>$N_{F2}$</td>
<td>202</td>
<td>173</td>
<td>108</td>
<td>55</td>
<td>435</td>
</tr>
<tr>
<td>$N_{F3}$</td>
<td>197</td>
<td>168</td>
<td>112</td>
<td>55</td>
<td>430</td>
</tr>
<tr>
<td>$N_{F4}$</td>
<td>293</td>
<td>288</td>
<td>278</td>
<td>129</td>
<td>470</td>
</tr>
<tr>
<td>$BS_1$</td>
<td>2036</td>
<td>1816</td>
<td>1330</td>
<td>665</td>
<td>4096</td>
</tr>
<tr>
<td>$BS_2$</td>
<td>154</td>
<td>119</td>
<td>58</td>
<td>30</td>
<td>475</td>
</tr>
<tr>
<td>$BS_3$</td>
<td>245</td>
<td>175</td>
<td>72</td>
<td>37</td>
<td>690</td>
</tr>
<tr>
<td>$BS_4$</td>
<td>125</td>
<td>104</td>
<td>57</td>
<td>26</td>
<td>293</td>
</tr>
</tbody>
</table>

Mean, median, mode and quantiles of posterior distribution samples for effective population sizes (original parameters) and bottleneck severities (composite parameters, see text for definition) for the *Zosterops lateralis lateralis* data set.
Figure 1

**SIMULATION STEP**

- Draw parameter values from prior distributions
- Simulate genetic data according to scenario and mutation model
- Compute summary statistics
- Record parameter and summary statistic values in a reference table file

**REJECTION STEP**

- Compute distances between observed and simulated summary statistics
- Retain simulated data sets closest to observed data

**ESTIMATION STEP**

- Estimate posterior probability of scenarios by logistic regression
- Estimate posterior distribution of parameter by local linear regression
Figure 2
Figure 3
Figure 4
Figure 5

NS NS NS NS NS
0 sample 1
0 sample 2
0 sample 3
0 sample 4
0 sample 5
1 sample 5
2 sample 3
t1-DB VarNe 5 NF1
t1 merge 3 5
t2-DB VarNe 4 NF2
t2 merge 2 4
t3-DB VarNe 3 NF3
t3 merge 2 3
t4-DB VarNe 2 NF4
t4 merge 1 2
Figure 6
Appendix 1: Computation of the posterior probability of a scenario and its credibility interval through a weighted polychotomic logistic regression

Consider an analysis in which we want to evaluate the posterior probability of $K$ scenarios. We have a reference table of $N$ simulated data sets and for each data set, there are $J$ summary statistics, noted $S^k_{ij}$ ($i \in [1, N]$, $j \in [1, J]$, $k \in [1, K]$) for the $j$-th summary statistic of the $i$-th data set simulated according to scenario $k$. Let $S_{obs,j}$ the $j$-th summary statistic of the observed data set.

Let $Y$ a matrix of $N$ rows and $K$ columns such that the element $y_{ik} = 1$ if the $i$-th data set has been simulated with scenario $k$ and 0 otherwise. Eventually, let $\omega_i$ = the Epanechnikov weight given to the $i$-th data set, the same as in the local linear regression use for parameters (Beaumont et al., 2002) but normalized such that $\sum_{i=1}^N \omega_i = N$.

We are looking for the relationship between the dependent variable $p_{ik}$, the posterior probability of scenario $k$ for the data set $i$ and the predictive variables $x_{ij} = S^k_{ij} - S_{obs,j}$.

Using a logit link function, the regression model used is:

$$
\text{logit}(p_{ik}) = \beta_{k0} + \beta_{k1}x_{i1} + \ldots + \beta_{kJ}x_{iJ} = \beta_k^T x_i
$$

with $\beta_k = (\beta_{k0}, \beta_{k1}, \ldots, \beta_{kJ})$ and $x_i = (1, x_{i1}, \ldots, x_{iJ})$

Because of the constraint $\sum_{k=1}^K p_{ik} = 1$, we consider only the first $K - 1$ scenarios, the posterior probability of the last one being deduced from the others, i.e.:

$$
p_k = Pr(\text{Scen} = k|X = x) = \frac{e^{\beta_k^T x}}{1 + \sum_{l=1}^{K-1} e^{\beta_l^T x}}
$$

for $k = 1, ..., K - 1$, and

$$
p_K = Pr(\text{Scen} = K|X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\beta_l^T x}}
$$

This regression provides an estimation of the relative probability $p_{obs,k}$ that the observed data set would be obtained with scenario $k$, considering that the actual scenario is one of the $K$. Since
by definition all $x_{obs,j}$ are zero, the estimates of $p_{obs,k}$ simplify to:

$$\hat{p}_{obs,k} = \frac{e^{\hat{\beta}_{k0}}}{1 + \sum_{l=1}^{K-1} e^{\hat{\beta}_{l0}}} \quad \text{for} \quad k \in [1, K-1], \quad \text{and} \quad \hat{p}_{obs,K} = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\hat{\beta}_{l0}}} \quad (1)$$

Noting $\beta_K = (0, 0, ..., 0)$, the likelihood is:

$$lik = \prod_{i=1}^{N} \left\{ \prod_{k=1}^{K} \left\{ \frac{e^{\hat{\beta}_k x_i}}{1 + \sum_{l=1}^{K-1} e^{\hat{\beta}_l x_i}} \right\} y_{ik} \omega_i \right\}$$

and the log-likelihood:

$$log(lik) = \sum_{i=1}^{N} \omega_i \left\{ \sum_{k=1}^{K} y_{ik} \log \left( \frac{e^{\hat{\beta}_k x_i}}{1 + \sum_{l=1}^{K-1} e^{\hat{\beta}_l x_i}} \right) \right\}$$

Let call $\beta$ the vector $(\beta_{10}, \beta_{11}, ..., \beta_{1J}, \beta_{20}, ..., \beta_{2J}, ..., \beta_{K-1,0}, ..., \beta_{K-1,J})$. The Newton Raphson algorithm is such that starting from an arbitrary value of $\beta$, the following equation is repeated until maximum likelihood is reached and convergence is obtained ($\beta^{new} = \beta^{old}$).

$$\beta^{new} = \beta^{old} + \left( \frac{\partial^2 \log(lik)}{\partial \beta_u \partial \beta_u} \right)^{-1} \times \left( \frac{\partial \log(lik)}{\partial \beta_u} \right)$$

Furthermore, $\hat{\beta} \approx N(\beta, V_\beta)$ where $V_\beta = \left( \frac{\partial^2 \log(lik)}{\partial \beta_u \partial \beta_u} \right)^{-1}$.

Because only $\hat{\beta}_{k0}$ are useful in formula (1), let define $\beta^*$ the sub-vector $(\beta_{10}, \beta_{20}, ..., \beta_{K-1,0})$, and $V^*$ the corresponding sub-matrix of $V_\beta$.

Noting $d = 1 + \sum_{l=1}^{K-1} e^{\hat{\beta}_l}$, equation (1) reduces to

$$\hat{p}_{obs,k} = \frac{e^{\hat{\beta}_k}}{d} \quad 1 \leq k \leq K - 1 \quad \hat{p}_{obs,K} = \frac{1}{d} \quad (2)$$

According to Slutsky’s formula, $\hat{p}_k \approx N(p_k, B_k^T V^* B_k)$, with $B_k = (\frac{\partial p_k}{\partial \beta_{10}}, \frac{\partial p_k}{\partial \beta_{20}}, ..., \frac{\partial p_k}{\partial \beta_{K-1,0}})$

Elements of $B_k$ being the following:

$$\frac{\partial p_k}{\partial \beta_{10}} = \frac{e^{\hat{\beta}_k (d-e^{\hat{\beta}_k})}}{d^2} \quad \text{si} \quad l = k \quad \text{et} \quad k < K$$

$$\frac{\partial p_k}{\partial \beta_{20}} = -\frac{e^{\hat{\beta}_k} e^{\hat{\beta}_l}}{d^2} \quad \text{si} \quad l \neq k \quad \text{et} \quad k < K$$

$$\frac{\partial p_k}{\partial \beta_{K-1,0}} = -\frac{e^{\hat{\beta}_k}}{d^2} \quad \text{si} \quad k = K$$
it is straightforward to derive approximate C.I. for $p_{obs,k}$. 
Appendix 2: Measures of dispersion computed in DIYABC

the Relative square Root of the Mean Square Error (RRMSE): the square root of the average square difference between the point estimate \((e)\) and the true value, divided by the true value,

\[
\frac{1}{v} \sqrt{\frac{1}{n} \sum (e - v)^2}
\]

the Relative square Root of the 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, \(\frac{1}{v} \sqrt{\frac{1}{n_x} \sum (\frac{1}{n_x} \sum (x - v)^2)}\). \(x\) is a value of the sample of the posterior distribution for a given test data set, \(n_x\) being the sample size. \(n\) is the number of test data sets and \(v\) is the true value.

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, \(\frac{1}{v} \frac{1}{n} \sum (\frac{1}{n_x} \sum |x - v|)\)

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