Inferring past demographic changes from contemporary genetic data: a simulation-based evaluation of the ABC methods implemented in DIYABC

Cabrera, A. A. & Palsbøll, P. J

Inferring the demographic history of species and their populations is crucial to understand their contemporary distribution, abundance, and adaptations. The high computational overhead of likelihood-based inference approaches severely restricts their applicability to large data sets or complex models. In response to these restrictions, Approximate Bayesian Computation (ABC) methods have been developed to infer the demographic past of populations and species. Here we present the results of an evaluation of the ABC-based approach implemented in the popular software package DIYABC using simulated data sets (mitochondrial DNA sequences, microsatellite genotypes and single nucleotide polymorphisms). We simulated population genetic data under five different simple, single-population models to assess the model recovery rates as well as the bias and error of the parameter estimates. The ability of DIYABC to recover the correct model was relatively low (0.49): 0.6 for the simplest models and 0.3 for the more complex models. The recovery rate improved significantly when reducing the number of candidate models from five to three (from 0.57 to 0.71). Among the parameters of interest, the effective population size was estimated at a higher accuracy compared to the timing of events. Increased amounts of genetic data did not significantly improve the accuracy of the parameter estimates. Some gains in accuracy and decreases in error were observed for scaled parameters (e.g., Neµ) compared to unscaled parameters (e.g., Ne and µ). We concluded that DIYABC-based assessments are not suited to capture a detailed demographic history, but might be efficient at capturing simple, major demographic changes.

Published in Molecular Ecology Resources, 2017
Introduction

Inferring the demographic history of con-specific populations is fundamental to our understanding of the impacts of past environmental and anthropogenic changes upon the distribution, abundance and adaptations of contemporary populations (Girod et al., 2011; Lorenzen et al., 2011). The demographic history can be inferred from the fossil record (e.g., Newbrey & Ashworth, 2004; Parducci et al., 2005). However, the collection of fossils (e.g., animal remains, pollen, environmental DNA) from historic populations is often challenging or impossible. Alternatively, the demographic history can be inferred from contemporary genetic data (Avise, 2004; Marjoram & Tavare, 2006; Lawton-Rauh, 2008).

Most population genetic approaches, aimed at inferring demographic history, utilize likelihoods (Griffiths & Tavare, 1994) and typically rely upon classical population genetic or coalescent theory in combination with Markov Chain Monte Carlo sampling (Beaumont & Rannala, 2004; Csillery et al., 2010). Likelihood-based methods are implemented in a variety of software, such as BATWING (Wilson & Balding, 1998; Wilson et al., 2003), MSVAR (Beaumont, 1999), IM and IMa (Hey & Nielsen, 2004b; Hey & Nielsen, 2007), LAMARC (Kuhner, 2006), VarEff (Nikolic & Chevalet, 2014) and BEAST (Drummond & Rambaut, 2007). The aforementioned approaches differ mainly in the assumptions they make regarding the underlying demographic model and the mutational processes generating genetic variation (Girod et al., 2011). However, the high computational overhead of likelihood-based approaches imposes severe constraints upon the amount of data and the complexity of the underlying models in the analysis. Although these computational issues might appear trivial, most current implementations of likelihood-based approaches are incapable of accommodating the large data sets emerging from the massively parallel sequencing methods (Marjoram & Tavare, 2006; Sunnaker et al., 2013). Consequently, the above issue has spurred the development of approximate, as opposed to full likelihood, methods capable of accommodating large data sets and complex models (Marjoram & Tavare, 2006). One popular implementation of such approximate methods is Approximate Bayesian Computation (ABC, Beaumont et al., 2002).

ABC is a Monte Carlo-based approach of approximate Bayesian statistical inference which employ summary statistics to estimate the posterior probability of a specific hypothesis, i.e., a specific combination of parameter values (Beaumont et al., 2002). ABC approach was first introduced to population genetics by Beaumont et al. (2002), and has gained in popularity, in part
due to the lower computational overhead enabling the analysis of larger data sets and more complex models (Beaumont et al., 2002; Aeschbacher et al., 2012).

A “generic” ABC-based estimation is comprised of the following steps (Csillery et al., 2010; Sunnaker et al., 2013): (i) an input parameter value is sampled at random from a pre-specified prior probability distribution and used to simulate data sets. (ii) Selected summary statistics are computed from the simulated data. (iii) The summary statistic values computed from the simulated data are compared to the summary statistic values computed from the empirical data. (iv) The input parameter value is retained if the simulated summary statistics are deemed close to the summary statistics of the empirical data. (v) Once a pre-specified number of simulated data sets have been retained, a posterior probability distribution is approximated from the distribution of accepted input parameter values.

During recent years, the use of ABC-based approaches has increased in population genetics to infer demographic history (e.g., Rovito, 2010; Hoffman et al., 2011; Lombaert et al., 2011; Row et al., 2011; Fontaine et al., 2012; Nadachowska-Brzyska et al., 2013; Phillips et al., 2013). Several software packages implement ABC approaches, such as DIYABC (Cornuet et al., 2008), ONeSAMP (Tallmon et al., 2008), PopABC (Lopes et al., 2009), ABC-SysBio (Liepe et al., 2010), ABCtoolbox (Wegmann et al., 2010), and several R-Cran packages (R-Development-Core-Team, 2013), such as ‘abc’ (Csillery et al., 2012) or “EasyABC” (Jabot et al., 2013). However, the assessment of the performance of these implementations of ABC to infer demographic history has so far been relatively limited (e.g., Cornuet et al., 2008; Cornuet et al., 2010; Robert et al., 2011; Pelletier & Carstens, 2014). The choice of which ABC implementation to employ depends upon the available computational resources, the nature of the molecular data and the complexity of the estimation. DIYABC is one of the most commonly employed ABC-based software, has a user friendly graphical interface and executables for all major operating systems, such as, Microsoft Windows™, Linux and OS X™. DIYABC enables the comparison of multiple complex historical models and accommodates (in a single estimation) commonly employed molecular data types (Cornuet et al., 2008; Cornuet et al., 2014). Cornuet and coworkers conducted an evaluation of DIYABC regarding model selection and parameter estimation of the effective population sizes, population divergence time and admixture rates from microsatellite genotypes. Employing a single replicate per demographic model, the authors concluded that DIYABC performed well, i.e., the “true” model was recovered at high rates and the estimates of the effective population size and admixture rates were accurate. However, the assessment also revealed some degree of bias in the estimates of temporal parameters, particularly the timing of the most recent event. In a subsequent assessment, Cornuet et al. (2010) looked at the effects of combining nuclear microsatellite
genotypes, mitochondrial DNA sequences and/or nuclear autosomal DNA sequences. The assessment revealed substantial differences in model recovery rates among different molecular data types. In line with expectations, data sets including data from different molecular data types performed comparatively better (Cornuet et al., 2010). The assessments undertaken by Cornuet et al., (2008; 2010) revealed some bias in the estimation of key parameters, which was inferred by the authors to stem from differences in the degree of “informativeness” among different molecular data types.

A common use of ABC approaches (DIYABC in particular) is to assess past population size changes in a single isolated population. Inferences regarding past population size changes from contemporary population genetic data has been widely employed to assess the effects of past environmental changes, such as the most recent glaciations (e.g., O’Corry-Crowe, 2008; Bisconti et al., 2011; González-Wevar et al., 2013) and impacts of human exploitation (e.g., Hoffman et al., 2011; Fontaine et al., 2012). Here, we presented the results of an evaluation of DIYABC’s ability to recover past population demographic events as well as the accuracy and error of the targeted parameter estimates (i.e., effective population size ($N_e$), time of population size changes ($T$), and mutation rates ($\mu$)). We aimed specifically at population size changes in the relatively distant past, i.e., during the Last Glacial Maximum (LGM) some 26,500 to 15,000 years ago (Clark et al., 2009), which has been the target of many studies. The climatic and ecological changes during and after the LGM have exerted significant and lasting effects upon the distribution and levels of genetic diversity we observe in contemporary populations across a wide range of habitats (Hewitt, 1996, 2000; Hofreiter & Stewart, 2009; Rovito, 2010; Fontaine et al., 2012; Cahill et al., 2013; Nadachowska-Brzyska et al., 2013). We simulated data representing the three most commonly employed molecular data types in DIYABC-based estimations (mitochondrial DNA sequences, microsatellite genotypes and single nucleotide polymorphism (SNP), e.g., Rovito, 2010; Lombaert et al., 2011; Fontaine et al., 2012; Lippens et al., 2017) under five different, and increasingly complex models of past population size changes.

Methods

We simulated data sets of population genetic data under five different demographic models. These population genetic data sets were subsequently employed to evaluate the ability of DIYABC to recover past demographic events and estimate the parameter values of interest.
Generation of pseudo-empirical data sets

Simulated data was generated under five different demographic models including growing, declining, and stable populations. The changes in population size were modeled to have taken place during the LGM in the Northern Hemisphere (Clark et al., 2009). The five models were: (CON) constant population size, (DEC) a single instantaneous decrease in population size, (INC) a single instantaneous increase in population size, (INCDEC) a single instantaneous increase followed by a single instantaneous decrease in population size, and finally (DECINC) a single instantaneous decrease followed by a single instantaneous increase in population size (Figure 1, Table 1). In the case of the models DEC and INC, the change in population size occurred after the LGM (i.e., approximately 11,250 years ago), and in the case of the models INCDEC and DECINC the changes in population size occurred just before and after the LGM (i.e., approximately 22,500 and 11,250 years ago respectively). We employed a generation time of 15 years, i.e., in the range of generation times estimated for large, long-lived mammals, such as marine mammals and primates (e.g., Pacifici et al., 2013).

![Figure 1. Schematic representation of the demographic models evaluated.](image)

The areas of the figures represent changes in population size through time. Effective population size \(N_e\) is represented by \(N_i\) (i.e., \(N_{01}, N_{11}, N_{13}, N_{114}, N_{24}, N_{15},\) and \(N_{215}\)). Most recent events are at the top of the figure. The time, \(T\), (in number of generations) when a demographic change occurred is as \(T_1(4-5), T_1(2-3), T_1(0-3), T_2(4-5),\) and \(T_2(0-5)\).

Each simulated data set was comprised of a sample of 100 individuals from one single panmictic population using DIYABC v. 2.0.4 (Cornuet et al., 2008; Cornuet et al., 2014). The simulated data sets were comprised of either: (i) mitochondrial (haploid) DNA sequences of 450
nucleotides; (ii) 20 diploid microsatellite genotypes; (iii) a combination of 20 diploid microsatellite genotypes and mitochondrial DNA sequences of 450 nucleotides; or (iv) 1,000 SNP genotypes. The mutation model and rates used to simulate each molecular data type are listed in Table 2. The values are similar to those estimated for large mammals (Hasegawa et al., 1985; Dallas, 1992; Tamura & Nei, 1993; Weber & Wong, 1993; Reyes et al., 1998; Schlötterer et al., 1998; Excoffier & Yang, 1999; Pesole et al., 1999; Estoup et al., 2002; Carapelli et al., 2008; Santos et al., 2008). Nuclear genotypes were unlinked and selectively neutral.

In order to evaluate the effect of effective population size and different proportional changes in effective population size, we simulated two or three different combinations of $N_e$ for each demographic model (two for the models INCDEC and DECINC and three for the models CON, DEC, and INC, denoted as submodels A to M in Table 1). Ten replicated simulated data sets were generated for each combination of molecular data types (four combinations, i to iv above, Figure 2) and $N_e$ (13 in total, submodels A to M above, Figure 2). In total, this amounted to 520 simulated data sets. Below we refer to these simulated data sets as the “pseudo-empirical” data sets.

**Figure 2. Schematic representation of the pseudo-empirical data sets generated per demographic model and genetic data set combination.** The first column represents the demographic model evaluated. The CON model is used as an example. The second column represents the different $N_e$ evaluated. We referred to these combinations of $N_e$ as submodels in the text. The third column represents the different genetic data sets evaluated: mitochondrial DNA sequences (MIT), microsatellite genotypes (MIC), a combination of mitochondrial DNA sequences and microsatellite genotypes (MITMIC) and SNP genotypes (SNP). For each combination of genetic data set per demographic model, ten simulated data sets were generated. The models CON, DEC and INC have the same schematic representation. The models INCDEC and DECINC have only two submodels. $N_{01}$ is the effective population size.
Defining hypothesized models, prior information and summary statistics

The model selection and parameter estimations for each pseudo-empirical data set was conducted employing DIYABC v.2.0.4 (Cornuet et al., 2014) following the procedure common to most published studies based upon DIYABC (e.g., Lombaert et al., 2011; Fontaine et al., 2012; Louis et al., 2014a). The five models, which guided the generation of the pseudo-empirical data sets (i.e., CON, DEC, INC, INCDEC, DECINC in Figure 1), were employed as candidate models during the DIYABC estimations. The prior distributions for the demographic and genetic parameter estimations are listed in Table 3 and 4, respectively. Employing the same simulation software to both generate and analyze the pseudo-empirical data implies that the pseudo-empirical data sets represent the best-case scenario. In other words, empirical data sets collected from natural populations that have undergone demographic changes similar to those simulated in our assessment, are likely subject to greater uncertainty and bias compared to the “pseudo-empirical” data employed in this study.

Table 1. Demographic models and model parameterization for the generation of the “pseudo-empirical data sets”

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Model 1: CON</th>
<th>Model 2: DEC</th>
<th>Model 3: INC</th>
<th>Model 4: INCDEC</th>
<th>Model 5: DECINC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
</tr>
<tr>
<td>$N_0$</td>
<td>100</td>
<td>1,000</td>
<td>10,000</td>
<td>100</td>
<td>1,000</td>
</tr>
<tr>
<td>$N_{12}$</td>
<td>10,000</td>
<td>10,000</td>
<td>1,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_{13}$</td>
<td>100</td>
<td>1,000</td>
<td>1,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_{14}$</td>
<td>10,000</td>
<td>10,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_{24}$</td>
<td>100</td>
<td>1,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_{15}$</td>
<td>100</td>
<td>1,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_{25}$</td>
<td>10,000</td>
<td>10,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$T_{(2-3)}$</td>
<td>750</td>
<td>750</td>
<td>750</td>
<td>750</td>
<td>750</td>
</tr>
<tr>
<td>$T_{(4-5)}$</td>
<td>750</td>
<td>750</td>
<td>750</td>
<td>750</td>
<td></td>
</tr>
<tr>
<td>$T_{(2-3)}$</td>
<td>1,500</td>
<td>1,500</td>
<td>1,500</td>
<td>1,500</td>
<td></td>
</tr>
</tbody>
</table>

Models are based on the five demographic models shown in Figure 1. $N_0$ is the current effective population size. $N_{ib}$, $N_{is}$, $N_{ibA}$, $N_{isA}$, $N_{ibB}$, and $N_{isB}$ are historical effective population sizes. $TT$, $T1$, $T2$: time in generations (Assuming a generation time of 15 years, $TT$ and $T1$ emulate an event after the LGM, 11,250 years ago, and $T2$ emulates an event before the LGM, 22,500 years ago). Letters A to M represent different parameters of the demographic models that were evaluated.

The summary statistics employed in ABC estimation are tabulated in Table 4 by molecular marker type. The selected summary statistics were those typically used in previous empirical studies (e.g., Lombaert et al., 2011; Fontaine et al., 2012; Louis et al., 2014a).
ABC computation

As the basis for the ABC estimation, we generated $10^6$ simulated data sets for each demographic model and combination of parameter values from prior distributions as described above. The $10^4$ prior parameter value combinations which yielded summary statistics closest to the values estimated from the pseudo-empirical data set were retained and the posterior probability distributions were estimated (Beaumont et al., 2002; Cornuet et al., 2008). These $10^4$ retained simulations represented the default and fixed threshold value of 0.01 in DIYABC. The posterior probability distributions were employed for the model selection and parameter estimation. In our assessment, we treated the model selection and the parameter estimation as two separate assessments.

Table 2. Mutation model parameterization for the generation of the “pseudo-empirical data sets”

<table>
<thead>
<tr>
<th>Autosomal microsatellite genotypes</th>
<th>Mitochondrial DNA sequences</th>
<th>SNP genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutation model GSM</td>
<td>Mutation model HKY</td>
<td></td>
</tr>
<tr>
<td>MEAN-$\mu_{MIC}$ 1.00E-04</td>
<td>MEAN-$\mu_{MIT}$ 5.00E-06</td>
<td></td>
</tr>
<tr>
<td>INDL-$\mu_{MIC}$ GA[1E-5, 1E-3, 2]</td>
<td>INDL-$\mu_{MIT}$ GA[1E-7, 1E-5, 2]</td>
<td></td>
</tr>
<tr>
<td>MEAN-$P$ 0.22</td>
<td>MEAN-$K1$ 10</td>
<td></td>
</tr>
<tr>
<td>INDL-$P$ GA[0.01, 0.9, 2]</td>
<td>INDL-$K1$ GA[0.05, 0.2]</td>
<td></td>
</tr>
<tr>
<td>MEAN-$SNI$ 1.00E-08</td>
<td>% invariable sites 80</td>
<td></td>
</tr>
<tr>
<td>INDL-$SNI$ GA[1E-9,1E-04, 2]</td>
<td>Shape 2</td>
<td></td>
</tr>
</tbody>
</table>

The mutation model parameters of the autosomal microsatellite genotypes were the mutation rate ($\mu_{MIC}$) per locus per generation, the parameter determining the shape of the gamma distribution of individual genotypes mutation rate ($P$), and the single insertion nucleotide rate ($SNI$). The mutation rate of the microsatellite genotypes was based on published mutation rate estimates for mammals, which are in the range of $10^{-3}$ to $10^{-5}$ per locus per generation (Dallas, 1992; Weber & Wong, 1993). The length and nucleotide composition of the mitochondrial DNA sequences were simulated based on commonly used values for the mitochondrial DNA control region or D-loop (Reyes et al., 1998; Carapelli et al., 2008). The mutation model parameters for the mitochondrial DNA sequences were the mutation rate ($\mu_{MIT}$) in substitution per site per generation, the transition/transversion ratio ($K1$) parameter, the fraction of constant sites (% invariable sites) and the shape of the gamma distribution of mutations among sites (Shape). HKY is the Hasegawa-Kishino-Yano model (Hasegawa et al., 1985). GSM is the generalized stepwise mutation model (Estoup et al., 2002). The mutation rate value of the mitochondrial DNA sequences is within the range of mutation rates found in literature for the control region of the mitochondrial DNA in mammals (Tamura & Nei, 1993; Excoffier & Yang, 1999; Pesole et al., 1999; Santos et al., 2008). MEAN denotes mean value of the parameter and INDL denotes the individual locus value with a gamma distribution (GA) around mean. Inside brackets are the minimum, maximum and shape of the parameter. Sex ratio for all simulated populations was 0.5.
Model selection and assessment of the confidence of the recovered model

The posterior probability for each demographic model was estimated from the polytomous weighted logistic regression (Beaumont, 2008) from which the point estimate and the 95% confidence interval was computed for each candidate model as described by Cornuet et al. (2008). The most probable demographic model was deemed successfully recovered if the maximum posterior probability of the model was excluded from the 95% confidence intervals estimated for all other candidate models (Cornuet et al., 2008). The model selection was classified as “undefined” in those cases where the above criterion failed. The model selection was conducted twice for each pseudo-empirical data set, i.e., with three (CON, DEC and INC) or five (CON, DEC, INC, INCDEC, DECINC) candidate models, resulting in a total number of assessments at 1,040.

Table 3. Demographic models and model parameterization in the ABC analysis

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Model 1: CON</th>
<th>Model 2: DEC</th>
<th>Model 3: INC</th>
<th>Model 4: INCDEC</th>
<th>Model 5: DECINC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_{0}(1-5)$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
</tr>
<tr>
<td>$N_{1l2}$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
</tr>
<tr>
<td>$N_{1s3}$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
</tr>
<tr>
<td>$N_{1l4}$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
</tr>
<tr>
<td>$N_{2s4}$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
</tr>
<tr>
<td>$N_{1s5}$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
</tr>
<tr>
<td>$N_{2l5}$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
<td>UN$[100, 200,000]$</td>
</tr>
<tr>
<td>$T_{1(4-5)}$</td>
<td>UN$[100, 2,000]$</td>
<td>UN$[100, 2,000]$</td>
<td>UN$[100, 1,000]$</td>
<td>UN$[100, 1,000]$</td>
<td>UN$[1,000, 2,000]$</td>
</tr>
</tbody>
</table>

According to models from Figure 1: $N_0, N_{1l2}, N_{1s3}, N_{1l4}, N_{2s4}, N_{1s5},$ and $N_{2l5}$ most recent and historical effective population sizes. $T_{1(2-3)}, T_{1(4-5)}$ and $T_{2(4-5)}$, time in generations. UN: uniform distribution with a minimum and maximum parameter inside brackets.

We estimated the classification error (referred to as the type I error in DIYABC) and the misclassification error (referred to as the type II error in DIYABC) from 500 pseudo-observed data sets (hereafter named pods) for each candidate demographic model (Cornuet et al., 2010). The pods (i.e., simulated data sets for which we knew the true values of the parameters) were simulated from the initial prior distributions and compared against the original $10^6$ ABC simulated data sets to calculate the posterior parameter distributions. The classification error was estimated as the proportion of pods among which the true model was incorrectly excluded. The misclassification error was estimated as the proportion of pods among which an incorrect model was recovered.
Parameter estimation, bias and error

We estimated both unscaled (i.e., $N_e$, $T$ and $\mu$) and scaled (i.e., $N_e\mu$ and $T\mu$) parameters (Beaumont et al., 2002; Cornuet et al., 2008; Cornille et al., 2012) for 200 pseudo-empirical data sets of the 540, i.e., ten replicated data sets for five demographic models (submodels B, E, H, K and M, in Table 1) and four combinations of genetic data set. The five selected demographic models represented a single value of $N_e$ per demographic model. In order to assess the accuracy and error of the parameter estimates, we compared the estimated values against the input parameter values used to generate the pseudo-empirical data sets (Table 1, supplementary material Figure S1.A and Tables S1.A-S1.E).

<table>
<thead>
<tr>
<th>Mutation model parameterization and summary (test) statistics</th>
<th>Microsatellite genotypes</th>
<th>Mitochondrial DNA sequence</th>
<th>SNP genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsatellite genotypes</td>
<td>GSM</td>
<td>Mutation model</td>
<td>HKY</td>
</tr>
<tr>
<td>Mutation model parameterization and summary statistics</td>
<td></td>
<td>No mutation model was required</td>
<td></td>
</tr>
<tr>
<td>MEAN-$\mu_{MIC}$</td>
<td>UN[1.0 E-5, 1.0 E-3]</td>
<td>MEAN-$\mu_{MIT}$</td>
<td>UN[1.00E-7,1.00E-5]</td>
</tr>
<tr>
<td>INDL-$\mu_{MIC}$</td>
<td>GA[1.00E-5, 1.00E-2, 2]</td>
<td>INDL-$\mu_{MIT}$</td>
<td>GA[1.00E-7,1.00E-5,2]</td>
</tr>
<tr>
<td>MEAN-\text{P}</td>
<td>UN[0.1, 0.3]</td>
<td>MEAN-\text{K1}</td>
<td>UN[0.050,20]</td>
</tr>
<tr>
<td>INDL-P</td>
<td>GA[0.01, 0.9, 2]</td>
<td>INDL-\text{K1}</td>
<td>GA[0.050,20,2]</td>
</tr>
<tr>
<td>MEAN-\text{SNI}</td>
<td>LU[1.00 E-9, 1.00E-4]</td>
<td>% invariable sites</td>
<td>80</td>
</tr>
<tr>
<td>INDL-\text{SNI}</td>
<td>GA[1.00 E-9, 1.00 E-4, 2]</td>
<td>Shape</td>
<td>2</td>
</tr>
</tbody>
</table>

The mutation model parameters of the autosomal microsatellite genotypes were the mutation rate ($\mu_{MIC}$) per locus per generation, the parameter determining the shape of the gamma distribution of individual loci mutation rate ($P$), and the single insertion nucleotide rate ($\text{SNI}$). The mutation model parameters for the mitochondrial DNA sequences were the mutation rate ($\mu_{MIT}$) in substitution per site per generation, the transition/transversion ratio ($K1$) parameter, the fraction of constant sites (% invariable sites) and the shape of the gamma distribution of mutations among sites (Shape). HKY is the Hasegawa-Kishino-Yano model (Hasegawa et al., 1985). GSM is the generalized stepwise mutation model (Estoup et al., 2002). MEAN denotes mean value of the parameter and INDL denotes the individual locus value. GA: gamma distribution around mean. UN: uniform distribution. LU: Log-uniform distribution. Inside brackets are the minimum and maximum in uniform distributions and minimum, maximum and shape of the parameter in gamma distribution.

The accuracy and error of the parameter estimates was estimated for 20 pseudo-empirical data sets, i.e., one data set for each of the five demographic models (submodels B, E, H, K and M) and four combinations of genetic data set. The pseudo-empirical data sets were selected at random...
among the pseudo-empirical data sets in which the true model was correctly recovered. We estimated the accuracy as the average relative bias (ARB) and the error as the relative median of the absolute error (RMAE). We employed the same ARB and RMAE measures to assess the effect of the pseudo-empirical data upon the posterior probability distributions (Cornuet et al., 2008).

Assessing the goodness-of-fit of the recovered demographic model

We assessed the goodness-of-fit of the recovered demographic model relative to the other four candidate demographic models for each of the 20 pseudo-empirical data sets described above (i.e., one data set for each of the five demographic models and the four kinds of genetic data sets).

The goodness-of-fit of each specific demographic model was estimated employing the model checking procedure implemented in DIYABC. The goodness-of-fit was based on the posterior distribution, which was used to estimate a posterior predictive distribution (Geisser & Eddy, 1979). The assessment was based upon $10^3$ new simulated data sets per candidate model and a new set of summary statistics that served as test statistics. The posterior distributions estimated from the pseudo-empirical data were employed as prior distributions for these new simulated data sets (Gelman et al., 2014). The posterior predictive $p$-value was estimated for each test statistic of the pseudo-empirical data sets (Cornuet et al., 2010). If the model was a good fit, the posterior predictive $p$-value was expected to be close to 0.5 (Gelman, 2013). In order to evaluate the risk of overestimating the quality of the fit by using as test statistics the same statistics as in the model selection (Cornuet et al., 2010), we employed two sets of test statistics: the same statistics employed during the model selection, as well as a new, different set of statistics. During the first assessment, we employed the statistics listed in Table 4, which were the same test statistics employed as summary statistics in the model selection. The second assessment was based upon test statistics that were different from the summary statistics in the model selection. In the latter case, the odd numbered statistics in Table 4 were employed as summary statistics in the model selection and the even numbered statistics as test statistics.

Results

Model choice: recovery rates

The overall ability of DIYABC to recover the correct model was low (255/520 assessments or 0.49, Figure 3). Among the five demographic models evaluated in this study (see Figure 1), three demographic models were “simple” (CON, INC and DEC) and two “complex” (INCDEC and
DECINC). The correct model was recovered at an estimated rate of 0.6 in the case of the simple models (Figure 3). In contrast, the correct model was recovered at a rate of 0.3 in the case of the complex models (Figure 3). The average classification error was substantially lower among the three simple models (between 0.44 and 0.47, Table 5) compared to the complex models (between 0.62 and 0.81, Table 5). In contrast, the average misclassification error was lower for the two complex models (between 0.07 and 0.09) compared to the simple models (between 0.15 and 0.20, Table 5). A chi-square test of goodness-of-fit was performed to assess if the five models were recovered at equal rates. When the assessment included all possible outcomes (i.e., correct, incorrect, undefined, $n = 520$), we observed a significant difference in the recovery rates $\chi^2 (df = 8) = 99.49$, $p <0.01$, as was the case when the undefined category was excluded from the assessment ($n = 494$) $\chi^2 (df = 4) = 92.20$, $p <0.01$.

In total, 130 pseudo-empirical data sets were analyzed for each of the four genetic data set combinations (see Figure 3). The recovery rate was higher for data sets comprised of SNP genotypes (0.55) in comparison to mitochondrial DNA sequences (0.47), microsatellites genotypes, or a combination of mitochondrial DNA sequences and microsatellite genotypes (0.47). The estimated recovery rates differed significantly among the four types of genetic data sets, $\chi^2 (df = 8) = 17.01$, $p <0.01$, when all outcomes were included in the assessment (i.e., correct, incorrect and undefined, $n = 520$). However, when we excluded undefined ($n = 26$, for a total of 494 assessments), we were unable to detect a statistically significant difference in the recovery rate among the four different types of genetic data sets ($\chi^2 (df = 4) = 2.23$, $p = 0.53$).

<table>
<thead>
<tr>
<th>Model</th>
<th>Classification error</th>
<th>Misclassification error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Molecular data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MIT</td>
<td>MIC</td>
</tr>
<tr>
<td>CON</td>
<td>0.653</td>
<td>0.557</td>
</tr>
<tr>
<td>DEC</td>
<td>0.493</td>
<td>0.511</td>
</tr>
<tr>
<td>INC</td>
<td>0.586</td>
<td>0.477</td>
</tr>
<tr>
<td>INCDEC</td>
<td>0.801</td>
<td>0.931</td>
</tr>
<tr>
<td>DECINC</td>
<td>0.633</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Table 5. Confidence in model selection for the different demographic models and molecular markers

Confidence in model selection was assessed employing the classification and misclassification error (referred as type I and II error in the software). Each estimate represents the mean value among three pseudo-empirical data sets selected at random per demographic model (CON, DEC, INC, INCDEC and DECINC) per genetic data set combination (MIT, MIC, MITMIC and SNP). All estimates were based on the logistic approach by Cornuet et al. (2010). Results are based on 500 simulated pseudo-observed data sets (pods) per model with parameter values drawn from the same distributions as the prior distributions given in Table 3 and 4.
The model recovery rates were highest for the DEC model (0.8, 0.9 and 0.7, Figure 3D, 3E and 3F) regardless of which combination of genetic data and demographic parameters was employed to generate the pseudo-empirical data. The lowest model recovery rates (0.2 and < 0.1, Figure 3J and 3K) were observed for the DECINC model.

Figure 3. Best supported model among five different competing demographic models in DIYABC. The columns represent the five demographic models evaluated. The rows correspond to the different combinations of demographic parameter values evaluated: three combinations for CON, DEC and INC models, and two combinations for INCDEC and DECINC model. The simulated parameters for each demographic model (A-M) are shown on top of each figure and correspond to Table 2. The bars with a gray dot represent the expected demographic model based on the pseudo-empirical data set. The term “undefined” was employed when none of the candidate demographic models was recovered as the most probable demographic model. Each demographic model was tested for four genetic data sets: MIT, MIC, MITMIC and SNP.
The model recovery rate among the simple models increased from 0.57 to 0.71 when the total number of candidate models was reduced from all five models to the three simplest models (Figure 3D, 3E, 3F and Figure 4D, 4E, 4F). This increase was observed for all combinations of genetic data and parameter values ($N_e$ and $T$) representing 360 pseudo-empirical data sets. The largest increase in model recovery rate, due to the reduction in the total number of candidate modes, was observed for the INC model (from 0.5 to 0.84).

Figure 4. Best supported model among three competing demographic models in DIYABC. The columns represent the five demographic models evaluated. The rows correspond to the different combinations of demographic parameters tested: three combinations for CON, DEC and INC model and two combinations for INCDEC and DECINC model. The simulated parameters for each demographic model (A-M) are shown on top of each figure and correspond to Table 2. The bars with a gray dot represent the expected model based on the pseudo-empirical data sets. The term “undefined” was employed when none of the candidate demographic models was recovered as the most probable demographic model. The demographic models J, K, L and M represent models in which the true model was not included in the comparison set. Each demographic model was tested for four genetic data sets: MIT, MIC, MITMIC and SNP.
When the true model was a complex model but the candidate models only contained the simple models, the most recent event of the true complex model was recovered with the highest rate. For example, the INC model when the true model was the DECINC model and the DEC model when the true model was the INCDEC model. In those cases, the recovery rates were between 0.6 and 0.9 (Figure 4J, 4K, 4L and 4M). For instance, among the 80 cases where the true model was the INCDEC demographic model, the DEC model was recovered as the most probable model 48 times, the CON model 30 times, the INC model once. One instance was undefined (Figure 4J and 4K). The outcomes were similar for the DECINC model (Figure 4L and 4M).

**Estimating effective population size, time and mutation parameter values**

The effective population size (e.g., $N_{o2}$, $N_{o3}$, $N_{1l2}$ and $N_{1s3}$, Figure 5) was generally estimated with higher accuracy and lower error compared to $T$ (e.g., $T_{T2}$ and $T_{T3}$, Figure 5). The smallest $N_e$ (recent $N_e$ for DEC, and historic $N_e$ for INC, Figure 5) was estimated most accurately and precisely, irrespective of the nature of the genetic data.

Increasing the amount of genetic data (i.e., number of loci and/or type of genetic markers) improved the precision by reducing the credible interval of the parameter estimates ($N_{o2}$, $N_{o3}$, $N_{1l2}$, $N_{1s3}$, $T_{T2}$ and $T_{T3}$, Figure 5, Supplementary material FigureS1), but did not appear to improve the accuracy (Fig. 5, Supplementary material Figure S1 and Tables S1.A-F). The unscaled parameter values were estimated with higher error compared with the scaled parameters values (Figure 6, Supplementary material Figure S1.B-S1.E and Tables S1.A-E).

We assessed the effect of the genetic data and demographic history upon the parameter estimates. We compared the difference between the posterior estimates obtained with and without the pseudo-empirical genetic data (Supplementary materials, Tables S2.A-B, Figure S2.A-B). The average ARB (bias) for estimates of $N_e$ and $T$ was improved from 1.49 to 0.64 (Supplementary material Table S2.A). Excluding the CON model, the average improvement in ARB for $N_e$ (from 2.12 to 0.74) was larger compared to the improvement observed in estimates of $T$ (from 0.45 to 0.40). We observed a similar trend, in the case of error, where the average RMAE for estimates of $N_e$ and $T$ was improved from 0.34 to 0.29 (Supplementary material Table S2.A). Excluding the CON, the average improvement in ARB was larger for $N_e$ (from 0.37 to 0.27) compared to $T$ (from 0.33 to 0.31).
Figure 5. Inference of the effective population size and time parameters of pseudo-empirical data sets using DIYABC. Four genetic data set combinations are presented for two demographic models (top: population decline (DEC.E), and bottom: population increase (INC.H)) as described in Table 1. Each graphic presents the results of ten replicates per demographic model. The black circles represent the mode of the posterior probability distribution and the grey vertical line the 95% credible interval for each parameter. The true parameter value is shown with a horizontal line. Parameter notation are the time of the demographic event ($T_{(2-3)}$) in generations, the most recent effective population size ($N_{(2-3)}$), the historic effective population size before the demographic decline ($N_{(12)}$) and before the demographic increase ($N_{(13)}$). See Table 3 and supplementary material S1 for all models.
Comparing the four different types of genetic data sets we observed the largest improvement in ARB in analyses based upon only microsatellite genotypes (from 1.61 to 0.42) or a combination of microsatellite genotypes and mitochondrial DNA sequences (from 1.48 to 0.47). Despite much more data, the improvement in ARB was slightly less for the data sets comprised solely of SNP genotypes (from 1.41 to 0.77) or mitochondrial DNA sequences (from 1.33 to 0.80, Supplementary material Table S2.B). The improvement in RMAE was larger for data sets comprised solely of SNP genotypes (from 0.35 to 0.25) or a combination of microsatellite genotypes and mitochondrial DNA sequences (from 0.30 to 0.25) compared to data sets comprised solely of microsatellite genotypes (from 0.35 to 0.30) or mitochondrial DNA sequences (from 0.35 to 0.31, Supplementary material Table S2.B).

**Figure 6. Marginal posterior distribution of unscaled and scaled parameters.** Parameters were estimated from data generated under a demographic decline based upon data emulating mitochondrial DNA sequences. Unscaled parameters: (A) most recent effective population size $N_{o2}$, (B) historic effective population size $N_{1l2}$, (C) time of the population decline in generations $T_{T2}$, (D) mutation rate in substitutions per site per generation $\mu_{MIT}$. Scaled parameters: (E) the product of most recent effective population size by mean mutation rate $N_{o2}\mu_{MIT}$, (F) The product of historic effective population size by mean mutation rate $N_{1l2}\mu_{MIT}$, (G) the product of time by mean mutation rate $T_{T2}\mu_{MIT}$. The demographic model corresponds to a population decline ($N_{o2}=1,000, N_{1l2}=10,000, T_{T2}=750$) based on MIT (model DEC.E, Table 1). A dark grey line represents the prior distribution of the parameter values and a light grey line the posterior distribution of the parameter values. The y-axis is the density of the distributions.

**Assessing model fit**

In general, the goodness-of-fit was higher for all test statistics for the recovered model (Supplementary material Tables S3.A-H). In cases when an incorrect model was recovered as the most probable model, the goodness-of-fit of the recovered model was usually only marginally higher relative to the goodness-of-fit of the true model. Among 100 assessments (i.e., the goodness-of-fit
of all five candidate models for each of the 20 selected pseudo-empirical data sets), 22 candidate models contained at least one test statistic estimate that was located below 0.05 or above 0.95. None of these 22 candidate models were recovered as the most probably model (Supplementary material Tables S3.A-D). We obtained comparable outcomes for both employing the same or different statistics to estimate the posterior distribution of the parameter values and the goodness-of-fit of the model, suggesting that the choice of test statistics employed to assess the goodness-of-fit might not be all crucial (Supplementary material Tables S3.A-H). However, this specific outcome might not apply to all demographic histories or and data sets.

Discussion

We evaluated the performance of the ABC approach implemented in the popular software package DIYABC (Cornuet et al., 2014) both, in terms of ability to recover the true demographic model and with regards to the error and accuracy of the parameter estimates. Our assessment focused on a few simple, single-population demographic models with one or two changes in population size in the relative distant past (i.e., the LGM), as the performance of DIYABC would likely decrease for more complex models (e.g. Geman et al., 1992; Briscoe & Feldman, 2011). The graphical user interface (GUI), necessary for all analysis except the generation of simulated data sets, employed to estimate the summary statistics, is likely one of the main reasons why DIYABC is a popular choice. However, the mandatory use of GUI for the last steps in model selection and parameter estimation also constitutes a severe limitation for simulation-based assessments of DIYABC. This was why we were forced to limit the number of replicated pseudo-empirical data sets to ten for each combination of genetic data set and demographic model. However, the total number of pseudo-empirical data sets (520) in our assessment is considerably larger to those employed in previous assessments of DIYABC (Cornuet et al., 2008; Cornuet et al., 2010). The earlier assessments of DIYABC were based upon a single replicate (either 10 autosomal microsatellite genotypes, 1,000 nucleotides of mitochondrial DNA sequences, five nuclear autosomal DNA sequences of 1,000 nucleotides each, and all combinations of two and three types of markers) per demographic model (for a total of seven assessments). Here, we evaluated ten replicate data sets per combination of genetic data set and demographic model, resulting in a total of 520 data sets (Fig. 2). Each pseudo-empirical data set was evaluated against five and three candidate models, representing a total of 1,040 assessments. In many instances, the differences among the aspects compared were such that ten replicates were sufficient. However, many comparisons were effectively conducted across multiple parameter
combinations and hence the actual number of pseudo-empirical data sets employed in an assessment was much larger. Similar to most of such simulation-based assessments, the pseudo-empirical data was generated under assumptions that are identical, or very similar to the assumptions underlying the subsequent data analysis. In other words, the results presented here represent the best-case scenario. Consequently, it is safe to assume that the model selection and parameter value estimates obtained from empirical data sets likely will be subject to higher uncertainty and error compared to the assessment of the data analyzed in this study.

In general, we found that DIYABC performs relatively poorly. In less than half of the assessments, DIYABC was able to recover the correct demographic model, and most demographic parameters were estimated with bias and with high error. Our findings are based upon sample sizes and data sets that closely resemble data sets in publications employing DIYABC (e.g., Fontaine et al., 2012; Besnard et al., 2014; Inoue et al., 2014; Kimura et al., 2014; Louis et al., 2014a).

Our findings suggested that the ability to recover the correct demographic model, as well as the error and accuracy of the parameter estimates was influenced by multiple factors, such as the number of candidate models; which parameters are targeted; the informativeness of each molecular marker type, and the specific demographic history. Below we discuss some of these aspects in further detail and provide recommendations of how to remedy some of the potential caveats.

**Selection of the demographic models: complexity, similarity and number of candidate demographic models**

*Model complexity*

The genetic make-up of contemporary populations is a product of the interaction of multiple, different underlying processes, which potentially may lead to similar outcomes. Acknowledging this underlying complexity encourages the inclusion of multiple and complex candidate models in the assessment of the demographic history. However, it was evident from our results that multiple changes of $N_e$ was challenging to identify compared to demographic histories with only a single change of $N_e$. This general observation was consistent across all combinations of genetic data sets. Consequently, it is probably overly optimistic to expect a DIYABC-based analysis to capture every detail of the demographic history; a more efficient strategy might instead be to conduct an estimation that focuses on a few simple candidate demographic models that aim to capture that major difference(s) with the large effect size(s).

*Similarity of the candidate models*

Reassuringly our assessment revealed that in those cases when an incorrect model was recovered, the model deemed most probable was usually that most similar to the true model. For
example, in the case of data generated under the INC model, the models recovered most frequently (when the INC model was recovered), were those with increases in $N_e$ as well (i.e., DECINC and INCDEC). As the similarity of the candidate models increased so did the failure to recover the true model. These trends were also evident in increased ability to recover the true model when increasing the dissimilarity among the candidate models. Similar results were reported in other ABC assessments where model selection was employed to make phylogeographic inferences (Pelletier & Carstens, 2014). In this case, the authors concluded that increasing the similarity among the candidate models decreased the probability of recovering the true model and the expected posterior probability of true model reduced. Consequently, at least in the initial analysis, it is probably most efficient to avoid the inclusion of similar candidate models.

**Number of candidate models**

In general, our results suggested that increasing the number of candidate models influences the rate of recovering the true demographic model. As noted by Pelletier and Carstens (2014), an increase in the number of candidate models typically increases the overall similarity among the candidate models, which, in turn, reduces the expected posterior probability of the true model. The obvious consequence is to recommend a reduction in the number of candidate models to the bare minimum. Although our study showed that reducing the number of candidate demographic models improved the probability of recovering the correct demographic model, such a reduction in the number of candidate demographic models harbors the inherent risk of excluding the true model from the analysis altogether.

Previous authors have proposed to circumvent this potential caveat by conducting the model selection in a hierarchical manner, making an initial model selection among relatively simple models and subsequently add complexity to a second set of the models deemed most probable (Chan *et al.*, 2014; Turner & Van Zandt, 2014). Although we did not evaluate such a hierarchical approach, we found that limiting the number of candidate models to the least similar models appeared to increase the ability to recover the correct model, a finding that is consistent with the suggestion of conducting the model selection in a hierarchical manner. However, our simulations also revealed that models very similar to the true model are recovered sometimes, as the true model. Ideally, the posterior probabilities of similar candidate models that are close to the true model should yield similar posterior probabilities and hence result in an undefined outcome. However, this may not always be the case according to our results. Conducting the model selection in a hierarchical manner might also come with the risk of recovering a model, which is not, globally, the most probable model akin to some of the criticism directed towards nested clade analysis. Accordingly, we suggest that once
a model has been recovered using or not a hierarchical approach, the assessment should be repeated with a different range of candidate models, including the selected model.

**Estimates of $N_e$ and $T$**

In general, we observed that estimates of $T$ were more biased and subject to higher error than estimates of $N_e$. Biased and imprecise estimates of $T$ appear to be a general feature of genetic estimates of $T$ and hence not necessarily due the ABC approach (e.g., Girod *et al.*, 2011; Nikolic & Chevalet, 2014). Cornuet *et al.* (2008) suggested that the comparatively poor performance, in terms of estimating $T$, is caused by insufficient data. In other words increasing the sequence lengths and/or the number of loci might improve the accuracy and error of $T$ (e.g., Abdo *et al.*, 2004; Hoban *et al.*, 2013). However, we did not observe a substantial improvement in data sets comprised of 1,000 SNP genotypes or a combination of mitochondrial DNA sequences and microsatellite genotypes. Furthermore, while increasing the amount of genetic data might result in a higher precision, more data may not necessarily improve the accuracy (e.g., Peery *et al.*, 2013).

**Unscaled versus scaled parameter estimation**

Most studies aimed at elucidating the demographic history of natural populations are interested in obtaining estimates of the unscaled parameters, i.e., $N_e$ and $T$. However, the underlying coalescent-based simulations are based upon scaled parameter values (i.e., the product of $N_e$ and $\mu$ or $N_e$ and generation time) and converting the scaled estimates to their unscaled values comes at a cost. Our simulations did indeed confirm that unscaled parameters are estimated with a higher error and bias compared to scaled parameters. Additional uncertainty is introduced during the conversion of scaled parameter estimates to their unscaled components since their conversion require estimates of additional nuisance parameters, such as $\mu$ or generation time, which are also subject to uncertainty which is often ignored (Girod *et al.*, 2011). Bertorelle *et al.* (2010) suggested to employ scaled parameter estimates to the largest possible extent. Most aspects of interest to many studies may be informed from the relative differences in scaled parameter estimates (e.g., Fontaine *et al.*, 2012).

**Molecular marker types and number of loci**

Our findings suggested that the error of the parameter estimates and model selection was improved either with the increasing amounts of data or by including multiple genetic data types. For example, data sets comprised of both microsatellite genotypes and mitochondrial DNA sequences or increasing the data to many hundreds of SNP genotypes, as opposed to relying only on mitochondrial DNA sequences or genotypes from a few microsatellite loci. While it may be that
additional loci or longer DNA sequences will improve precision and bias, increasing the amount of
data did not always result in a corresponding improvement in accuracy or precision. For instance,
in the case of the DEC model (Fig. 5), employing genotypes from 1,000 SNP genotypes greatly
improved the error of $N_{e2}$, $N_{I12}$ and $T_{T2}$ in comparison to data sets comprised only of mitochondrial
DNA sequences or genotypes from 20 microsatellite loci. However, the accuracy of same estimates
did not improve in a similar manner. The mode of the posterior probability distributions of $N_{e2}$, $N_{I12}$
and $T_{T2}$ based upon the 1,000 SNP data sets was upward bias in comparison to the posterior
probability distributions obtained from data sets comprised solely of mitochondrial DNA sequences
or microsatellite genotypes.

This trade-off between error and bias was previously noted by Beaumont et al. (2002) who
showed that an increase in the tolerance in the ABC estimation reduces error since the sample size
from which the posterior distribution is estimated increases. However, the larger sample sizes may
increase the bias due to uncorrected departures from additivity and linearity (Beaumont et al., 2002).
Increasing data will also add variance and hence diminishing the return in terms of improving the
error and bias (Robinson et al., 2014; Shafer et al., 2015). For example, Shafer et al. (2015) reported
that 1,000 SNP genotypes performed equally well to 50,000 SNP genotypes in terms of parameter
estimates such as time. An additional, unaccounted source of error, which increases with DNA
sequence length, is recombination. Unfortunately, most implementations, including ABC-based, in
population genetics ignore recombination. Simulations, as in this study, can be employed as an
effective manner to identify the optimal balance between accuracy and error for each specific case.

During our simulations we noted some consistently incorrect outcomes in some assessments
based upon genetic data sets comprised of SNP loci (e.g., submodels D, F and K), which suggested
a potential computational bug in the simulations conducted by DIYABC.

The ability of all estimation approaches depends crucially upon the information content in
the data that serves as the basis of estimation and, in the case of ABC-based approaches, the
suitability of the specific summary statistics. Summary statistics differ in terms of the population
genetic aspects they capture, the nature of the genetic data (e.g., nucleotide sequence or
microsatellite loci) and the true population history. The primary aim of this study was not to evaluate
the choice of individual summary statistics. Instead, we employed a suite of summary statistics that
are common to past studies based on DIYABC. The very nature (and advantage in a ABC context)
of summary statistics is to reduce the overall information content, which in turn introduces
additional bias, particularly in complex models (Robert et al., 2011; Aeschbacher et al., 2012;
Sunnaker et al., 2013). Given this fundamental feature of ABC-based estimations may make it seem
appealing to increase the number of summary statistics, thereby presumably improving the precision
and bias of the parameter estimates. However, we observed bias across all four types of generic data sets, suggesting that the bias is not in order to improve the parameter estimates. However, we observed bias across all four kinds of genetic data sets, suggesting that the bias is not necessarily related to a particular set of summary statistics if one can assume that the same summary statistic results in different bias for different molecular data types. Additionally, the “curse of dimensionality” (Beaumont et al., 2002) also limits the benefits of increasing the number of summary statistics. Consequently, aiming for a few “optimal” summary statistics is likely a better strategy.

Conclusions and suggested guidelines

Genetic diversity in natural populations is a product of multiple interacting evolutionary processes, which makes it a challenging exercise to discern among those processes and to obtain estimates of unbiased parameters with low error. However, the development of specific parameterized hypotheses as demographic models will aid a study to focus on the key parameters.

Our simulated-based evaluation of DIYABC demonstrated a low ability to recover the true demographic model and that some parameters were estimated with considerable error and bias. However, an ABC-based approach, such as that implemented in DIYABC is an effective manner by which to obtain a general (as opposed to very specific) insight into the past of contemporary populations.

Our results suggested not to focus on capturing every detail and not to aim for an as complex and realistic model as possible. Instead, our findings suggest to focus on simple contrasting models that are likely to capture the key demographic events of large effect sizes. Our assessment also suggested that it might be counterproductive to include very similar candidate models in the estimation, which might increase the probability of recovering an incorrect model. Similarly, the number of candidate models in itself proved important in terms of recovering the true model. Accordingly, it is recommendable to reduce the number of candidate models (e.g., less than five candidate demographic models). Additional models could be included later in the assessment by conducting the model selection, for example, in a hierarchical manner. Finally, it is advisable to evaluate different combinations of candidate models in order to assure that the same candidate model is recovered consistently (akin to convergence in other Bayesian-based approaches).

The apparent difficulty in obtaining accurate and precise estimates of some parameter values, such as $T$, may be circumvented, to some extent, by focusing on scaled parameters or the confidence interval of the estimate rather than a point estimate. Although, most studies aim to estimate unscaled
parameters, scaled parameters were estimated with comparatively less bias and lower error. Accordingly, it might prove beneficial to employ scaled parameter estimates when possible.

The development of massively parallel sequencing technologies has enabled the generation of very large data sets in most species. Although, an increase in the amounts of genetic data seems a logical next step to improve the estimation, our simulations showed that “more is not always better”. Increasing the number of loci improved the error in the parameter estimates, but not necessarily the accuracy. The optimal number of samples and data is case-specific and depending upon aspects, such as, the objectives and the population history. Accordingly, applying generic rules may prove somewhat counter-productive. A reasonably and readily available approach is to employ a simulation framework, as in this study, to estimate the optimal amount and kind of data given the logistic and monetary constraints in each particular case (e.g., Hoban et al., 2012).

The results reported here provide a general insight into key aspects of ABC-based approaches in terms of model recovery rates and the bias and precision of the parameter estimates. Our assessment showed that inferences drawn from ABC model comparisons may require additional simulations in order to assess if the results are robust and likely to recover the true demographic history and parameter estimates. Shafer et al. (2015) recommended to apply both ABC and likelihood approaches. However, consistent results between two different analytical approaches, which fundamentally utilize the same aspects of the data, should not be considered as two independent tests. We considered that assessing the “sensitivity” of ABC-based (and likelihood) approaches in each specific case using simulations harbors substantial potential to gain insights into the specific estimation and identify which aspects are likely captured reliably. Unfortunately, the study of natural populations comes with limited prior knowledge and hence a considerable degree of uncertainty. Hence, it is most efficient to focus upon a few demographic events with large effects sizes, as it is almost impossible, at this time, to represent and resolve the enormous complexity of nature in a single model with contemporary genetic data.

ACKNOWLEDGMENTS

We thank the members of the MarECon group for their helpful comments and discussions throughout this work, in particular Dr. Jungkoo Kang, Dr. Anna Kopps, Tom Oosting, Xênia Lopes and Vania Rivera. We are also indebted to Lionel Morgado and Bob Dröge for their support with the execution of scrips and jobs on the University of Groningen High Performance Computing cluster Peregrine. The University of Groningen provided the funding for Andrea Cabrera’s PhD. We would also like to thank Brian Lavin, Hielko van der Hoorn, the five anonymous reviewers, and the editor, Dr. Sean Schoville, for their constructive criticisms, which greatly improve our manuscript.
Supplementary material

Additional supporting information may be found online in the supporting information tab for this article: Cabrera, A. A., & Palsbøll, P. J. (2017). Inferring past demographic changes from contemporary genetic data: A simulation-based evaluation of the ABC methods implemented in DIYABC. Molecular Ecology Resources, 17(6), e94-e110. doi: 10.1111/1755-0998.12696