

INVITED REVIEW

ABC as a flexible framework to estimate demography over space and time: some cons, many pros

G. BERTORELLE,* A. BENAZZO* and S. MONA*†‡

*Department of Biology and Evolution, University of Ferrara, Via Borsari 46, 44100 Ferrara, Italy, †CMPG, Institute of Ecology and Evolution, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland, ‡Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

Abstract

The analysis of genetic variation to estimate demographic and historical parameters and to quantitatively compare alternative scenarios recently gained a powerful and flexible approach: the Approximate Bayesian Computation (ABC). The likelihood functions does not need to be theoretically specified, but posterior distributions can be approximated by simulation even assuming very complex population models including both natural and human-induced processes. Prior information can be easily incorporated and the quality of the results can be analysed with rather limited additional effort. ABC is not a statistical analysis *per se*, but rather a statistical framework and any specific application is a sort of hybrid between a simulation and a data-analysis study. Complete software packages performing the necessary steps under a set of models and for specific genetic markers are already available, but the flexibility of the method is better exploited combining different programs. Many questions relevant in ecology can be addressed using ABC, but adequate amount of time should be dedicated to decide among alternative options and to evaluate the results. In this paper we will describe and critically comment on the different steps of an ABC analysis, analyse some of the published applications of ABC and provide user guidelines.

Keywords: approximate Bayesian computation, likelihood-free inference, molecular ecology, population demography, population genetics, population history

Received 19 January 2010; revision received 20 April 2010; accepted 21 April 2010

Introduction

Population genetics is the analysis and understanding of genetic variation within and between populations. Early population geneticists, possibly also as a consequence of the paucity of empirical data, were mainly concerned with the theoretical framework of this discipline. Assuming simple demographic and evolutionary models, expected genetic variation patterns were theoretically predicted and sometimes compared with the available genetic information. During an intermediate phase from the 1970s to the early 1990s, when classical genetic markers were easily typed and the use of molec-

ular markers began to spread following the introduction of the PCR, descriptive analyses of genetic variation dominated. Methods such principal component analysis (PCA), spatial autocorrelation, and analysis of molecular variance (AMOVA) were widely used to describe patterns and informally compare hypotheses (e.g. Menozzi *et al.* 1978; Sokal *et al.* 1987; Excoffier *et al.* 1992). Parameter estimation and probability-based comparison of different scenarios were limited and imprecise, due to the fact that contemporary models were unrealistic and that more complex demographic and genetic models were theoretically intractable or computationally prohibitive. More recently, the increased speed and power of personal computers favoured the spread of Monte Carlo algorithms. Likelihood functions can be approximated thanks to Markov Chain Monte Carlo (MCMC) methods

Correspondence: Giorgio Bertorelle, Fax: +390532249771; E-mail: ggb@unife.it

(e.g. Kuhner *et al.* 1995; Nielsen & Wakeley 2001; Drummond *et al.* 2002) and brute power can be used to simulate gene genealogies under virtually any demographic and genetic model and to approximate the likelihood functions even without explicitly defining them (e.g. Fu & Li 1997; Tavaré *et al.* 1997; Beaumont *et al.* 2002). This latter approach, called approximate Bayesian computation (ABC) in its Bayesian version, is the topic of this review. We believe that ABC is matching, for the first time in population genetics studies, abundant genetic data and realistic (which usually means complex) evolutionary scenarios, allowing (i) the simultaneous estimation of posterior distributions for many parameters relevant in ecological studies; (ii) the probabilistic comparison of alternative models; and (iii) the quantitative evaluation of the results' credibility.

Approximate Bayesian computation is intuitively very easy: millions of genealogies are simulated assuming different parameter values and under different models and the simulations that produce genetic variation patterns close to the observed data are retained and analysed in detail. Parameter values and model features in the retained simulations are of course interesting since they are able to generate data sets with some properties, measured by summary statistics (SuSt hereafter), found in the observed data. At the same time, even if software packages are now available (e.g. Cornuet *et al.* 2008; Wegmann *et al.* 2010), ABC is not (yet?) user-friendly. Users are typically required to: (i) carefully consider each step in the ABC protocol since consensus on the best way to proceed has not been reached; and (ii) estimate the quality of the final results. In short, ABC is mathematically graceless and rather intricate to apply, but very flexible and powerful. In this review we will describe and critically comment on the different steps of an ABC analysis, analyse some of the published applications of ABC and provide throughout the paper some user guidelines. We will not discuss the recent criticisms to ABC and in general to Bayesian methods (Templeton 2010a,b). Detailed answers can be found, for example, in Beaumont *et al.* (2010).

First of all, we present the main ABC concepts in a historical perspective.

ABC: main concepts and history

Origins

The basic idea of ABC can be found in two papers published in February 1997. Stimulated by Templeton (1993) to find a correct estimator of the time to the most recent common ancestor (TMRCA) for a set of DNA sequences and assuming a simple demographic model of a single demographically stable population, Fu & Li

(1997) and Tavaré *et al.* (1997) proposed simulating artificial data-sets and using SuSt to select among them. The selected data-sets, used to estimate the posterior distribution of the TMRCA, were either those having exactly the same maximum number of pairwise differences k_{\max} as the observed data set (Fu & Li 1997) or those having a gene genealogy whose total length was compatible with the observed number of segregating sites, S (Tavaré *et al.* 1997). The former approach can be almost considered 'theory-free', since knowledge of probability functions is not needed to approximate likelihood or posterior densities of the quantities of interest under any specified demographic and mutational model. This is the reason why the Fu & Li (1997) idea can, in principle, be applied to any demographic scenario, favouring its spread and extension with little theoretical effort. On the other hand, the algorithm proposed by Tavaré *et al.* (1997) had the merit of explicitly introducing the Bayesian component [the parameter $\theta = 4N\mu$ was not fixed as in Fu & Li (1997), but sampled from a prior distribution], which is a key aspect of modern ABC.

All the information contained in the data is not captured by a single SuSt. Also, if simulated data-sets are retained only when they show a SuSt identical to the SuSt observed in the real data, a large number of simulations are discarded. Weiss & Von Haessler (1998) addressed these two different but related problems suggesting that more SuSt should be used to better compute the distance between simulated and observed data sets and only the simulations in which the distance between simulated and observed data sets was higher than a specific threshold should be discarded. In particular, Weiss & Von Haessler (1998) used S and k as SuSt, where k is the mean pairwise difference between DNA sequences, and applied the distance threshold to k excluding the simulations where $|k' - k|$ was larger than 0.2 ($| \cdot |$ indicates the absolute value, and the presence or absence of the prime refers to the SuSt in the simulated and real the data sets, respectively). Weiss & Von Haessler (1998) also pioneered the use of simulations and SuSt to compare alternative demographic models, but did not incorporate, as was done a year later by Pritchard *et al.* (1999), the Bayesian step suggested by Tavaré *et al.* (1997).

In synthesis, the most important aspect of ABC which favoured its rapid development is that the likelihood function does not need to be specified. Using ABC, the posterior distribution of a parameter given the observed data, $P(\theta|D)$, can be empirically reconstructed since the likelihood is positively related to the distance between summary statistics computed in real and simulated data sets. More formally, when data are replaced by summary statistics, the reconstructed distribution is

$P(\theta | \rho(\text{SuSt}_{\text{sim}}, \text{SuSt}) \leq \varepsilon)$ (hereafter, $P(\theta | \rho \leq \varepsilon)$), where ρ is any distance metrics between observed and simulated SuSt and ε an arbitrary threshold. In the limit of $\varepsilon \rightarrow 0$ and if SuSt are sufficient (i.e. they capture all the relevant features of the data), $P(\theta | \rho \leq \varepsilon)$ will match exactly $P(\theta | D)$. The idea of ABC is that a good balance between accuracy and efficiency can be reached for small values of ε .

The formal definition of ABC

Beaumont *et al.* (2002) formalized and generalized the ABC approach. They introduced a series of improvements, evaluated the performance of ABC finding a reasonably good agreement with full-likelihood methods under some simple scenarios and discussed in some detail the challenging aspects associated with the choice of SuSt and of the most appropriate distance threshold ε . The actual birth of ABC coincides with this study.

The major improvement introduced by Beaumont *et al.* (2002) is the regression step. Roughly speaking, the slope of the regression line (regression is linear) between a parameter and the vector of SuSt, estimated using the retained simulations (regression is local) and giving more weight to the simulations producing SuSt closer to the observed values (regression is weighted), is used to modify the retained parameters' values and thus mimics a situation in which all simulations produce SuSt equal the observed values. If the chosen ε is very low, the regression step is unnecessary, but the acceptance rate will be very low and a very large numbers of simulations will be required in most cases. Increasing ε , the acceptance rate obviously increases, but in this case the regression step becomes important to improve the approximation of $P(\theta | \rho = 0)$ by $P(\theta | \rho < \varepsilon)$. For multiple SuSt, ρ is usually computed as the Euclidean distance between observed and simulated SuSt. The regression step aims specifically at reducing this discrepancy between simulated and observed SuSt by weighting and adjusting the parameters in the retained simulations, thus requiring fewer simulations. In these circumstances, Beaumont *et al.* (2002) showed that the regression method clearly outperforms the simple rejection algorithm, in which retained parameters are directly used to reconstruct their posterior distribution.

Recently, Leuenberger & Wegmann (2010) reformulated the regression step using the General Linear Model (GLM). SuSt are here response variables with explicit causes within the model, whereas the regression model introduced by Beaumont *et al.* (2002) considered the SuSt as explanatory variables. Some pros and cons of this approach are discussed in the 'Step 8' section. Under a simple one-population model which allows (for comparison) the analytical computation of the

results, the ABC-GLM approach provide a good approximation of the posterior probability of the parameters [i.e. it produces $P(\theta | \rho < \varepsilon)$ close to $P(\theta | D)$], even when the chosen ε was moderately large (Leuenberger & Wegmann 2010).

ABC, MCMC and importance sampling

All simulations are independent under the ABC approach. This means that if a simulated genealogy produces an interesting data-set, i.e. a data-set with SuSt very similar to the observed values, the next simulation can be absolutely useless. In other words, approaching by chance the real values of the parameters during the simulations does not affect the machinery of the method. This sounds inefficient and Marjoram *et al.* (2003) introduced an algorithm to link simulations along a Markov chain path. The parameters for each new simulation are no longer sampled randomly from their prior distributions but are obtained starting from the values used in the previous simulation. The parameter space is explored as in classical MCMC methods, but a substantial difference is introduced. In the Metropolis–Hasting ratio, which is used to decide whether or not to accept a proposed parameter value, the likelihood term is replaced by an indicator function that takes a value of 1 if a simulated data set produces a distance between observed and simulated SuSt below ε and 0 otherwise. As expected, the acceptance ratio and thus the algorithm speed increase, but simulations are not independent any more. One practical advantage of ABC, that simulations for a single analysis can be run on many independent computers and simply pooled at the end, is therefore lost with the introduction of MCMC (but see Wegmann *et al.* 2009 for a possible solution). Embedding the ABC analysis in a MCMC setting raises new problems, some of which are common to any MCMC analysis (e.g. determining the length of the chain, monitoring its mixing and assessing the convergence) and some others are specific of ABC–MCMC. Among the latter, the choice of ε and the definition of the proposal distribution appear crucial to prevent the chain to stick to regions of low likelihood (Sisson *et al.* 2007). Bortot *et al.* (2007) proposed to augment the parameter space by treating ε as an additional parameter and Wegmann *et al.* (2009) introduced a preliminary simulation step to select the threshold ε and to set the proposal distribution.

Additional Monte Carlo schemes, such as population (Cappé *et al.* 2004) and sequential (Doucet *et al.* 2001) Monte Carlo, are under development. Here, importance sampling arguments in various flavours and with various acronyms (ABC-PRC, ABC-PMC, ABC-SMC) are used with the same purpose of MCMC settings to better

explore the parameter space, avoiding the simulation (and the analysis) of unrealistic scenarios (see e.g. Sisson *et al.* 2007; Beaumont *et al.* 2009; Toni *et al.* 2009). Preliminary simulations are used to identify a set of parameters vectors, called *particles*, which are within a certain distance ε from the observed data. The particles are then repeatedly re-sampled (according to a weighting scheme that considers the prior distributions), perturbed (using a transition kernel) and filtered (on the basis of new set of simulations and a decreased threshold ε). The particles after this iterative process tend to converge to a sample from the posterior distribution of the parameters. A final regression adjustment on the retained parameters can be easily applied to all these, as well as MCMC, algorithms (Beaumont *et al.* 2009; Wegmann *et al.* 2009; Leuenberger & Wegmann 2010).

The performances of ABC modified via MCMC or importance sampling have been analysed on simple simulated or real data sets, but the few results available appear controversial. For example, standard ABC, ABC-PMC (ABC with population Monte Carlo, Beaumont *et al.* 2009) and ABC-MCMC (under the Bortot *et al.* 2007 implementation) behave similarly when the computing times are kept identical [Fig. 2 in Beaumont *et al.* (2009)], whereas the ABC-MCMC implemented by Wegmann *et al.* (2009) seems to reach the performances of conventional ABC with a reduction of computational time.

ABC and model selection

Selecting among alternative models under the conventional ABC framework is, at least in principle, even simpler than parameter estimation. The mechanism of the direct method introduced by Weiss & Von Haessler (1998) and Pritchard *et al.* (1999) is straightforward. After pooling all the simulations generated by different models and retaining only those within a distance threshold from the real data, the posterior probabilities of each model is approximated by the fraction of simulations produced by each of them. Accuracy can be very low if the distance threshold ε is not close to 0, but can be improved using the logistic regression approach introduced by Beaumont (2008). The direct and the logistic approaches have been used and compared in various studies (Beaumont 2008; Cornuet *et al.* 2008; Guillemaud *et al.* 2010) and the possible advantages of some recent and more complex alternatives (see Toni *et al.* 2009; Leuenberger & Wegmann 2010) are under investigation.

ABC in nine steps

Here we update, extend and generalize the ABC scheme reported in Excoffier *et al.* (2005). The steps of a

standard ABC analysis, which should be more technically defined as 'rejection ABC', are reported in Fig. 1. Running such ABC analysis rigorously requires careful development of each module, assemblage and validation.

Recently, two implementations of non-standard ABC (using MCMC and PMC) have become available for general users within the set of programs called ABCtoolbox (Wegmann *et al.* 2010). The use of these variants implies the replacement of a specific ABC module, but the general scheme and strategy for the whole analysis does not vary. We have therefore limited our description to the standard ABC.

Step 1: setting the scene

The *model*, i.e. the history and the demography of the populations with the associated parameters together with the genetic parameters relevant for the typed loci, needs to be clearly specified. Unsourced populations can and should be included in the model if they are potentially relevant for the sampled populations. In principle, the complexity of the scenario is not a limiting factor. Almost any demographic event, including migration, colonization, extinctions, divergence, population size changes, mass migrations or translocations, can be easily simulated and thus considered by ABC. Given this opportunity offered by ABC, it is easy to understand why classical population genetics models such as the stepping stone model (Kimura & Weiss 1964) or the divergence-with-isolation model (Wakeley & Hey 1997) appear unrealistic.

The parameters used to specify the model for an ABC analysis are the classic demographic and ecological parameters (e.g. population sizes, migration/growth/admixture rates, carrying capacities), the ages of any sort of natural or human-mediated population event (e.g. population split, translocation, invasion, bottleneck) and the genetic parameters (mutation and recombination rates with associated sub-parameters if needed). Under the hyperprior approach, particularly suitable for situations in which many loci and/or many species are simultaneously analysed (Excoffier *et al.* 2005; Hickerson *et al.* 2006; Beaumont 2008) the parameterization is hierarchical: hyper-parameters define some general feature (e.g. the mean mutation rate at a certain number of microsatellites) and single parameters are defined conditionally. In this way, the parameter space is explored more efficiently and more meaningfully. In principle, even aspects strictly related to the structure of the model, such as the size of river segments (see Neuenschwander *et al.* 2008) or the number of populations, can be defined as parameters to be estimated. This approach can be useful especially in

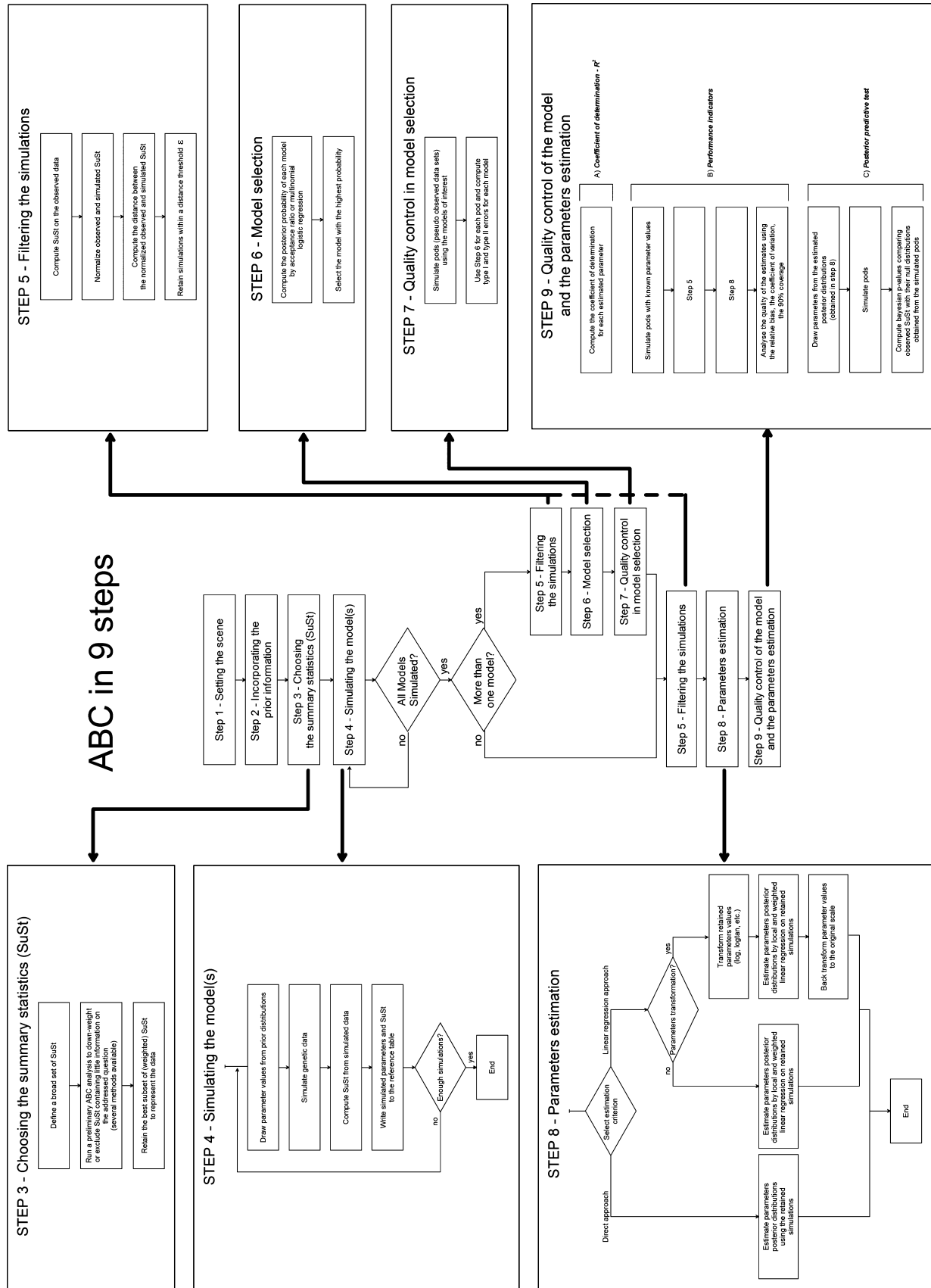


Fig. 1 ABC in nine steps.

preliminary ABC tests to identify and then fix, some aspects of the model.

Depending on the available data and on the relative influence of different events and parameters on the genetic variation pattern, increasing the complexity of the model can be either useless, time consuming or result in poor exploration of the parameter space. Some simple reasoning or preliminary simulations can be used to better understand the impact of different aspects of the model on the genetic variation pattern and, consequently, used to fix the value of some parameters and/or simplify the model (e.g. Estoup & Clegg 2003; Pascual *et al.* 2007; Ludwig *et al.* 2008). We believe, however, that ABC should be, at least at the beginning of the analysis, used at its maximum potential power, i.e. devising realistic models. A detailed analysis of the results, also comparing different runs (e.g. Hickerson *et al.* 2006), will help to determine their robustness.

Clearly, if different models are going to be compared, all of them need to be defined at the outset. Models can be nested or non-nested and it may be also interesting to compare the differences between the results provided by ABC under realistic and simplified versions of a model.

Step 2: incorporating the prior information

As in typical Bayesian settings, prior information can and should be incorporated in the ABC analysis. Prior beliefs regarding the parameters and the models (if different models are compared) will be used to modify the information contained in the data to obtain the posterior distributions. These beliefs are incorporated in standard ABC analyses in the simulation step (Step 3), i.e. when the parameters used to simulate each genetic variation data will have values sampled from their prior distributions and the number of simulations performed using each of the different models will be proportional to the prior probability assigned to each model.

Prior distributions should be obviously large enough to include all the values which are considered at least possible and their shape may well vary among parameters. For example, mutation rates are often sampled from gamma distributions, migration rates from exponential distributions, times from uniform or exponential distributions and population size from uniform or log-uniform distribution. These choices may reflect previous knowledge on some parameters (e.g. the gamma distribution usually fit well real mutational data) and/or the need to homogeneously sample the parameter across different orders of magnitude (e.g. migration rates between 10^{-5} and 10^{-1}). Of course, if the value of a parameter is known with relatively high precision [e.g.

the starting time of an invasion (Pascual *et al.* 2007)], the parameter should be fixed in the simulations. When different scenarios are compared, they are usually considered with the same prior probability.

Sometimes, prior distributions are slightly modified if 'first shot' simulations produce data sets very different from the observed data. This strategy can be necessary in some circumstances, it can be regarded (Gelman 2008) as a test of prior beliefs when combined with appropriate quality controls (see steps 7 and 9), but it should be honestly and carefully adopted. There is clearly a potential difficulty in using the data twice, both for estimation and to 'refine' the priors and the resulting posterior distributions will not be 'true' Bayesian combinations of prior beliefs and likelihoods.

There are clear computation and logical advantages in using prior distributions and the Bayesian approach compared for example, to maximum likelihood methods, even when the prior knowledge is very limited and consequently flat and wide prior distributions are used (Huelsenbeck *et al.* 2001; Holder & Lewis 2003; Beaumont & Rannala 2004). However, we believe that more efforts should be dedicated to identifying information to incorporate with confidence in the prior distributions, using previous genetic or non-genetic studies. These efforts can be facilitated by the hyperprior approach whereby at least the hyperprior distributions can be narrowed. For multilocus microsatellite data for example, the mean and the variance (the hyper-parameters) of the mutation rates are reasonably well known, whereas the single-locus rates are not. Incorporating robust prior beliefs will produce more accurate and precise estimations and it will also facilitate the interpretation of the results. When prior definitions are based on vague information, the effects of errors in prior beliefs can, and should, be efficiently investigated with a sensitivity analysis within the ABC framework (e.g. Pritchard *et al.* 1999; Estoup *et al.* 2001; Hickerson *et al.* 2006; Verdu *et al.* 2009; Guillemaud *et al.* 2010).

Step 3: choosing the summary statistics

The whole ABC machinery is based on the comparison between observed and simulated data sets and this comparison is made after reducing data sets to summary statistics, SuSt. Unfortunately, there is still no general rule as to which and how many SuSt should be used, although the importance of this step was already recognized since the formal introduction of ABC (Beaumont *et al.* 2002; Marjoram *et al.* 2003). The selected SuSt should be able to capture the relevant features of the data. Ideally, SuSt should be *sufficient*, i.e. the posterior probability of a parameter given these SuSt

should be the same as its posterior distribution given the complete data set (Marjoram & Tavaré 2006). In practice, a SuSt should not be included in this set if it does not provide any additional information about the data useful for the estimation process. Easy to say, but very difficult to realize. The sufficiency of a set of SuSt is strictly dependent on the model, parameters and data, meaning that some preliminary analysis is required.

A single or few SuSt are almost always a very crude representation of the data and likely produce biases in ABC analyses (Marjoram *et al.* 2003). On the other hand, too many SuSt (especially those providing little information regarding the parameter being estimated) introduce stochastic noise, reducing the fraction of retained simulations and increasing the errors both when the distance between observed and simulated data sets is estimated and during the regression step (Beaumont *et al.* 2002). More than 100 SuSt were used by Rosenblum *et al.* (2007) to reconstruct the historical demography of a lizard colonization process, but the usual numbers of SuSt in published empirical studies range between 5 and 20.

When several loci are typed, SuSt are usually means and variances of single locus statistics (e.g. Ross-Ibarra *et al.* 2009) or indices correlated to the shape of the distribution of phenotypes (e.g. AFLP data, see Foll *et al.* 2008) or allele (e.g. SNP data) frequencies. At least three methods, not yet implemented for practical use, have been suggested to identify the best set of SuSt. Hamilton *et al.* (2005) used the determination coefficients between each SuSt and each parameter, estimated from a set of preliminary simulations, to weight differentially the SuSt. The distance between observed and simulated data sets is thus computed separately for each parameter. This is not the same as selecting a subset of SuSt, but is a criteria to avoid this selection and to almost exclude by weighting some SuSt from the estimation process. Joyce & Marjoram (2008) have introduced a 'sufficiency' score to be assigned to each SuSt in a sort of preliminary experiment. The whole ABC estimation step is performed several times adding and removing different SuSt and retaining only those that significantly modify the posterior distribution of the parameter of interest. Wegman *et al.* (2009) suggest extracting a limited number of orthogonal components, appropriate to explain the parameters variation, from a large number of SuSt. These new variables, estimated by a partial least square regression approach with coefficients estimated on the basis of a set of preliminary simulations, are then used as SuSt. So far, only modest advantages of these approaches have been demonstrated. Considering the actual state of the art, we recommend a selection of the SuSt known to be informative about the

parameters of interest, an appropriate number of simulations (see below) and, in particular, some preliminary tests showing that the selected SuSt can be used to reasonably recover models and parameters in data sets simulated under scenarios relevant for the addressed question (see e.g. Becquet & Przeworski 2007; Rosenblum *et al.* 2007; Neuenschwander *et al.* 2008).

Step 4: simulating the model(s)

A large number of data sets should be simulated under the model(s) defined at Step 1, with each simulation using a different set of parameter values sampled from the corresponding prior distribution. Simulation is the time-consuming step, but an important advantage of standard ABC (but not of ABC coupled with MCMC or importance sampling) is that the data sets generated by simulation can be used for estimation or model selection on many different data sets. The simulated data sets, which are commonly reduced to the values of the chosen SuSt due to disk space limitations, are stored in the *reference table*. The same reference table can then be used for inference on the real data sets but also, for example, on pseudo-observed data sets, or *pods*. Pods are specific data sets generated with known parameter values by simulation and are very useful for investigating the bias/accuracy of the analysis (see steps 7 and 9). The reference table is therefore very valuable, both because it usually takes lot of computing time to generate it and because it will be recycled several times. It seems thus a good idea to select accurately the software for the simulations most appropriate for the scenario and genetic markers of interest, to avoid hurried decisions about the prior distributions and not to economize on the number of simulations.

In principle, both backward coalescent and forward classical simulations of the genetic data can be used for ABC. In practice, only the former seem to have, today, the required time-efficiency. Forward genetic simulations have the advantage to substantially simplify the implementation of natural selection and for this reason they may spread for specific ABC implementations (e.g. Itan *et al.* 2009). We expect that efficient forward genetic simulator (Chadeau-Hyam *et al.* 2008; Hernandez 2008; Carvajal-Rodriguez 2010) coupled with ABC will be used in the near future to analyse complex scenario involving both selective and demographic processes.

The available coalescent simulation programs, reported in Table 1 with their main characteristics, can be classified in two major groups: ABC integrated and ABC independent simulators. ABC integrated simulators are assembled within a larger package designed to perform all the ABC analyses. These user-friendly pack-

Table 1 Features of the main online backward coalescent simulators available for ABC analysis. Other software are developed for specific purposes and available upon request to the authors (see e.g. Przeworski 2003)

Name	Type of markers	Demographic model				Recombination	Selection	Serial sampling ¹	Consider explicitly spatial and environmental heterogeneity	ABC integrated	Reference
		One/many populations	Population divergence	Migration	Change in population size						
MS	DNA sequence	Many	Yes	Yes	Yes	Yes	No ²	No	No	No	Hudson (2002)
Simcoal2	RFLP, STR, DNA sequence, SNP	Many	Yes	Yes	Yes	Yes	No	No	No	No	Laval & Excoffier (2004)
Selsim	STR, DNA sequence	One	No	No	No	Yes	Yes	No	No	No	Spencer & Coop (2004)
SPLATCHE	RFLP, STR, DNA sequence ³	Many	Yes	Yes	Yes	No ³	No	No	Yes	No	Currat <i>et al.</i> (2004)
Bayesian Serial Simcoal	RFLP, STR, DNA sequence	Many	Yes	Yes	Yes	No	No	Yes	No	No ⁴	Anderson <i>et al.</i> (2005)
AQUASPLATCHE	RFLP, STR, DNA sequence, SNP	Many	Yes	Yes	Yes	No	No	No	Yes	No	Neuenschwander (2006)
msBayes	DNA sequence	Many ⁵	Yes	Yes	Yes	Yes	No	No	No	Yes	Hickerson <i>et al.</i> (2007)
DIY ABC	STR, DNA sequence	Many	Yes	No	Yes	No	No	Yes	No	Yes	Cornuet <i>et al.</i> (2008) ⁶
ONEsAMP	STR	One	No	No	No	No	No	No	No	Yes	Tallmon <i>et al.</i> (2008)
PopABC	STR, DNA sequence	Many	Yes	Yes	No	Yes	No	No	No	Yes	Lopes <i>et al.</i> (2009)
ABCtoolbox ⁷	RFLP, STR, DNA sequence, SNP	Many	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes/No ⁷	Wegmann <i>et al.</i> (2010)

¹Samples with different ages can be simulated (relevant if ancient DNA data or time series are available).²MS has been modified by Jensen *et al.* (2008) to include selection.³The new version of SPLATCHE, including recombination and SNP markers, will be available soon (L. Excoffier, pers. comm.).⁴Data can be by simulated sampling parameter values from prior distributions.⁵Only a vicariance model can be simulated.⁶The new version of DIYABC is available online.⁷ABCtoolbox is a collection of independent command-line programs which facilitate the development of a pipeline to estimate model parameters and compare models; several external simulation programs can be pipelined.

ages have obvious advantages since the complete ABC analysis can be accomplished using a single program. Users considering these packages should, however, realize that very complex models with specific prior distributions of the parameters, as well as some kind of genetic markers, may not be simulated. Similarly, other steps of the analysis are constrained to specific functions, whereas the ABC is by nature almost like *bricolage*, requiring frequent small adjustments (including suggestions coming from new studies) and specific tests for different data-sets. ABC non-integrated simulators are independent programs simulating genetic variation patterns. The best choice is probably to look initially at the general ABC packages with integrated simulators (see Table 1), figure out if the models and markers of interest can be simulated, and, if not, find out if the authors will release an updated version soon (the field is moving fast). However, if some experience with programming and script development is available (e.g. in R, Python, or C++), choose an ABC independent simulator. MS (Hudson 2002) and Simcoal2.0 (Laval & Excoffier 2004) are widely used ABC independent simulators. For models which consider explicitly spatial or environmental heterogeneity, i.e. where large numbers of demes and their migration/colonization relationships through time and space are assumed, Splatche (Currat *et al.* 2004) or Aquasplatche (Neuenschwander 2006) are more appropriate. Serial SimCoal (Anderson *et al.* 2005) can be used if ancient DNA data are available. All these ABC independent simulators are very flexible and allow access to the code, but need of course to be 'pipelined' within all the other steps of the ABC analysis. The recent introduction of a series of programs within a single ABC tool box (Wegmann *et al.* 2010) will likely alleviate this problem in the future.

Finally, we have to address another question with no general answer: how many simulations? Empirical studies seem to converge towards the order of magnitude of 10^6 . Clearly, the complexity of the model and the dimensions of the parameter affect the number of simulations necessary to explore them. Our view is that some preliminary simulations, for example testing the convergence by comparing the results obtained in a few independent analyses with 10^4 – 10^5 simulations, can be very useful. In some cases, for example in the relatively simple scenarios analysed by Guillemaud *et al.* (2010) even 10^4 simulations appear sufficient to reach accuracy in model selection. However, if alternative versions of ABC (for example ABC-MCMC) are not considered, brute power rather than style is the main feature of ABC analyses. We suggest therefore using large CPU clusters (now relatively cheap and commonly available in computing departments) and performing several millions of simulations in the final analysis for both model

selection and parameter estimates. The current development of specific implementations of genetic analyses using graphics processing units (GPU) (e.g. Suchard & Rambaut 2009) will possibly reduce the need for large clusters soon.

Step 5: filtering the simulations

Simulations are retained when a multivariate distance between observed and simulated SuSt is below a certain distance threshold. In general, a simple Euclidean distance is computed on normalized SuSt and the threshold is defined such that a small fraction of the simulations (0.1–3%), corresponding to the smallest distances, are retained for the estimation step. A conditional threshold (e.g. Thornton & Andolfatto 2006; Putnam *et al.* 2007) can also be devised, implying that simulations (see step 3) are repeated until a certain number (in the order of 10^3 – 10^4) of accepted simulations is reached. The distance threshold can be different in model selection (step 6) and parameter estimation (step 8). As underlined by Guillemaud *et al.* (2010), the choice of the threshold should be always validated. Pods should be simulated under a model (or models) relevant for the question addressed and the threshold producing reasonably stable and accurate reconstruction of the known scenarios should be adopted. In any case, it is a good idea always to check the effect of using different threshold values on the distribution of Euclidean distances, on the comparison between observed SuSt (separately or combined for example using PCA) and the corresponding simulated distributions and, obviously, on the estimated posterior distributions of the parameters and the models.

Step 6: model selection (if different models are compared)

In step 6 of ABC, some results are obtained at last. Comparing models, which actually means comparing alternative hypotheses about a process, is the key to the work of scientists. The ABC framework allows the computation of the relative weight, i.e. the posterior probability, of different hypotheses (i.e. different models). Probably for the first time, genetic variation data can be used not only to reconstruct a plausible historical or demographic scenario, often combining many different analyses and tests, but also to assign a quantitative 'belief score' to each of many alternative and possibly complex scenarios. ABC *should* also favour a reduction in the length of manuscripts, since elaborated arguments supporting or opposing each hypothesis can be summarized by a corresponding set of meaningful probability scores, the sum of which is always equal to

one. The fear is that this reduction will be compensated by extensive, technical and somewhat boring description of ABC options and validation, but general readers will be able to skip these sections and easily comprehend the results and main conclusions.

All the models are generally simulated the same number of times. This is equivalent to giving the same prior probability to each model under comparison and zero probability to any other model. Clearly, errors in the latter assumption may produce incorrect conclusions regarding the models support (see e.g. Templeton 2009), but the ABC framework allows for the evaluation of the effects of excluding some models in the specific situation under investigation (Guillemaud *et al.* 2010). In the final set of retained simulations, the data sets produced by the more probable models will be over-represented and the data sets produced by the less probable models will be under-represented or even absent. Intuitively, the probability of a model is proportional to the relative frequency of the data sets it produces that are among the retained simulations (Weiss & von Haeseler 1998; Pritchard *et al.* 1999). This frequency is actually the direct estimator of the posterior probability of a model, but this estimator is rarely accurate in complex scenarios when, inevitably, the retained simulations are either too few or also contain data sets not closely matching the observed data. Recently, Leuenberger & Wegmann (2010) proposed the use of a parametric General Linear Model to adjust the model frequencies in the retained simulations. However, the most reliable and tested method, also available in ABC packages such as DIYABC (Cornuet *et al.* 2008), is still the adjustment based on the weighted multinomial logistic regression introduced by Beaumont (2008). The coefficients for the regression between a model indicator (response) variable and the simulated SuSt (the explanatory variables) can be estimated, allowing the estimation of the posterior probability for each model at the intercept condition where observed and simulated SuSt coincide. CIs of the probabilities can be computed as suggested by Cornuet *et al.* (2008).

The posterior probability of each model is of course an intuitive score of our belief in that model. An additional index, comparable with a standard table of reference values where the evidence is assigned to categories from 'not worth more than a bare mention' to 'decisive', is the Bayes factor. The Bayes factor can be easily computed in an ABC analysis, being the ratio between the posterior probabilities estimated in any pair of models, divided by the ratio of their prior probabilities. The latter ratio is of course equal to one if all models have the same prior probability. This index is a summary of the evidence provided by the data if favour of a model as opposed to another and it can be inter-

preted as a measure of the relative success of the models at predicting the data (Kass & Raftery 1995). The Bayes factor is also the ratio of the posterior odds to the prior odds, meaning that it actually measures the change of relative probabilities of the various scenarios tested in the ABC analysis due to the knowledge obtained from the genetic data.

Step 7: quality control in model selection

The ABC framework can be used to investigate the robustness of model selection and parameter estimation with relatively little additional effort (e.g. Fagundes *et al.* 2007; Guillemaud *et al.* 2010). Data sets simulated under specific scenarios with known parameter values are tested against the same reference table (the large number of simulated data sets, see step 4) used in the analysis of the real data set.

Some hundreds of pseudo-observed data sets (the pods, see step 3) are generated using each of the scenarios considered in the model selection analysis. Obviously, other scenarios can be analysed to investigate the effects of incomplete model specification on the inference (see step 6). The values of the parameters used for generating pods are generally restricted to the best estimates obtained from the analysis of the real data, but they can be other values of interest. Pods generated with fixed parameters will provide information about the quality of the estimated model probabilities which is restricted to a specific parameter set. The ability of ABC to identify the correct model in a larger space of parameter values can be analysed by generating pods using parameter values sampled from, for example, the prior distribution (Fagundes *et al.* 2007; Cornuet *et al.* 2008; Verdu *et al.* 2009) or the posterior distributions estimated from the real data set.

Even if the definitions here are not rigorous, type I and type II errors can be estimated for each scenario, using, in turn, each scenario as the null or alternative hypothesis. In practice, the type I error for, say, scenario A is estimated as the fraction of pods generated under scenario A that support other scenarios, whereas the type II error for scenario A is estimated as the fraction of pods generated under all the other scenarios that support scenario A. A single pod is considered to be supporting a scenario simply if the posterior probability of this scenario is the largest. So, these are not really type I and II errors in the classical frequentist framework, whereby the null hypothesis is never accepted and is rejected only if the data are manifestly incompatible with it. These estimated errors can be very useful when small, but otherwise their joint interpretation may not be straightforward. Some additional insight into the accuracy and power of the analysis can be obtained by

computing mean and standard deviation of the posterior probability of each model using the probabilities estimated in the pods or the frequency of pods where the CI of the model with the highest probability does not overlap with the CI of the next supported model (see Guillemaud *et al.* 2010). Pods could be also used to estimate the distribution of the Bayes factor for each simulated scenario also and thus to better interpret the Bayes factor computed from the real data set.

Step 8: parameters estimation

For the single model considered in the analysis or for the most supported model if different models are considered, the posterior distributions of the parameters can be reconstructed by ABC. In many cases it is a good idea to start looking at the estimated distribution of the hyper-parameters or the composite parameters, the latter obtained by combining, in each simulation, the parameters which are difficult to estimate separately (for example, the population-mutation parameter $N\mu$ which combines the population size N and the mutation rate μ).

As already explained, retained simulations have data sets closer (but not identical) to the real data than do non-retained simulations. Therefore, using the parameter values of the retained simulations as a sample from their posterior probability distribution (the direct approach), still maintains an undesirable component of the prior. If all simulations are retained, i.e. with a threshold of tolerance equal to infinity, the prior will be recovered. At the other extreme, when the threshold is proximate to 0, the direct approach works well, but huge numbers of simulations are needed to obtain a reasonable sample size from the posterior. We can imagine that in the near future, especially if different groups will share their reference tables, billions of simulations will be available for the direct method. In the meantime, the most commonly used method to adjust the imperfect retained simulations is the local linear weighted regression introduced by Beaumont *et al.* (2002). The coefficients of a linear regression between each parameter and the vector of the chosen SuSt are estimated from the retained simulations (the local aspect) assigning to each point a weight based on a function increasing as the distance between the observed and simulated data sets decreases (the weighting aspect). The regression slope is then used to adjust each parameter value from the retained simulations towards the value expected in correspondence with the observed SuSt. The intercept corresponds to the posterior mean estimate of the parameter. This approach, which can be applied to all the parameters simultaneously, assumes local linearity between parameters and SuSt (see Blum & Francois 2009

for an extension to non-linear regression models), additivity and a multivariate normal distribution. However, its use in the last 8 years after the original development (Beaumont *et al.* 2002) suggest that small violations of these assumptions only marginally affect the results. The accuracy of the posterior distributions, when evaluated under simple scenarios which allows also the use of full-likelihood methods, is drastically increased by the regression step compared to the direct approach (Beaumont *et al.* 2002; Leuenberger & Wegmann 2010). The commonly-used weighting function is the Epanechnikov kernel, but the effects on the final estimates of applying other weighting schemes are probably limited. Parameters are usually transformed before the regression step (and then back transformed after it) by a log (e.g. Estoup *et al.* 2004; Hamilton *et al.* 2005; Crestanello *et al.* 2009), logtan (e.g. Kayser *et al.* 2008; Ross-Ibarra *et al.* 2008) or logit function (e.g. Cornuet *et al.* 2008). Logtan and logit functions avoid adjustments outside the prior distribution.

The general linear model (GLM) approach recently proposed by Leuenberger & Wegmann (2010) and implemented in ABCtoolbox (Wegmann *et al.* 2010) can be used as an alternative method to estimate the posterior distributions from the retained simulations. Additional testing is however necessary to identify the best adjustment procedure under different conditions, since GLM have both advantages (it considers the correlation among SuSt and never produces posterior distributions outside the priors) and disadvantages (it assumes normal distributions for the SuSt and is computationally more intensive) compared to the earlier approach.

Of course, when a sample from the posterior distribution is available, point estimators and a relative measure of accuracy are needed. Usually, a smoothed-posterior density is fitted to the sample of adjusted parameter values using specific methods (e.g. local-likelihood) and after specifying a bandwidth. This fitting step is embedded in the DIYABC package (Cornuet *et al.* 2008), but we suggest analysing the rough-frequency distribution of adjusted parameters to identify possible distortions introduced by the fitting algorithm.

The point estimators usually computed from the posterior densities are the mean, the mode, the median and the intercept estimated in the regression step. No consensus has been reached about the point estimator with the smallest bias and variance and, as usual, the analysis of simulated data sets relevant for the scenario of interest can be useful. In general, however, the differences between point estimators are quite small if compared with the width of their posterior distribution and their choice is therefore almost irrelevant. Most importantly, the posterior density can be used to compute the confidence of the estimates. Typical measures are the

SD and the credible interval, the latter being the Bayesian equivalent of the frequentist CI. A commonly used credible interval is the X% highest posterior density or HPD. The X% HPD interval is the interval which includes the X% of the parameter values and within which the density is never lower than the density outside it. Typically, 90 or 95 HPD limits are reported in ABC analyses (e.g., Fagundes *et al.* 2007; Ludwig *et al.* 2008; Ray *et al.* 2010).

Step 9: quality control of the model and the parameters estimation

The quality of the parameter estimates can be initially evaluated by the proportion of parameter variance that is explained by SuSt (see e.g. Fagundes *et al.* 2007; Neuenschwander *et al.* 2008). This is the classical coefficient of determination. If only a small fraction of parameter variation in the reference table can be explained by variation in SuSt, it is hard to imagine that the parameter will be accurately estimated for the model(s) under consideration and the number of individuals and markers typed. It is possible that different SuSt might improve the estimation (and this hypothesis can be tested), but it is also possible that the parameter cannot be estimated for that model/data package even if the full likelihood could be computed. The coefficient of determination should be taken with caution. Even a small fraction of the explained variation can be sufficient for reasonably precise estimates given enough data. In analogy, population assignment of single individuals can be quite accurate even when a small fraction of the variation is attributed to between-population differences (e.g. Latch *et al.* 2006; Colonna *et al.* 2009).

A more important evaluation of the quality of parameter estimates under the specific scenario (or scenarios) under investigation can be performed within the ABC framework in exactly the same way we described for model selection: generating pods, i.e. simulating data-sets using known parameter values or parameter distributions and analysing them (e.g. Excoffier *et al.* 2005; Jensen *et al.* 2008). The best estimates of the parameters obtained in step 8 (or their posterior distributions) are of course interesting candidates for generating pods in this analysis. For each pods, parameters are estimated using the same reference table and the same procedure applied to the real data and are then compared to the true known values used to generate them. In fact, after the analysis of, say, 1000 pods, 1000 posterior densities will be available for each parameter. From each of these distributions, a point estimator (e.g. the mode) and a credible interval (e.g. the 90% HPD) can be computed and several measures of the estimator quality can be estimated (see e.g. Cornuet *et al.* 2008) by simply com-

paring these 1000 point estimators and credible intervals with the true value of the parameter (i.e. the value used in the simulations). The relative bias, the coefficient of variation and the 90% or the 50% coverage (which are the fraction of 90% or 50% HPD intervals in the 1000 simulations which include the true value), are usually informative to ascertain the quality of the estimates. The analysis and interpretation of other highly-correlated measures is likely useless and confounding. Some caution is also needed in general for the interpretation of these performance measures. For example, a relative bias of 1 (100%) when estimating a true population size of 1000, meaning that on the average the estimated value will lay at a distance of 1000 individuals from the true value, would appear enormous. But if the prior knowledge on this parameter was entirely missing and a uniform prior distribution ranging between 100 and 100 000 was defined, such bias should be considered small. If possible, it is always very instructive to estimate at least some parameters from the pods using other non-ABC approaches and then compare the quality measures across methods (Guillemaud *et al.* 2010). A large bias or variance using ABC can become acceptable in comparison with the performance of other methods.

A third way to investigate the quality of the ABC results is to compare the SuSt observed in the real data with the posterior distribution of SuSt (e.g. Pascual *et al.* 2007; Ingvarsson 2008). The posterior distribution of SuSt is the SuSt distribution computed from pods generated by simulation with parameters values sampled from their estimated posterior distribution (Gelman *et al.* 2004). The rationale of this comparison, which is testing the goodness-of-fit of the combination 'scenario + posterior distributions of the parameters' to the data, is simple: if the estimated parameters under a specific model have anything to do with what happened in the history of the real samples, then histories simulated using these values should produce pods similar to the data. If this is not the case, either the parameter estimation is bad and/or the model is wrong. Using a simple graphical inspection, the goodness-of-fit should be considered high if the distance between observed SuSt and the SuSt in their posterior distribution is low. A principal component analysis of the SuSt in the real data set, the SuSt from their posterior distribution and also the SuSt in the reference table can provide additional insights on the quality of the estimation (Guillemaud *et al.* 2010; A. Estoup, pers. comm.). Also, the performance of the estimation can be quantified by computing bias and variance, relative to the observed SuSt, of the posterior distributions of SuSt (see e.g. Neuenschwander *et al.* 2008).

The comparison between observed SuSt and their posterior distributions is also the basis for a posterior

predictive test, which is the Bayesian analogous of the parametric bootstrap under the frequentist framework (Gelman *et al.* 2004). The goodness-of-fit of the inferred combination 'scenario + the posterior distribution of the parameters' and the observed data is quantitatively measured by the posterior predictive P value. This Bayesian P value corresponds to the probability that data replicated using the estimated posterior distributions of the parameters are more extreme than the observed data (Gelman *et al.* 2004). It can be specific for each SuSt (e.g. Thornton & Andolfatto 2006) or appropriately combined across SuSt (e.g. Ghirotto *et al.* 2010) and it can be also viewed as the probability of observing a less good fit between the model and the data.

A posterior predictive test is a good practice in any Bayesian model-based analysis, since it is the most straightforward way to understand if the estimated parameters are at least meaningful. In general, however, its power to reject the null hypothesis under reasonable population genetic condition and particularly when few loci are analysed, is limited. Nonetheless, this test can be useful to identify deviant SuSt with significant P values, possibly related to specific poorly estimated parameters or erroneous aspects of the demographic model. The use of posterior predictive tests in ABC is therefore recommended (e.g. Becquet & Przeworski 2007; Ingvarsson 2008; Neuenschwander *et al.* 2008; Ghirotto *et al.* 2010).

Applications

The number of published applications of ABC to genetic variation data increased rapidly following the formal definition of the methodology in 2002 (Beaumont *et al.* 2002), doubling for example between 2007 and 2008. A bibliographic Endnote list of 107 papers on ABC, with about two-thirds of them presenting applications to real data, is provided as 'Supporting information'.

Approximate Bayesian computation has been applied to very different types of organisms, from bacteria (e.g. Luciani *et al.* 2009; Wilson *et al.* 2009) to plants (e.g. Francois *et al.* 2008; Ross-Ibarra *et al.* 2009) and animals (e.g. Voje *et al.* 2009; Lopes & Boessenkool 2010). Microsatellite markers are the most commonly used source of genetic information (43% of the studies), followed by nuclear and mitochondrial DNA sequences (each of them analysed in about 30% of the studies). DNA data from ancient samples is included in about 10% of the studies. The number of loci varies widely among papers, but the median value for STR and nuclear sequences is 9 and 19, respectively.

After surveying the many options available when running an ABC analysis, we outline in the 'Supporting

information' the main trends. In general, we estimated that if all the steps discussed in the previous section were applied, about 60% of the published ABC applications would have significantly improved their robustness. As positive examples of studies in the field of molecular ecology where, in our opinion, the ABC framework was properly used to estimate parameters, to compare models and to evaluate the quality of the model settings and the results, we would like to mention Neuenschwander *et al.* (2008) and Guillemaud *et al.* (2010). Neuenschwander *et al.* (2008) reconstructed the dynamic of the post-glacial colonization of a river basin in Switzerland by the European bullhead (*Cottus gobio*). Guillemaud *et al.* (2010), after extensively investigating the capabilities of ABC in reconstructing different aspects of an invasion process, applied this method to investigate alternative scenarios for the introduction to Europe of the North American pest of corn *Diabrotica virgifera*.

Using a subset of 14 relatively homogeneous studies and 152 parameter estimates, we also compared prior and posterior distributions to obtain some general indications about the fraction of uncertainty about a species that was reduced by ABC using genetic information. The difference in width and location between prior and posterior distributions was not clearly related to any general features of the data sets, the model or the ABC setting (such as the number of loci, the number of sampled individuals, the number of parameters to be estimated and the number of SuSt). If applied to a single study, this result would appear counter intuitive, since increasing for example the number of markers should narrow the credible intervals (Excoffier *et al.* 2005). The power of our analysis is clearly limited, but it is also possible that large differences in the informativeness of the data sets and in the complexity of the scenarios across studies blurred the expected pattern. At any rate, our analysis of more than 150 posterior distributions seems to confirm the reasonable idea that guidelines regarding the number of individuals and markers to analyse, the maximum allowed complexity of the model and the number of SuSt sufficient to summarize the data, cannot be easily identified. Such guidelines can be very specific only for the process and species of interest, meaning that the set of preliminary simulations described throughout this paper can be very useful to plan the sampling, the typing and the final ABC setting. Additional results and details of this analysis are provided as 'Supporting information', Table S1 and Fig. S1.

Conclusions

Approximate Bayesian computation has a short history and very likely a long future. Molecular variation data

are useful to reconstruct past events, estimate biological parameters and compare alternative scenarios. The ABC approach has the potential to become a standard approach in molecular ecology, as well as in other fields (Lopes & Beaumont 2010; Lopes & Boessenkool 2010). It allows, for the first time, the efficient exploitation of the enormous progresses in genetic typing and computing speed to investigate very complex population models including both natural and human-induced processes.

Throughout this paper we have summarized the theoretical and practical aspects of this methodology, which should not be considered as a statistical analysis *per se* but rather as a statistical framework. We also briefly analysed its behaviour reviewing the published applications. Overall, we tried to stimulate the general reader to consider ABC as a possible instrument for analysing their data and also for planning sampling and typing strategies. Many pros and cons, together with some practical suggestions, were given. We schematically recall and integrate them in this final section.

Complex and specific models can be analysed

The likelihood of models and parameters does not need to be theoretically derived. We believe that, to a reasonable extent, initial investigations under the ABC framework should always take advantage of this quality and challenge the data using very realistic models. Our analysis of empirical studies suggests that even a few markers can be useful to substantially increase the knowledge about the process of interest. This is possible in many cases by limiting the inference to well-estimated composite and hyper-parameters. Importantly, ABC provides the instruments to understand if the set data + models can be used or not to reconstruct the most likely scenario and if the estimate can progress from composite to single parameters.

Quality of estimates and model selection can be measured

After becoming familiar with the ABC framework, quality controls and power analyses can be performed with rather limited additional effort. The simulation results, stored in a single reference table, allow the analysis of the real data set as well as many other simulated data sets of interest and the feasibility of an accurate estimation or a model selection can be analysed. It is recommended that large reference tables stored by different research groups become accessible, possibly promoting a shared repository, since some preliminary analysis before the sampling and typing using simulated data sets or real data sets with properties (sample sizes, number and type of loci, etc.) matching as closely as

possible the data sets in the reference table, could be very useful. Interestingly, all these analyses in specific ABC settings for different species and questions will also help in understanding the general properties of the method. Unfortunately, quality control is more difficult under the ABC-MCMC and related serial approaches, since the dependency among simulated data sets prevents the compilation of a reference table.

Difficulties in performing an ABC analysis are decreasing

The conceptual scheme of ABC is quite simple and modular, implying that end users who do not find enough flexibility in complete ABC packages [e.g. DIY-ABC, Cornuet *et al.* (2008)] can develop specific implementations using the appropriate simulator (e.g. MS) pipelined with relatively simple algebraic or statistical computations. The post-simulation analyses are already coded in available and easily modifiable R scripts or C/C++ programs. A great collection of command-line modules useful for an ABC analysis, which includes also samplers based on MCMC (Marjoram *et al.* 2003) and PMC (Beaumont *et al.* 2009), is also available in the ABCtoolbox (Wegmann *et al.* 2010). Important ABC implementations for estimating for example selection coefficients or individual-based parameters will likely imply the development of more efficient simulators (see e.g. Przeworski 2003; Jensen *et al.* 2008; Leblois *et al.* 2009).

Probabilities are approximated

More studies are needed to better evaluate the degree of approximation of ABC estimates compared to full-data likelihood methods. These studies, however, are restricted to scenarios where explicit likelihoods functions are tractable. Fortunately, the ABC approach has the potential to be used, case by case and if properly implemented, as a self-evaluator of its performances under known simulated scenarios. This potential, clearly related to the fact that an ABC application is actually a hybrid between a simulation and a data-analysis study, can be used to understand general questions regarding, for example, the ability of ABC to select the true among many simulated models even when the parameters of the model are poorly estimated.

Time and patience are required

It is very important to realize that performing an ABC study is not at all like using other methods for the analysis of genetic variation data. Even if questions of interest can be addressed using a single complete package such

as DIYABC (Cornuet *et al.* 2008), adequate amount of time should be dedicated to initially define the model(s) and then decide among the many alternative options. Crucial steps such as deciding on the number of simulations, the SuSt and the acceptance threshold cannot be based on general rules. The effects of these choices and the performances of the estimates should be evaluated and tested in each study. Fortunately, even if planning a rigorous ABC analysis in all its steps, as well as modifying the initial plan when needed, requires time and patience, computer speed (for example exploiting hundreds of CPUs and possibly, in the future, GPUs) and data storage possibilities (terabytes are very cheap today) are not a limiting factor in many studies. Elegant methods to efficiently reduce the number of relevant simulations needed for the estimation process, such as ABC coupled with MCMC, are under development and testing, but their standardization and spread to non-experts might not be rapid.

Acknowledgements

We thank Arnaud Estoup and two anonymous reviewers for useful and detailed comments and suggestions. This research was supported by the Italian Ministry for Research and Education and by the University of Ferrara, Italy.

References

- Anderson CN, Ramakrishnan U, Chan YL, Hadly EA (2005) Serial SimCoal: a population genetics model for data from multiple populations and points in time. *Bioinformatics*, **21**, 1733–1734.
- Beaumont MA (2008) Joint determination of topology, divergence time and immigration in population trees. In: *Simulation, Genetics and Human Prehistory* (eds Matsamura S, Forster P, Rrenfrew C), pp. 135–154. McDonald Institute for Archaeological Research, Cambridge, UK.
- Beaumont MA, Rannala B (2004) The Bayesian revolution in genetics. *Nature Review Genetics*, **5**, 251–261.
- Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.
- Beaumont MA, Cornuet J-M, Marin JM, Robert CP (2009) Adaptivity for ABC algorithms: the ABC-PMC scheme. *Biometrika*, **96**, 983–990.
- Beaumont MA, Nielsen R, Robert C *et al.* (2010) In defence of model-based inference in phylogeography. *Molecular Ecology*, **19**, 436–446.
- Becquet C, Przeworski M (2007) A new approach to estimate parameters of speciation models with application to apes. *Genome Research*, **17**, 1505–1519.
- Blum MGB, Francois O (2009) Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, **20**, 63–73.
- Bortot P, Coles SG, Sisson SA (2007) Inference for stereological extremes. *Journal of the American Statistical Association*, **102**, 84–92.
- Cappè O, Guillin A, Marin JM, Robert C (2004) Population Monte Carlo. *Journal of Computing Graphics and Statistics*, **13**, 907–929.
- Carvajal-Rodriguez A (2010) Simulation of genes and genomes forward in time. *Current Genomics*, **11**, 58–61.
- Chadeau-Hyam M, Hoggart CJ, O'Reilly PF *et al.* (2008) Fregene: simulation of realistic sequence-level data in populations and ascertained samples. *BMC Bioinformatics*, **9**, 364.
- Colonna V, Natile T, Ferrucci RR *et al.* (2009) Comparing population structure as inferred from genealogical versus genetic information. *European Journal of Human Genetics*, **17**, 1635–1641.
- Cornuet JM, Santos F, Beaumont MA *et al.* (2008) Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics*, **24**, 2713–2719.
- Crestanello B, Pecchioli E, Vernesi C *et al.* (2009) The genetic impact of translocations and habitat fragmentation in chamois (*Rupicapra*) spp. *Journal of Heredity*, **100**, 691–708.
- Curat M, Ray N, Excoffier L (2004) SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. *Molecular Ecology Notes*, **4**, 139–142.
- Doucet A, de Freitas JFG, Gordon NJ (2001) *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York, NY.
- Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, **161**, 1307–1320.
- Estoup A, Clegg SM (2003) Bayesian inferences on the recent island colonization history by the bird *Zosterops lateralis lateralis*. *Molecular Ecology*, **12**, 657–674.
- Estoup A, Wilson IJ, Sullivan C, Cornuet JM, Moritz C (2001) Inferring population history from microsatellite and enzyme data in serially introduced cane toads, *Bufo marinus*. *Genetics*, **159**, 1671–1687.
- Estoup A, Beaumont M, Sennedot F, Moritz C, Cornuet JM (2004) Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, *Bufo marinus*. *Evolution*, **58**, 2021–2036.
- Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.
- Excoffier L, Estoup A, Cornuet JM (2005) Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics*, **169**, 1727–1738.
- Fagundes NJ, Ray N, Beaumont M *et al.* (2007) Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences, USA*, **104**, 17614–17619.
- Foll M, Beaumont MA, Gaggiotti O (2008) An approximate Bayesian computation approach to overcome biases that arise when using amplified fragment length polymorphism markers to study population structure. *Genetics*, **179**, 927–939.
- Francois O, Blum MG, Jakobsson M, Rosenberg NA (2008) Demographic history of european populations of *Arabidopsis thaliana*. *PLoS Genetics*, **4**, e1000075.
- Fu YX, Li WH (1997) Estimating the age of the common ancestor of a sample of DNA sequences. *Molecular Biology and Evolution*, **14**, 195–199.

- Gelman A (2008) *Rejoinder. Bayesian Analysis*, **3**, 467–478.
- Gelman A, Carlin JB, Stern HS, Rubin DB (2004) *Bayesian Data Analysis*. Chapman & Hall/CRC, London, UK.
- Ghirotto S, Mona S, Benazzo A *et al.* (2010) Inferring genealogical processes from patterns of Bronze-Age and modern DNA variation in Sardinia. *Molecular Biology and Evolution*, **27**, 875–886.
- Guillemaud T, Beaumont MA, Ciosi M, Cornuet JM, Estoup A (2010) Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data. *Heredity*, **104**, 88–99.
- Hamilton G, Currat M, Ray N *et al.* (2005) Bayesian estimation of recent migration rates after a spatial expansion. *Genetics*, **170**, 409–417.
- Hernandez RD (2008) A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, **24**, 2786–2787.
- Hickerson MJ, Stahl EA, Lessios HA (2006) Test for simultaneous divergence using approximate Bayesian computation. *Evolution*, **60**, 2435–2453.
- Hickerson MJ, Stahl E, Takebayashi N (2007) msBayes: pipeline for testing comparative phylogeographic histories using hierarchical approximate Bayesian computation. *BMC Bioinformatics*, **8**, 268.
- Holder M, Lewis PO (2003) Phylogeny estimation: traditional and Bayesian approaches. *Nature Review Genetics*, **4**, 275–284.
- Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.
- Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, **294**, 2310–2314.
- Ingvarsson PK (2008) Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*. *Genetics*, **180**, 329–340.
- Itan Y, Powell A, Beaumont MA, Burger J, Thomas MG (2009) The origins of lactase persistence in Europe. *PLoS Computing and Biology*, **5**, e1000491.
- Jensen JD, Thornton KR, Andolfatto P (2008) An approximate Bayesian estimator suggests strong, recurrent selective sweeps in *Drosophila*. *PLoS Genetics*, **4**, e1000198.
- Joyce P, Marjoram P (2008) Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, **7**, 26.
- Kass RE, Raftery AE (1995) Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.
- Kayser M, Lao O, Saar K *et al.* (2008) Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians. *American Journal of Human Genetics*, **82**, 194–198.
- Kimura M, Weiss GH (1964) The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, **49**, 561–576.
- Kuhner MK, Yamato J, Felsenstein J (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, **140**, 1421–1430.
- Latch E, Dharmarajan G, Glaubitz J, Rhodes O (2006) Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conservation Genetics*, **7**, 295–302.
- Laval G, 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.
- Leblois R, Estoup A, Rousset F (2009) IBDSim: a computer program to simulate genotypic data under isolation by distance. *Molecular Ecology Resources*, **9**, 107–109.
- Leuenberger C, Wegmann D (2010) Bayesian computation and model selection without likelihoods. *Genetics*, **184**, 243–252.
- Lopes JS, Beaumont MA (2010) ABC: a useful Bayesian tool for the analysis of population data. *Infection, Genetics and Evolution*, In Press.
- Lopes JS, Boessenkool S (2010) The use of approximate Bayesian computation in conservation genetics and its application in a case study on yellow-eyed penguins. *Conservation Genetics*, **11**, 421–433.
- Lopes JS, Balding D, Beaumont MA (2009) PopABC: a program to infer historical demographic parameters. *Bioinformatics*, **25**, 2747–2749.
- Luciani F, Sisson SA, Jiang H, Francis AR, Tanaka MM (2009) The epidemiological fitness cost of drug resistance in *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences, USA*, **106**, 14711–14715.
- Ludwig A, Arndt U, Lippold S *et al.* (2008) Tracing the first steps of American sturgeon pioneers in Europe. *BMC Evolutionary Biology*, **8**, 221.
- Marjoram P, Tavaré S (2006) Modern computational approaches for analysing molecular genetic variation data. *Nature Review Genetics*, **7**, 759–770.
- Marjoram P, Molitor J, Plagnol V, Tavaré S (2003) Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences, USA*, **100**, 15324–15328.
- Menozi P, Piazza A, Cavalli-Sforza L (1978) Synthetic maps of human gene frequencies in Europeans. *Science*, **201**, 786–792.
- Neuenschwander S (2006) AQUASPLATCHE: a program to simulate genetic diversity in populations living in linear habitats. *Molecular Ecology Notes*, **6**, 583–585.
- Neuenschwander S, Lurgiader CR, Ray N *et al.* (2008) Colonization history of the Swiss Rhine basin by the bullhead (*Cottus gobio*): inference under a Bayesian spatially explicit framework. *Molecular Ecology*, **17**, 757–772.
- Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.
- Pascual M, Chapuis MP, Mestres F *et al.* (2007) Introduction history of *Drosophila subobscura* in the New World: a microsatellite-based survey using ABC methods. *Molecular Ecology*, **16**, 3069–3083.
- Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, **16**, 1791–1798.
- Przeworski M (2003) Estimating the time since the fixation of a beneficial allele. *Genetics*, **164**, 1667–1676.
- Putnam AS, Scriber JM, Andolfatto P (2007) Discordant divergence times among Z-chromosome regions between two ecologically distinct swallowtail butterfly species. *Evolution*, **61**, 912–927.
- Ray N, Wegmann D, Fagundes NJ *et al.* (2010) A statistical evaluation of models for the initial settlement of the

- American continent emphasizes the importance of gene flow with Asia. *Molecular Biology and Evolution*, **27**, 337–345.
- Rosenblum EB, Hickerson MJ, Moritz C (2007) A multilocus perspective on colonization accompanied by selection and gene flow. *Evolution*, **61**, 2971–2985.
- Ross-Ibarra J, Wright SI, Foxe JP *et al.* (2008) Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. *PLoS ONE*, **3**, e2411.
- Ross-Ibarra J, Tenaillon M, Gaut BS (2009) Historical divergence and gene flow in the genus *Zea*. *Genetics*, **181**, 1399–1413.
- Sisson SA, Fan Y, Tanaka MM (2007) Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences, USA*, **104**, 1760–1765.
- Sokal RR, Oden NL, Barker JSF (1987) Spatial structure in *Drosophila buzzatii* populations: simple and directional spatial autocorrelation. *American Naturalist*, **129**, 122–142.
- Spencer CC, Coop G (2004) SelSim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics*, **20**, 3673–3675.
- Suchard MA, Rambaut A (2009) Many-core algorithms for statistical phylogenetics. *Bioinformatics*, **25**, 1370–1376.
- Tallmon DA, Koyuk A, Luikart GH, Beaumont MA (2008) ONeSAMP: a program to estimate effective population size using approximate Bayesian computation. *Molecular Ecology Resources*, **8**, 299–301.
- Tavare S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. *Genetics*, **145**, 505–518.
- Templeton AR (1993) The Eve hypotheses: a genetic critique and reanalysis. *American Anthropologist*, **95**, 51–72.
- Templeton AR (2009) Statistical hypothesis testing in intraspecific phylogeography: nested clade phylogeographical analysis vs. approximate Bayesian computation. *Molecular Ecology*, **18**, 319–331.
- Templeton AR (2010a) Coalescent-based, maximum likelihood inference in phylogeography. *Molecular Ecology*, **19**, 431.
- Templeton AR (2010b) Coherent and incoherent inference in phylogeography and human evolution. *Proceedings of the National Academy of Sciences, USA*, **107**, 6376–6381.
- Thornton K, Andolfatto P (2006) Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. *Genetics*, **172**, 1607–1619.
- Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MP (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, **6**, 187–202.
- Verdu P, Austerlitz F, Estoup A *et al.* (2009) Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Current Biology*, **19**, 312–318.
- Voje KL, Hemp C, Flagstad O, Saetre GP, Stenseth NC (2009) Climatic change as an engine for speciation in flightless Orthoptera species inhabiting African mountains. *Molecular Ecology*, **18**, 93–108.
- Wakeley J, Hey J (1997) Estimating ancestral population parameters. *Genetics*, **145**, 847–855.
- Wegmann D, Leuenberger C, Excoffier L (2009) Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, **182**, 1207–1218.
- Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics*, **11**, 116.
- Weiss G, von Haeseler A (1998) Inference of population history using a likelihood approach. *Genetics*, **149**, 1539–1546.
- Wilson DJ, Gabriel E, Leatherbarrow AJ *et al.* (2009) Rapid evolution and the importance of recombination to the gastroenteric pathogen *Campylobacter jejuni*. *Molecular Biology and Evolution*, **26**, 385–397.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 The ratio of posterior vs. prior distribution range (RR) and the ratio of the largest vs. the smallest location measure in the posterior and the prior distributions (ER) computed from 14 ABC studies and 152 parameter estimates. n: number of parameter estimates subdivided into four groups; Q1 and Q3: first and third quartile. Sample sizes are not the same for RR and ER because the information required to compute them was not homogeneous across studies. See text for additional details.

Fig. S1 The relationship between ER and RR when the median value of ER is computed separately within six RR bins. Bars are quartiles, and the numbers indicate the fraction of point belonging to each bin. Codes are as in Table 2.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.