Heredity

Heredity (2012) 108, 521–530

www.nature.com/hdy

ORIGINAL ARTICLE

Population divergence with or without admixture: selecting models using an ABC approach

VC Sousa[1,2,5], MA Beaumont[3], P Fernandes[1], MM Coelho[2] and L Chikhi[1,4]

Genetic data have been widely used to reconstruct the demographic history of populations, including the estimation of migration rates, divergence times and relative admixture contribution from different populations. Recently, increasing interest has been given to the ability of genetic data to distinguish alternative models. One of the issues that has plagued this kind of inference is that ancestral shared polymorphism is often difficult to separate from admixture or gene flow. Here, we applied an approximate Bayesian computation (ABC) approach to select the model that best fits microsatellite data among alternative splitting and admixture models. We performed a simulation study and showed that with reasonably large data sets (20 loci) it is possible to identify with a high level of accuracy the model that generated the data. This suggests that it is possible to distinguish genetic patterns due to past admixture events from those due to shared polymorphism (population split without admixture). We then apply this approach to microsatellite data from an endangered and endemic Iberian freshwater fish species, in which a clustering analysis suggested that one of the populations could be admixed. In contrast, our results suggest that the observed genetic patterns are better explained by a population split model without admixture.
Heredity (2012) **108,** 521–530; doi:10.1038/hdy.2011.116; published online 7 December 2011

Keywords: admixture; shared ancestral polymorphism; ABC; model choice; microsatellite

INTRODUCTION

The use of genetic data to reconstruct the demographic history of populations is now well established (see, for example, Goldstein and Chikhi, 2002; Hey and Machado, 2003). Many inferential methods have been developed in the past 20 years that allow biologists to detect, date or quantify population size changes (see, for example, Cornuet and Luikart, 1996; Storz and Beaumont, 2002), and to estimate the time at which different populations separated, migration rates (see, for example, Nielsen and Wakeley, 2001) or the relative contribution of parental populations in admixture models (see, for example, Chikhi *et al.*, 2001, 2002). In most studies published to date, a particular demographic model is assumed and their aim is to determine the most likely values for the parameters, say the splitting time, conditional on observed genetic data (Beaumont *et al.*, 2010). The underlying methods are usually evaluated with simulated data first, but the final objective is to apply them for the analysis of real data sets. One important assumption of this approach is that the model chosen is a reasonable approximation of the main demographic events that have affected the populations under study (Chikhi *et al.*, 2001; Hey and Machado, 2003). This general approach has proven to be useful (see, for example, Marjoram and Tavaré, 2006; Beaumont *et al.*, 2010), but recent advances in population genetics have now made it easier to compare alternative models (Estoup *et al.*, 2004; Johnson and Omland, 2004; Fagundes *et al.*, 2007; Beaumont, 2008; Guillemaud *et al.*, 2009).

One example where it is important to distinguish alternative models is in the study of admixed populations. Among other genetic signatures, admixed populations are expected to exhibit allele frequencies that appear intermediate in relation with the putative parental populations (Chikhi *et al.*, 2001). This underlies many of the model-based clustering algorithms that have become popular to study and identify admixed populations (see, for example, Pritchard *et al.*, 2000; Corander *et al.*, 2004). However, most of these clustering methods do not model explicitly the demographic history, and these patterns characterized by intermediate allele frequencies could have arisen simply by drift or some other demographic scenario. The question is thus whether observed data are better explained by a past admixture event or an alternative scenario. Model-choice methods are useful as they quantify the evidence of the data in favor of different models.

Approximate Bayesian computation (ABC) methods (Beaumont *et al.*, 2002; Marjoram *et al.*, 2003; Beaumont, 2010; Csilléry *et al.*, 2010) have seen major recent developments allowing inference of demographic parameters under complex demographic models, involving several populations and up to two independent admixture events in the case of admixture models (Excoffier *et al.*, 2005; Cornuet *et al.*, 2008; Sousa *et al.*, 2009; Bray *et al.*, 2009b). Here we used the ability of ABC methods to assess the relative probability of alternative demographic models (Estoup *et al.*, 2004; Fagundes *et al.*, 2007; Beaumont, 2008; Cornuet *et al.*, 2008; Guillemaud *et al.*, 2009).

Although several studies have used the ABC framework to compare alternative demographic models, their performance remains poorly understood. From a theoretical perspective, there has been recent

[1]Instituto Gulbenkian de Ciência, Rua da Quinta Grande, Oeiras, Portugal; [2]Universidade de Lisboa, Faculdade de Ciências, Centro de Biologia Ambiental, Campo Grande, Lisboa, Portugal; [3]School of Mathematics and School of Biological Sciences, University of Bristol, Bristol, UK and [4]Laboratoire Evolution et Diversité Biologique–UMR CNRS UPS 5174, Université Paul Sabatier, Toulouse, France
[5]Current address: Department of Genetics, Rutgers University, Piscataway, NJ 08854, USA.
Correspondence: Dr VC Sousa or L Chikhi, Instituto Gulbenkian de Ciência, Rua da Quinta Grande, N.6, P-2780-156 Oeiras, Portugal.
E-mail: vsousa@rci.rutgers.edu or chikhi@cict.fr and chikhi@igc.gulbenkian.pt
Received 18 April 2011; revised 7 August 2011; accepted 9 August 2011; published online 7 December 2011

debate about the reliability of ABC for model choice (Didelot *et al*., 2011; Robert *et al*., 2011). The question hinges on the potential sensitivity of the method to the choice of summary statistics. We revisit this issue in the Discussion, but note that both these studies stress the importance of performing simulation studies before analyzing real data, in order to characterize ABC performance when applying it to a particular set of models.

The study of Guillemaud *et al*. (2009), where ABC was used to study invasive species and assess the relative probability of different models of species introduction, was one of the first to perform an extensive simulation study to assess the performance of ABC as a model-choice method in population genetics. Their results suggested that ABC may be a successful tool in comparing alternative models, confirming previous results with limited simulation studies (Estoup *et al*., 2004; Fagundes *et al*., 2007; Beaumont, 2008).

In the present study we apply an ABC approach to select the model that best fits microsatellite data among a set of alternative splitting and admixture models. We show that with reasonably large data sets it is possible to determine with high probability the model that most likely produced the data. We then apply this approach to data from an endangered fish species, in which a clustering analysis suggested that a population could be admixed. Our approach suggests that the apparent admixture is more likely the result of shared ancestral polymorphism between differentiated populations. We therefore show the importance of accounting explicitly for the demographic history of populations in the case of admixture models.

MATERIALS AND METHODS

Demographic models

Admixture occurs when two or more differentiated populations are brought together into contact, creating hybrid or admixed populations. For instance, admixture may occur during the colonization of already occupied areas and after the domestication of plants and animals (for example, formation of new breeds; Bray *et al*., 2009a). Also, these can be particularly common in freshwater species, when changes in the drainage system of rivers allow for secondary contact between divergent populations. Admixture events involving more than two parental populations have been reported in humans (Wang *et al*., 2008) and breeds (Bray *et al*., 2009a). Note that these situations may not be well modeled by island models or isolation with migration models where an ongoing gene flow between two or more diverging populations is assumed. We thus considered two population split models and four admixture models (Figure 1). Figure 1 shows models with either three or four populations. In all models it is assumed that an ancestral population of size N_A split at t_{split} generations ago into two, three or four populations, depending on the model, with sizes N_i, i=(1, 2, 3, H). Under the population split models, the populations remain isolated from each other after the split event and evolve independently (with no gene flow; Figures 1a and d). The admixture models can involve one or two admixture events, and either two or three parental populations. Under the admixture models with one admixture event there is a unique admixture event creating a hybrid population t_{adm1} generations ago (Figures 1b and e). If there are two parental populations, called P_1 and P_2, they will contribute genes to the hybrid population in proportions p_1 and p_2 such that $p_1+p_2=1$ (Figures 1b and c). If there is a third parental population P_3, contributing p_3, then we will have $p_1+p_2+p_3=1$ (Figure 1e). In the models with two admixture events, the first admixture event will take place t_{adm1} generations ago and will only involve two parental populations, P_1 and P_2, such that $p_1+p_2=1$. The second admixture event is then assumed to occur t_{adm2} generations ago. In the model with two parental populations, P_2 is assumed to contribute again to the gene pool of the hybrid a proportion p_3 such that $0 \leqslant p_3 \leqslant 1$ (Figure 1c). In the model with three parental populations, it is the third population P_3 that is assumed to contribute p_3 (Figure 1f). In the admixture models, the admixed (or hybrid) population is assumed to have an effective size N_H. We note that in all models the loci are assumed to have the same per locus mutation rate μ and to evolve according to the stepwise mutation model, as is usually assumed for microsatellites (see, for example, Calabrese and Sainudiin, 2005).

ABC principles

The principle of ABC is to obtain the joint posterior distribution of parameters using simulations under a demographic model of interest (Beaumont *et al*., 2002; Marjoram *et al*., 2003; Beaumont, 2010; Csilléry *et al*., 2010). ABC methods are very flexible as they can be applied to demographic models for which there are no explicit likelihood functions (Marjoram and Tavaré, 2006). Data sets are simulated with parameter values drawn from prior distributions. The corresponding parameters are then accepted if the simulated data are

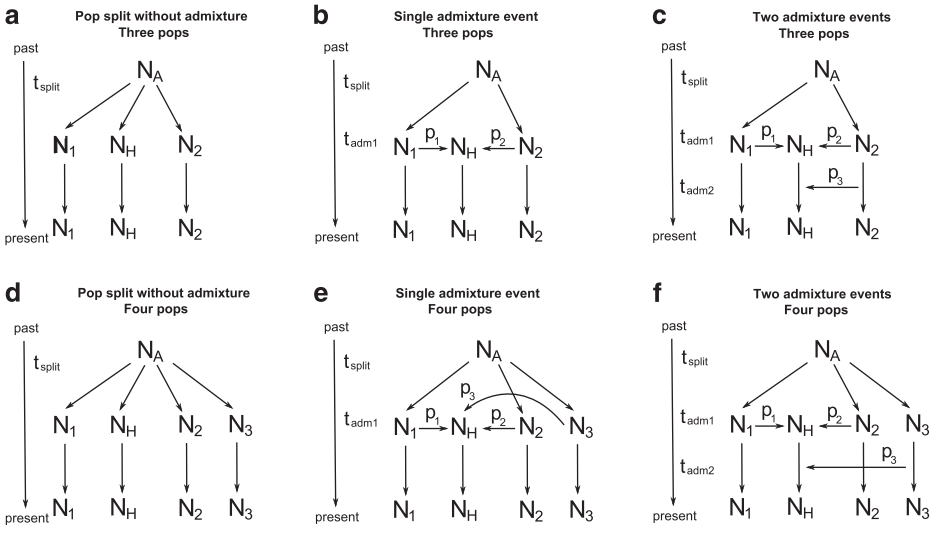

Figure 1 Admixture and population split models. (**a**) Population split model with three populations and without admixture. (**b**) Admixture model with two parental populations and one admixture event. (**c**) Admixture model with two parental populations and two admixture events. (**d**) Population split model with four populations and without admixture. (**e**) Admixture model with three parental populations and one admixture event. (**f**) Admixture model with three parental populations and two admixture events. In all models, the populations are allowed to have different effective sizes N_i, (i=1, 2, 3, H). The admixture and split events occurred at t_{adm1}, t_{adm2} and t_{split} generations ago.

similar to the vectors of observed data, according to a certain distance measure $d(.)$, and rejected otherwise. The values of the parameters θ that generated the closest data sets to the observed data are then taken as an approximation of the posterior distribution $P(\theta|d(D_{sim},D_{obs}) < \delta)$ where D_{sim} and D_{obs} are the simulated and observed data, respectively, and δ is an arbitrary tolerance level. In most ABC methods, instead of using the observed data D directly (allele or genotype frequencies) the data are summarized by a set of summary statistics S, such as expected heterozygosity (H_e), number of alleles or F_{ST}. Therefore, most ABC methods provide an approximate estimate of the posterior $P(\theta|d(S_{obs},S_{sim}) < \delta)$, where S_{obs} and S_{sim} represent the observed and simulated summary statistics, respectively.

Model choice and ABC

As first suggested by Pritchard *et al.* (1999), it is possible to use ABC to compute the posterior probability of a given demographic model among a set of alternatives. Performing the ABC rejection algorithm, the posterior probability of a model is given by simply counting the proportion of corresponding simulated statistics that lie within the tolerance region (defined by $d(S_{obs},S_{sim}) < \delta$). Beaumont (2008) suggested an improvement on this simple approach by using a weighted multinomial *logit* regression. The principle of the multinomial regression is to obtain the relation between categorical variables $Y=1, 2, \ldots$ indicating different demographic models k and the corresponding accepted summary statistics S_{sim}. By using a *logit* function, the regression describes the dependence of the posterior probability of a given model p_k as a function of the accepted summary statistics (Fagundes *et al.*, 2007; Beaumont, 2008; Cornuet *et al.*, 2008). Therefore, after performing the regression with the summary statistics accepted in the rejection step it is possible to assess the posterior probability of model k given the observed summary statistics $P(Y=k|S=S_{obs})$. It is noteworthy that assuming equal prior probabilities for the alternative models, the ratio of the posteriors obtained with the ABC approximates the Bayes factor (Didelot *et al.*, 2011), which is defined as the ratio of the marginal likelihoods of data D_{obs} under models M_1 and M_2 ($BF=p(D_{obs}|M_1)/p(D_{obs}|M_2)$). For simplicity, all our model comparisons were performed by comparing two models at a time (for example, no-admixture versus one-admixture event, see below).

Summary statistics

The different models for which the ABC approach was performed were compared using the following summary statistics, averaged over loci: (1) expected heterozygosity (H_e) estimated following Nei (1978) for each population and overall populations; (2) number of private alleles for each population (p_a); (3) number of alleles for each population (n_a); (4) microsatellite allele range for each population and overall populations (a_r) and (5) pairwise F_{ST} and overall populations F_{ST}, with the F_{ST} value computed as $(H_{total}-H_{local})/H_{total}$, where H_{local} is the mean H_e of the populations considered and H_{total} is computed by pooling together the different population samples. Altogether, models with three and four parental populations were summarized by 18 (4 H_e, 3 p_a, 3 n_a, 4 a_r and 4 F_{ST}) and 25 (5 H_e, 4 p_a, 4 n_a, 5 a_r and 7 F_{ST}) summary statistics, respectively.

Simulation study

The performance of our ABC-based model-choice approach was assessed with simulated data sets under known models. Data sets simulated under a particular model were used as test data sets and two different models were chosen (the true model and an alternative one) to determine whether the ABC method was able to identify the true model as the most likely. We tested our ABC approach under the following cases: (1) single admixture vs no admixture; (2) two admixture events vs no admixture and (3) single admixture vs two admixture events. This was done with three and four parental populations, making a total of six pairwise comparisons. For each pair of models, we analyzed 10 000 independent simulated test data sets generated for each of the two alternative models, corresponding to 12 model comparisons and 120 000 data sets in total. The values of the parameters used to simulate these data sets were sampled from the prior distributions. The aim was to explore the entire parameter space rather than focusing on particular parameter values. Each data set consisted of 25 diploid individuals sampled from each population and typed at 20 independent microsatellite loci. For the models with two parental populations, the effect of varying the number of loci was investigated by repeating the analyses with 5 loci, hence making a total of 180 000 analyzed data sets.

For each set of model pairs the ABC rejection step was performed simulating 10^6 data sets under each model, accepting the closest 25 000 simulations (tolerance level=0.0125). The regression step of Beaumont (2008) was performed on these accepted simulations, and a point estimate for the probability of a given model was obtained for each simulated test data set. We thus obtained 10 000 such estimates for each model, for each set of model pair comparison. These 10 000 values were used to produce the distribution of the posterior probability of each model, allowing us to quantify whether a particular model was correctly identified (Figure 2).

In addition, we performed a receiver operating characteristic (ROC) curve analysis, assuming that the ABC model-choice procedure can be seen as a classifier, as in Bazin *et al.* (2010). The method ranks the posterior probabilities for one model, say the no-admixture model, from highest to lowest. For each of these posterior probabilities we know whether or not the data actually came from the no-admixture model. If we had set a posterior probability of 1.0 as the threshold for deciding whether to classify the data as coming from the no-admixture model, we would then have a proportion 0.0 of the simulated no-admixture cases correctly classified (true positives), but also a proportion 0.0 of admixture cases incorrectly classified (false positives). We would plot this as a point on the lower left corner of the ROC curve. On the other hand, if we set 0.0 as the threshold, then we would have a proportion 1.0 of all the no-admixture instances classified correctly (true positives) and a proportion 1.0 of all the admixture case incorrectly classified (false positives), and this would be plotted as a point on the top right corner of the ROC curve. The ROC curve is constructed by successively taking the posterior probabilities in the list from highest to lowest and plotting the proportion of no-admixture cases that are correctly classified (true positives) and the proportion of admixture cases that are incorrectly classified (false positives). The ideal case occurs when all the no-admixture cases occur first in the list, followed by all the admixture cases, in which case the area under the ROC curve (AUC) would be 1. A random classifier would fluctuate around the diagonal and have an AUC of 0.5. The ROC curves in Figure 3 are based on 10 000 simulations, which is why they are quite smooth. The sampling error is very small (of the order of the line thickness), and error bars are omitted from the plots. The ROC analysis was done using the method implemented in the ROCR R package (Sing *et al.*, 2005).

All the simulations performed here were done using a modified version of the code developed for the program 2BAD, which allows simulating data under different models of population split and admixture described here and perform ABC (for details, see Bray *et al.* (2009b)). We also used 2BAD at the latest stages of this study to test the consistency between the two codes and identify bugs. The results between the two versions were identical and were used for the real freshwater fish data.

Prior distributions

The prior probabilities of the two alternative models were set as 0.5, meaning that *a priori* both models were equally likely to explain the data. For all models, the effective sizes N_i, $i=(1, 2, 3, H)$ of all populations were taken from a uniform $U[10^3, 10^4]$, the mutation rates (per locus per generation) were sampled from $U[10^{-5}, 10^{-3}]$, and t_{split} (in generations) from $U[10^3, 10^4]$. For the models with admixture, the times of admixture events (in generations) were drawn from $U[10^2, 10^3]$ for t_{adm1} and $U[1, 10^2]$ for t_{adm2}. In the case of models with a single admixture event, t_{adm1} was assumed to be sampled from a uniform $U[1, 10^3]$. The prior distributions for p_1, p_2 and p_3 were sampled from a uniform $U[0, 1]$, such that the sum of all admixture contributions was one.

Iberochondrostoma lusitanicum data

The data consisted of 129 individuals sampled in three rivers (one from the Samarra drainage, SM1 $n=43$, and two from the Tejo drainage, TJ1 $n=48$, and TJ2 $n=40$) and genotyped at five microsatellite loci (see details in Sousa *et al.*, 2008). *I. lusitanicum* (Cyprinidae) is a critically endangered freshwater fish species only found in lower Tejo, Sado and other small drainages in Portugal.

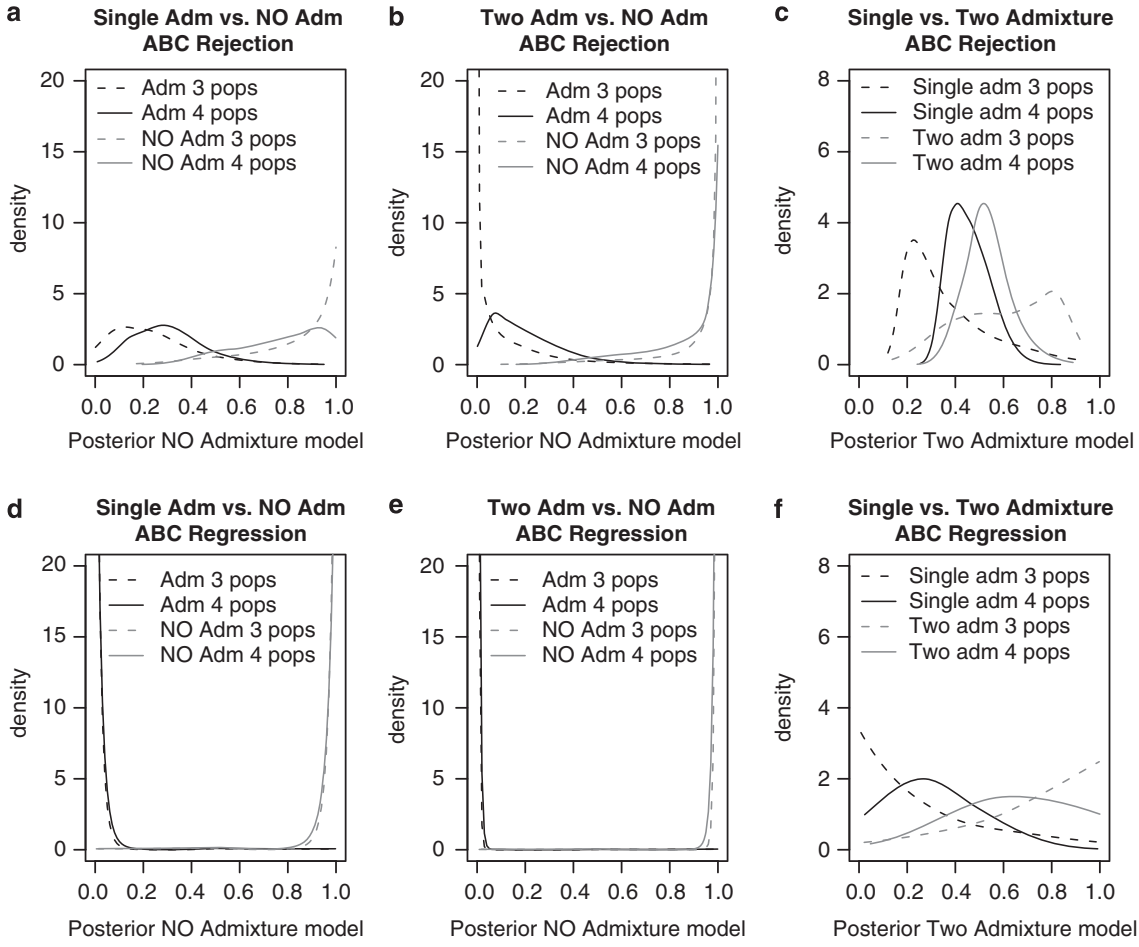

Figure 2 Distribution of the posterior probabilities for the population split without admixture model (NO admixture; **a**, **b**, **d**, **e**), and for the two-event admixture model obtained with the simulated data sets (**c**, **f**). Each curve was obtained with the analysis of 10 000 simulated data sets. In (**a**), (**b**), (**d**) and (**e**), black lines correspond to data sets generated under the admixture models, whereas gray lines correspond to data sets generated under the NO admixture models. In (**c**) and (**f**), black lines correspond to data sets generated under the single-admixture models (single adm), and gray lines correspond to data sets generated under the two-event admixture models (two adm). Solid lines correspond to the four-population models and dashed lines to the three-population models. The simulated test data sets were simulated with parameters sampled from the priors of each model.

In a recent study we found that most populations were highly differentiated from each other with medium to large pairwise F_{ST} between the three samples analyzed here. The Weir and Cockerham (1984) estimates ranged from 0.22 to 0.41 and Nei (1978) pairwise G_{ST} ranged from 0.09 to 0.21. Moreover, by performing a genetic clustering analysis using the STRUCTURE program (Pritchard *et al*., 2000) we found that one population of the Tejo drainage (TJ1) could potentially be the result of admixture between populations from the Tejo (TJ2) and the Samarra (SM1) drainages. The estimates obtained indicated an admixture contribution of 0.31 from SM1 and 0.69 from TJ2. The STRUCTURE analysis suggested that the three samples could also correspond to three independent clusters but this appeared less likely than the admixture model with only two clusters. Given that STRUCTURE makes no explicit assumptions regarding demographic history of the species analyzed, it was unclear whether the genetic patterns found were due to an ancient admixture event or simply due to the differentiation of populations without admixture, that is, shared ancestral polymorphism. The same data sets were thus re-analyzed using the ABC approach described here to assess the most likely scenario: single admixture event versus population split without admixture. We performed 10^6 simulations under each model with uniform prior distributions. The priors were specified according to previous estimates obtained with the MSVAR program (Storz and Beaumont, 2002). The results of this analysis suggested a recent population decrease from an ancestral effective size larger than 10^4 to current sizes of $\sim$10–100. We also allowed for both recent and ancient admixture and population split events. The effect of the priors used was tested repeating the analysis with two sets of priors. In the first, the priors were $U[10, 10^4]$ for N_1, N_2, N_H and N_A, $U[5\times10^{-5}, 5\times10^{-4}]$ for the mutation rate μ, $U[10, 5\times10^4]$ for t_{split}, $U[10, 10^4]$ for t_{adm1} and $U[0, 1]$ for p_1. In the second analysis, smaller effective sizes for current populations and mutation rates were allowed, with the priors $U[1, 10^3]$ for N_1, N_2 and N_H, $U[1, 10^4]$ for N_A, $U[10^{-6}, 10^{-4}]$ for the mutation rate μ, $U[1, 10^5]$ for t_{split}, $U[1, 10^3]$ for t_{adm1} and $U[0, 1]$ for p_1. In both cases, the same priors were used for the population split model without admixture. In addition to the point estimates of the posterior probability for each model obtained with the regression, the 95% confidence intervals were computed following Cornuet *et al*. (2008).

RESULTS

As shown in Figure 2, the method is able to identify the correct model with high posterior probabilities in the case of single or two admixture events versus population split without admixture (Figures 2a and b). This can be seen by the fact that the posterior probabilities are close to one when the no-admixture model is true, or close to zero when the admixture model is true. In other words, for most simulated test data sets we were able to determine with extremely high confidence whether the data came from a pure split model or from a model

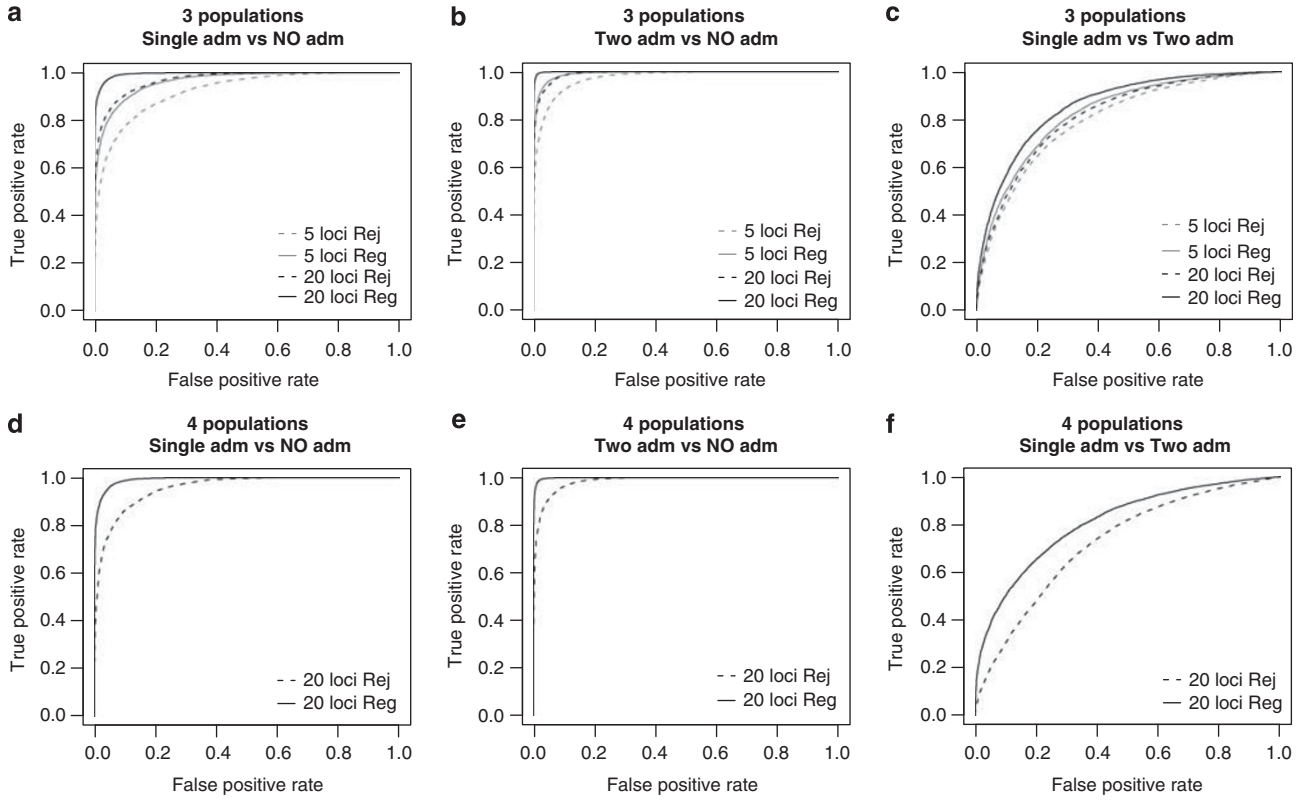

Figure 3 ROC analysis of the ABC method for the different scenarios tested. The top panel shows the ROC curves for the comparison of alternative models with three populations (**a**–**c**), and the bottom panel for models with four populations (**d**–**f**). The curves compare the results obtained with a pure ABC rejection (dashed lines) and applying the logistic regression step (solid lines). For the three-population models, the curves obtained with 5 and 20 loci are shown as gray and black lines, respectively. (**a**) Single admixture versus no admixture (two parental); (**b**) two admixture events versus no admixture (two parental); (**c**) single admixture versus two admixture events (two parental); (**d**) single admixture versus no admixture (three parental); (**e**) two admixture versus no admixture (three parental); (**f**) single admixture versus two admixture (two parental).

with at least one admixture event. It also appears to be easier to identify a two-admixture event against a split model as compared with a single-admixture event (Figures 2a versus b). Another result shown in this figure is that it is usually easier to separate splitting from admixture models when there are three rather than four populations involved (dashed versus solid lines). Also, we found that the separation of admixture from no admixture is not symmetrical (black versus gray lines). Indeed, it is easier to identify a sample generated by a population split model (gray lines), than a sample generated under an admixture model (black lines). This figure also shows that the regression method of Beaumont (2008) greatly improved our ability to identify the correct models (Figures 2d and e compared with Figures 2a and b, respectively). After the use of the regression step, there is still an asymmetry between admixture and no-admixture models but it is much more limited. When we compared the admixture models (Figures 2c and f) we found that the posteriors are shifted toward the value of 0.5, suggesting that in most simulations the data sets could not be clearly attributed to one admixture model or the other. Thus, it indicates that it is difficult to separate models involving one or two admixture events. This effect was more pronounced in the model involving four populations (solid versus dashed lines in Figures 2c and f).

The results obtained with the ROC analysis are shown in Figure 3. Each plot corresponds to the comparison of two alternative models, and the curves obtained with the rejection and regression steps are shown. A good classifier is characterized by a true positive rate close to one, and a false positive rate close to zero. This is reflected by a ROC curve close to the upper left corner (Fawcett, 2006). Thus, our results indicate that ABC performs well as a classifier to identify the model that most likely generated the data, especially for the cases of admixture versus no admixture (Figures 3a, b, d and e). The AUC values ranged from 0.924 to 0.999 for the case of admixture versus no admixture. Moreover, we found that with 20 loci the AUC was higher than 0.993 when the regression step was applied. For the comparison of single admixture versus two admixture events, the AUC ranged from 0.726 to 0.860. Overall, these results are in agreement with the distribution of the posterior probabilities shown in Figure 2.

Tolerance, *logit* regression and the number of loci

Figure 4 shows the effect of the tolerance (that is, accepting data sets that are increasingly closer to the simulated test data sets), of the number of loci, and of the regression step on the ability to identify the correct model. This is represented by the average posterior probability (that is, the mean of the posterior distributions similar to those shown in Figure 2). Figure 4a compares the population split model with no admixture with a single-admixture model. This figure shows that using the rejection method (dashed lines) the average accuracy of the method increases when the tolerance decreases, as the posterior probabilities tend to one (when the no admixture is true) and zero (when there was admixture). When the *logit* regression is applied the dependence on the tolerance is much weaker and good results can be

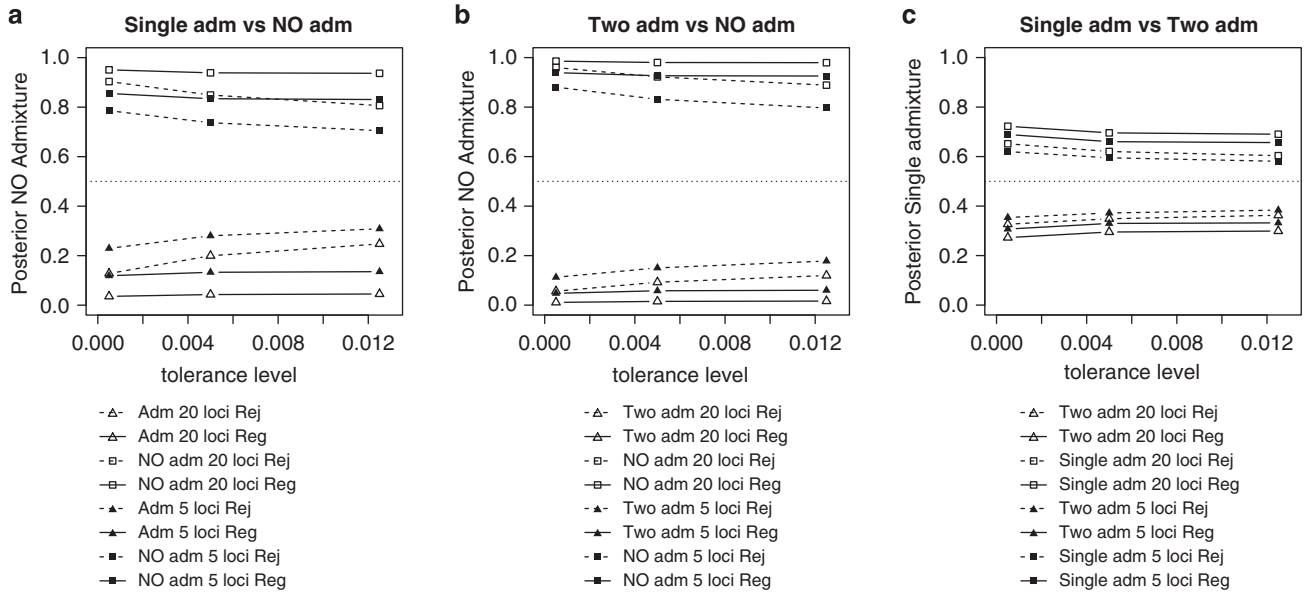

Figure 4 Effect of tolerance and regression in the mean posterior distributions. (**a**, **b**) Average posterior distribution for the population split without admixture model (NO admixture) as a function of the tolerance, and average posterior distribution for the single-event admixture model (Single admixture) in (**c**). Each point represents the average of 10 000 posterior probabilities. The simulated test data sets were simulated with parameters sampled from the prior. The horizontal dotted line corresponds to the prior probability of 0.50, meaning that both models are equally likely.

obtained with larger tolerance levels, and hence with fewer simulations. However, we note that even with small tolerance levels, the regression step improves the results (solid versus dashed lines). We also see that with 5 loci the average accuracy decreases in comparison with the results obtained with 20 loci (open versus filled symbols), but they still provide accurate estimates that are usually good enough to identify the most likely model. Figure 4b shows similar results for the comparisons between the splitting and two-admixture events. The main difference with Figure 4a is that the identification of the correct model is even better than in the single-admixture model, and provides good results even with only five loci (average posterior probability >0.80). In Figure 4c, we also find the same trend for the comparison between the two admixture models. However, the results are not as good, with average probabilities <0.80 even with 20 loci, and using the logistic regression and low tolerance levels.

Application to the critically endangered Iberian minnow *I. lusitanicum*

As Figure 5 shows, our analysis indicates that, after the regression step, the probability of the admixture model ranged from 0.10 to 0.40, depending on the tolerance level and prior set. Thus, data favor the population split model without admixture over a model with admixture, independently of the priors used. However, the posterior probabilities obtained for the higher tolerance level correspond to Bayes factors in favor of no admixture of 0.73/0.27=2.7 (first prior set), or 0.77/0.23=3.35 (second prior set), where 0.73 and 0.77 refer to the posterior of the model without admixture, and 0.27 and 0.23 refer to the posterior of the admixture model, for the two prior sets, respectively. These Bayes factors values are low and, according to the Jeffrey's classification, are on the borderline between 'barely worth a mention' and 'substantial' (see, for example, Didelot *et al.*, 2011). Moreover, we note that the results obtained have wide confidence intervals with an upper limit that goes beyond 0.50 for some tolerance levels. Hence, these results are not completely conclusive.

DISCUSSION

Disentangling admixture from ancestral polymorphism

In this study we examined the ability of ABC methods to infer the relative probability of alternative models involving admixture events and population splits. We performed a simulation study showing that it is possible to identify the model that generated the data from a pair of alternative population split or admixture models (Figure 1). In particular, the accuracy of our approach to separate scenarios with admixture events from scenarios with population split without admixture was very high in the simulation study. We believe that it is a significant result, as it suggests that populations that are thought to be the result of admixture events can be identified as the result of splitting events without admixture with high probabilities, and vice versa. This suggests that it is possible to distinguish genetic patterns due to past admixture events from those due to shared ancestral polymorphism.

At the same time, our results showed that it is much more difficult to provide conclusive posterior distributions when comparing pairs of alternative models comprising one or two admixture events (Figure 2). This is not very surprising as our comparisons were made across a wide range of parameter combinations, and given that these two scenarios can be seen as nested models. In particular, we used priors for t_{adm1} and t_{adm2} such that the two times could be close from each other. Also, our ability to determine that there had been a second admixture event increased with the contribution of population P_3. Indeed, when p_3 is low, a two-event admixture model is similar to a single-event admixture. We also found that our ability to identify the correct model was dependent on the time since the last admixture event (t_{adm2}) (Supplementary Figure 1). Finally, increasing the number of loci was also shown to be crucial to separate single- and two-event admixture models.

For all the pairs of models compared we found that the use of the logistic regression provided better results than the rejection method, significantly increasing the posterior probability of the correct model

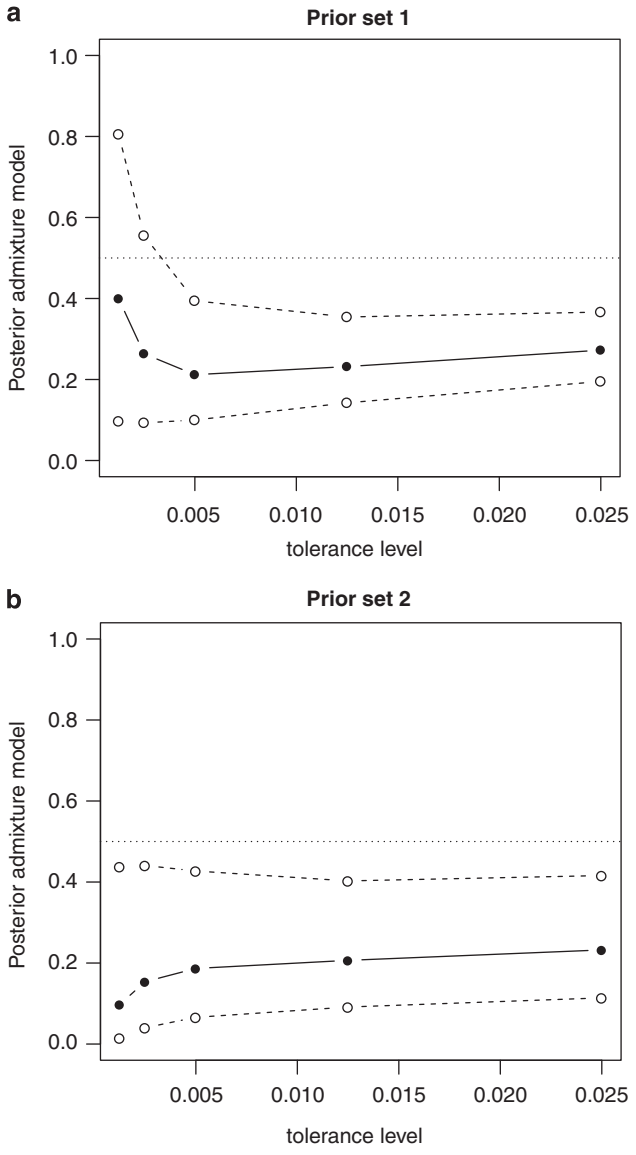

Figure 5 Posterior probability for the single-event admixture model (admixture) obtained for *I. lusitanicum*, for the model comparison of single-event admixture vs population split without admixture. Posterior probabilities for the admixture model shown as function of the tolerance level, for: (**a**) results obtained with the first prior set; (**b**) results obtained with the second prior set. Solid lines correspond to the results obtained with the regression step and the dashed lines correspond to the 95% confidence interval. The horizontal dotted lines correspond to the prior probability of 0.50, meaning that both models are equally likely.

(Figure 2). We also found that applying the regression step decreased the dependence on the tolerance level δ (Figure 4). This limited dependency on δ was also found when using ABC algorithms to estimate parameters within a single model (Beaumont *et al.*, 2002; Excoffier *et al.*, 2005; Sousa *et al.*, 2009; Wegmann *et al.*, 2009). The fact that the regression step decreases the dependence on the acceptance rate means that the number of simulations needed to separate models can be significantly reduced without losing much power (Figure 4).

Although our results were much better with 20 loci compared with 5 loci, the method was able to distinguish admixture from splitting models even with 5 loci (Figures 4a and b). This result was surprising as previous ABC methods developed to estimate admixture proportions required relatively large numbers of loci to provide precise estimates (Excoffier *et al.*, 2005; Sousa *et al.*, 2009). This indicates that a limited number of loci may be enough to assess the relative probability of alternative models, but that more loci are needed to estimate with precision the parameters of the models, once the latter have been identified. This should clearly be better investigated as this depends on the models that are compared. For instance, the separation of the two admixture models was not very good, even with 20 loci. Also, when full-likelihood methods were used to estimate admixture parameters for admixture models, it was found that good results could be obtained with between 5 and 10 loci (Chikhi *et al.*, 2001; Choisy *et al.*, 2004). At this stage, more work is needed to determine the conditions under which small data sets can be useful for model comparisons, and also when they are likely to be misleading.

Another parameter that seemed to be important is the effective size of the hybrid or admixed population (Supplementary Figure 2). When its effective size is small, genetic drift will be very important and it will become difficult to estimate the original contributions of the parental populations and hence to determine whether the data were generated with or without admixture. This is in agreement with the results found in several studies (Chikhi *et al.*, 2001; Wang, 2003; Choisy *et al.*, 2004) reporting that increasing drift since the admixture event increases the uncertainty around the admixture estimates (see also Bray *et al.*, 2009a for an application to breeds). Also related with the amount of drift, we found that the t_{adm} and t_{split} were also correlated with the ability to distinguish models with admixture from models without admixture (Supplementary Figure 2). Therefore, it is expected that population bottlenecks and expansions that affect the hybrid population, and hence its effective size, may influence the ability of this ABC method to separate admixture from population split models (see below).

Application to *I. lusitanicum*

When we applied our model-choice approach to the *I. lusitanicum* data we found that the most likely model was a population split model. This is particularly interesting, as in our original study (Sousa *et al.*, 2008) the results obtained with the STRUCTURE program were suggesting the existence of individuals with admixed genotypes in one of the Tagus population (TJ1, Sousa *et al.*, 2008). The STRUCTURE result was very surprising based on the fact that the potential parental populations are located in two different river drainages that do not communicate and the fact that *I. lusitanicum* has very limited dispersal ability. The conclusion of the Sousa *et al.* (2008) study was that this admixture was either due to an ancient admixture event when the rivers were connected, to ongoing migration between the populations (perhaps through undocumented human-driven translocations) or due to shared ancestral polymorphism. The current results suggest that the latter is the most likely explanation. This was further examined by analyzing data sets simulated under the population split models with and without admixture with STRUCTURE. As can be seen in Supplementary Figure 3, the STRUCTURE results indicate admixture when data were actually simulated without admixture, whereas the ABC model-choice approach identifies the correct model. In fact, we note that STRUCTURE identifies admixed individuals in the parental populations, even though they cannot be admixed in any of the models used. However, for the *I. lusitanicum*, the estimates of the posterior probabilities obtained with the ABC approach were not completely conclusive, and more loci would be required to confirm them. Also, other factors not considered in the

models may favor the no-admixture model, such as local selection and demographic events such as bottlenecks. Thus, some of the model assumptions may not hold for this specific data set. In particular, variation of mutation rate among loci and demographic events such as bottlenecks are not included in any of the alternative models. Given that field observations (Alves and Coelho, 1994) and genetic data (Sousa *et al*., 2008) indicated that *I. lusitanicum* populations suffered recent declines, it is possible that none of the two alternative models is a good enough approximation of the demographic history of this species (see below). As a simple test we simulated data sets under a two-phase mutation model, allowing for variation of the mutation rate among loci and population bottlenecks. The analyses of these data sets suggest that the ABC model choice is to some extent robust to these deviations, and that our conclusion that there was no admixture should hold for *I. lusitanicum* (Supplementary Figure 4).

Limitations

Although the model-choice approach used in this paper provides a way to separate the effects of admixture from those of pure divergence of populations, the interpretation of the results may be influenced by the following caveats.

Specification of models. First, in all our comparisons, the data were always generated under at least one of the two models compared. Thus, when analyzing real data our method will tend to identify one of the models as the most likely, even when none of the alternative models fit the observed data. This could be the case with the *I. lusitanicum* data analyzed here. One useful approach to assess the fit of the data to the model is to compare the distances obtained with data sets that fit the model (simulated data sets) with the distance distribution obtained with real data (see, for example, Sousa *et al*., 2009). When the real data distances are much greater than those of the simulated data sets, it suggests that the model identified as the most likely might not be appropriate, and other models should be investigated. This principle is similar to the one used by Ratmann *et al*. (2009) to perform model criticism. We applied this approach to the
I. lusitanicum data by obtaining the distribution of distances between the observed data and 10 000 data sets simulated according to the no-admixture model, with parameter values sampled from the prior. This distribution was then compared with 100 distributions obtained for data sets simulated according to the no-admixture model, whose distances were measured against the same 10 000 data sets used with the observed data. We found that the observed data are within the set of distributions obtained with simulated data sets (Figure 6a). We repeated this procedure with the second set of priors used and, as Figure 6b shows, the distribution of observed distances is very different from the distances obtained with the model without admixture using these new priors. Although this cannot be considered as a proof that the true model has been found, this is a strong suggestion that the population split without admixture model of Figure 6a captures important aspects of the *I. lusitanicum* demographic history. As a further analysis, we looked at the posterior distributions for the demographic parameters. These have peaks within the prior limits for all parameters except the ancestral population size, which has a peak close to the upper limit (Supplementary Figures 5 and 6), suggesting that the most likely parameter values are within the prior limits.

Deviations from model assumptions. A second caveat is that the current implementation of the method is based on a varied but still limited set of demographic models and simplifying assumptions (Bray *et al*., 2009b). For instance, the models assume that all loci evolve according to the stepwise mutation model and have the same mutation rate, which may be unrealistic for certain real data sets. Also, the models do not take into account other demographic events such as bottlenecks and expansions, which are likely to have occurred in many species and which may influence our ability to separate scenarios. The analysis of data sets generated under a two-phase mutation model, with variation of the mutation rate among loci and including

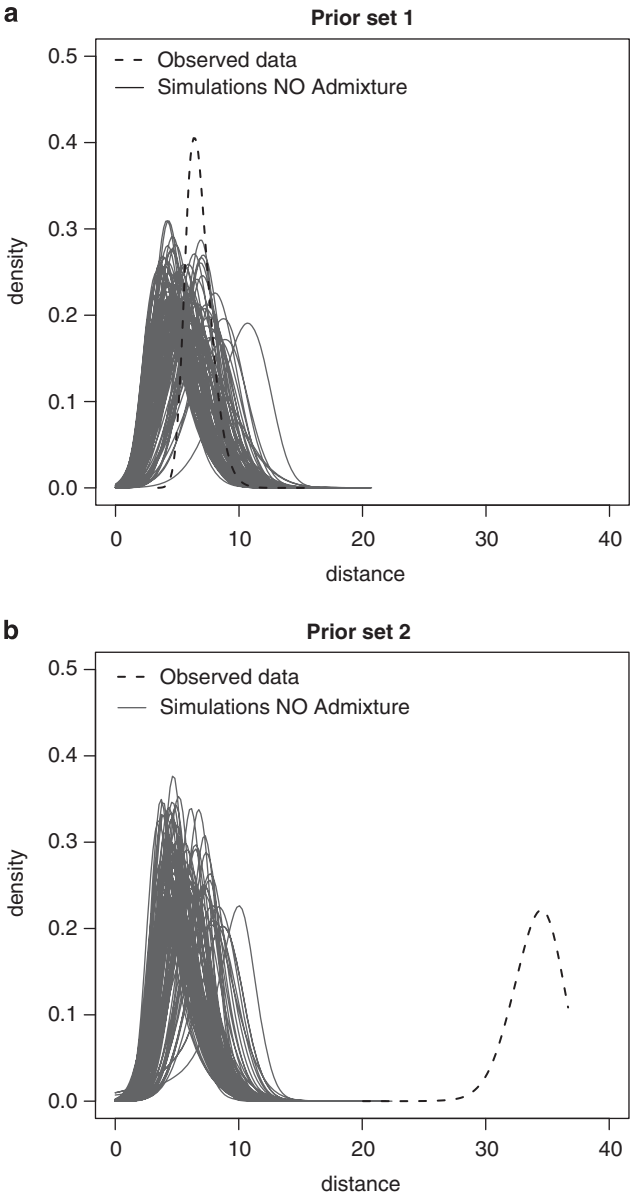

Figure 6 Comparison of the distance distributions between the *I. lusitanicum* data set and simulated data under the population split admixture model (NO admixture), with the distance distributions obtained for 100 data sets simulated under the most likely model (NO admixture). The simulated and observed data sets were compared with the same 10 000 simulations under the NO admixture model, with the parameters values sampled from the prior distributions. This model was selected as the most likely given the *I. lusitanicum* data. (**a**) Analysis with the first prior set tested; (**b**) analysis with the second prior set tested (see text for details).

population bottlenecks, suggest that the ABC model-choice procedure is robust, at least, to the deviations tested here. It would require a further study to determine under which conditions the models examined here are robust, and when increasing the complexity of the models would be useful and necessary. It is noteworthy that, in principle, the ABC model-choice framework allows the inclusion of more complex microsatellite mutation models, such as the generalized stepwise mutation model, which were shown to work well (Guillemaud *et al.*, 2009).

Prior distributions. The third caveat is related with the specification of the prior distributions, which is a general problem in Bayesian model choice. This is a critical point that has been recently discussed and studied by Guillemaud *et al.* (2009). Different models may appear as more or less likely depending on the range of the parameter values and the weight given to different parameter values specified by the priors. However, in most ABC algorithms there is a tradeoff between the width of the priors and the number of simulations needed to obtain a good approximation for the posterior. Thus, the analysis should be repeated with different sets of priors to assess its effect on the posterior. This is exemplified here, as the data set from *I. lusitanicum* was analyzed under two different prior sets. Both prior sets lead to posteriors that favored the population split without admixture model (Figure 5). However, when we performed the distance analysis (Figure 6) we found that the range of parameters specified by one of the prior sets did not fit the observed data. These results illustrate that more work is required on this general issue of selection of priors in model-choice approaches.

Choice of summary statistics. As mentioned earlier, Robert *et al.* (2011) and Didelot *et al.* (2011) have noted that model choice under ABC may be sensitive to the summary statistics (see also Grelaud *et al.*, 2009). Didelot *et al.* (2011) point out that for summary statistic S, computed from data x, to be sufficient one needs the distribution of the data conditioned on the summary statistic to be independent of the parameters ($p(x|S,\theta)=p(x|S)$). In the context of model choice, for any S (sufficient or not)

$$p(x|M) = P(x, S|M) = P(S|M)P(x|S, M)$$

The Bayes factor is then

$$\frac{p(x|M_1)}{p(x|M_2)} = \frac{P(S|M_1)P(x|S, M_1)}{P(S|M_2)P(x|S, M_2)}$$

Standard ABC model choice ignores the ratio $P(x|S,M_1)/P(x|S,M_2)$, and will therefore only be accurate if this is 1, that is, for a given summary statistic S, the distribution of data sets consistent with it is the same under both models (a similar treatment is in Robert *et al.*, 2011). Of course, there is no guarantee that this is the case, given that for most cases the probabilities $P(x|S,M_1)$ and $P(x|S,M_2)$ are unknown, and therefore one should be cautious. However, if one regards Bayesian model choice as a classifying procedure, as implied in our ROC analysis, then a purely pragmatic argument behind the use of the ABC approach is that we can examine its performance, and compare it with other classifiers (not necessarily Bayesian), simply through simulation. Our results (and those of Guillemaud *et al.*, 2009; Beaumont, 2010; and Bazin *et al.*, 2010) indicate that it works quite reliably, at least for the models under consideration. However, it may well be the case that our estimated Bayes factors are not identical to those that would be obtained under a full-likelihood analysis.

Other approaches to Bayesian model choice have been suggested. It should be noted that the Bayesian Information Criterion (BIC) is an asymptotic approximation to the Bayes factor (Robert, 2007). Although convenient in simple cases where only maximum likelihood estimates are available, there are a number of drawbacks to the use of BIC (Robert, 2007). In particular, it requires unimodal distributions, and the dependence on the priors for parameters within the models is lost. Recently, an ABC model choice using the Deviance Information Criterion has been suggested by François and Laval (2011), and appears to work well. In addition, classification and calibration methods in machine learning could be useful to predict the correct model for real data. In that case, the training sets would comprise data sets simulated under alternative models. Further work would be necessary to understand whether such approaches would be reliable using the full data sets or selecting a set of summary statistics as in the ABC framework.

Conclusion

In conclusion, we believe that this study contributes to a better understanding of the power of ABC methods as model-choice procedures, which is crucial as ABC methods are starting to be widely used in population genetics and other areas (Ratmann *et al.*, 2009; Bertorelle *et al.*, 2010; Beaumont, 2010; Csilléry *et al.*, 2010). We focused on models with either one or two admixture events and with up to four different populations. Our results suggest that it is possible to separate the effect of admixture from that of shared polymorphism. This is particularly important as admixture events are likely to have occurred in many species after the last glaciations during the colonization of new regions from several refugia or when populations encountered habitats that were already occupied (see, for example, Chikhi *et al.*, 2002; Fraser and Bernatchez, 2005). Admixture is also likely to have happened during the domestication of plants and animals and is still an ongoing process between breeds (see, for example, Bray *et al.*, 2009a). Moreover, admixture has been invoked in a number of genetic studies based on clustering methods. These methods (Pritchard *et al.*, 2000; Corander *et al.*, 2004) are very useful and have been very popular in the last decade to group individuals according to their genotypes under relatively simple population genetic models. However, the admixture parameter provided by these methods is in most cases of difficult biological interpretation, and currently it cannot discriminate shared polymorphism from proper admixture, as we saw here for the *I. lusitanicum* data. The main reason is that the demographic and evolutionary history of the populations is not explicitly modeled. For instance, the fact that populations may have different effective sizes is not taken into account in clustering methods. More work is required to find the situations where clustering and ABC methods are best applied. The former appears to be more suited for cases of ongoing gene flow, and the latter when ancient admixture and population split events have been important and hence need to be explicitly modeled. Regarding the detection of admixture events, some improvements are likely to come from the information about linkage disequilibrium, as admixture is known to generate linkage disequilibrium (Nordborg and Tavaré, 2002). The use of summary statistics based on the statistical association of alleles at different loci may thus prove very useful to separate scenarios with different numbers of admixture events, and perhaps to separate admixture from gene flow models. We clearly look forward to seeing these improvements in the next few years.

DATA ARCHIVING

Data deposited at Dryad: doi:10.5061/dryad.7tb527m3.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank three anonymous reviewers for their constructive comments. Also, we thank João Lopes and Franck Jabot for helpful discussions. The analyses were performed using the High-Performance Computing Centre (HERMES, FCT Grant H200741/re-equip/2005). This work was supported by SFRH/BD/22224/2005 granted to VCS by 'Fundação Ciência e Tecnologia' (FCT (Portuguese Science Foundation)). MMC is funded by the FCT Project PTDC/BIA-BDE/69769/2006. LC is funded by the FCT Projects PTDC/BIA-BDE/71299/2006 and PTDC/BIA-BEC/100176/2008, by the 'Institut Français de la Biodiversité', 'Programme Biodiversité des îles de l'Océan Indien' N. CD-AOOI-07-003, and by the 'Laboratoire d'Excellence (LABEX)' entitled TULIP (ANR-10-LABX-41). We also thank the Egide Alliance Programme (Project number: 12130ZG to LC and MAB) for funding visits between Toulouse and Reading.

Alves MJ, Coelho MM (1994). Genetic variation and population subdivision of the endangered Iberian cyprinid *Chondrostoma lusitanicum*. *J Fish Biol* **44**: 627–636.

Bazin E, Dawson KJ, Beaumont MA (2010). Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. *Genetics* **185**: 587–602.

Beaumont M (2010). Approximate Bayesian computation in evolution and ecology. *Annu Rev Ecol Evol Syst* **41**: 379–406.

Beaumont MA (2008) Joint determination of tree topology and population history. In: Matsumura S, Forster P, Renfrew C (eds) *Simulation, Genetics, and Human Prehistory*. McDonald Institute for Archaeological Research: Cambridge. pp 135–154.

Beaumont MA, Nielsen R, Robert C, Hey J, Gaggiotti O, Knowles L *et al.* (2010). In defence of model-based inference in phylogeography. *Mol Ecol* **19**: 436–446.

Beaumont MA, Zhang W, Balding DJ (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**: 2025–2035.

Bertorelle G, Benazzo A, Mona S (2010). ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Mol Ecol* **19**: 2609–2625.

Bray T, Chikhi L, Sheppy A, Bruford M (2009a). The population genetic effects of ancestry and admixture in a subdivided cattle breed. *Anim Genet* **40**: 393–400.

Bray T, Sousa V, Parreira B, Bruford M, Chikhi L (2009b). 2BAD: an application to estimate the parental contributions during two independent admixture events. *Mol Ecol Resour* **10**: 538–541.

Calabrese P, Sainudiin R (2005). Models of microsatellite evolution. In: Nielsen R (ed) *Statistical Methods in Molecular Evolution*. Springer: New York. pp 289–305.

Chikhi L, Bruford MW, Beaumont MA (2001). Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics* **158**: 1347–1362.

Chikhi L, Nichols RA, Barbujani G, Beaumont MA (2002). Y genetic data support the neolithic demic diffusion model. *Proc Natl Acad Sci USA* **99**: 11008–11013.

Choisy M, Franck P, Cornuet JM (2004). Estimating admixture proportions with microsatellites: comparison of methods based on simulated data. *Mol Ecol* **13**: 955–968.

Corander J, Waldmann P, Marttinen P, Sillanpaa M (2004). BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics* **20**: 2363–2369.

Cornuet J, Luikart G (1996). Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* **144**: 2001–2014.

Cornuet J, Santos F, Beaumont M, Robert C, Marin J, Balding D *et al.* (2008). Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics* **24**: 2713–2719.

Csilléry K, Blum M, Gaggiotti O, François O (2010). Approximate Bayesian computation (ABC) in practice. *Trends Ecol Evol* 410–418.

Didelot X, Everitt R, Johansen A, Lawson D (2011). Likelihood-free estimation of model evidence. *Bayesian Analysis* **6**: 49–76.

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, Estoup A, Cornuet JM (2005). Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics* **169**: 1727–1738.

Fagundes NJR, Ray N, Beaumont M, Neuenschwander S, Salzano FM, Bonatto SL *et al.* (2007). Statistical evaluation of alternative models of human evolution. *Proc Natl Acad Sci USA* **104**: 17614–17619.

Fawcett T (2006). An introduction to ROC analysis. *Pattern Recog Lett* **27**: 861–874.

François O, Laval G (2011). Deviance Information Criteria for model selection in approximate Bayesian computation. *Stat Appl Genet Mol* **10**: 33.

Fraser D, Bernatchez L (2005). Allopatric origins of sympatric brook charr populations: colonization history and admixture. *Mol Ecol* **14**: 1497–1509.

Goldstein DB, Chikhi L (2002). Human migrations and population structure: what we know and why it matters. *Annu Rev Genomics Hum Genet* **3**: 129–152.

Grelaud A, Marin J, Robert C, Rodolphe F, Tally F (2009). Likelihood-free methods for model choice in Gibbs random fields. *Bayesian Analysis* **3**: 427–442.

Guillemaud T, Beaumont M, Ciosi M, Cornuet J, Estoup A (2009). Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data. *Heredity* **104**: 88–99.

Hey J, Machado CA (2003). The study of structured populations–new hope for a difficult and divided science. *Nat Rev Genet* **4**: 535–543.

Johnson J, Omland K (2004). Model selection in ecology and evolution. *Trends Ecol Evol* **19**: 101–108.

Marjoram P, Molitor J, Plagnol V, Tavare S (2003). Markov chain Monte Carlo without likelihoods. *Proc Natl Acad Sci USA* **100**: 15324–15328.

Marjoram P, Tavaré S (2006). Modern computational approaches for analysing molecular genetic variation data. *Nat Rev Genet* **7**: 759–770.

Nei M (1978). Estimation of average heterozygosity and genetic distance from a small sample of individuals. *Genetics* **89**: 583–590.

Nielsen R, Wakeley J (2001). Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**: 885–896.

Nordborg M, Tavaré S (2002). Linkage disequilibrium: what history has to tell us. *Trends Genet* **18**: 83–90.

Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol* **16**: 1791–1798.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Ratmann O, Andrieu C, Wiuf C, Richardson S (2009). Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proc Natl Acad Sci USA* **106**: 10576–10581.

Robert C (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Verlag: New York.

Robert C, Cornuet J, Marin J, Pillai N (2011). Lack of confidence in ABC model choice. *Proc Natl Acad Sci USA* **108**: 15112–15117.

Sing T, Sander O, Beerenwinkel N, Lengauer T (2005). ROCR: visualizing classifier performance in R. *Bioinformatics* **21**: 3940–3941.

Sousa V, Penha F, Collares-Pereira MJ, Chikhi L, Coelho MM (2008). Genetic structure and signature of population decrease in the critically endangered freshwater cyprinid *Chondrostoma lusitanicum*. *Conserv Genet* **9**: 791–805.

Sousa VC, Fritz M, Beaumont MA, Chikhi L (2009). Approximate Bayesian computation without summary statistics: the case of admixture. *Genetics* **181**: 1507–1519.

Storz JF, Beaumont MA (2002). Testing for genetic evidence of population contraction and expansion: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. *Evolution* **56**: 154–166.

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

Wang S, Ray N, Rojas W, Parra M, Bedoya G, Gallo C *et al.* (2008). Geographic patterns of genome admixture in Latin American Mestizos. *PLoS Genetics* **4**: e1000037.

Wegmann D, Leuenberger C, Excoffier L (2009). Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics* **182**: 1207–1218.

Weir BS, Cockerham CC (1984). Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)