DOI: 10.1534/genetics.109.102509

Efficient Approximate Bayesian Computation Coupled With Markov Chain Monte Carlo Without Likelihood

Daniel Wegmann,*,† Christoph Leuenberger‡ and Laurent Excoffier*,†,1

**Computational and Molecular Population Genetics Laboratory, Institute of Ecology and Evolution, University of Bern, 3012 Bern, Switzerland and* †*Swiss Institute of Bioinformatics,* ‡*Ecole d'Ingénieurs de Fribourg, 1705 Fribourg, Switzerland*

Manuscript received March 6, 2009
Accepted for publication May 31, 2009

ABSTRACT

Approximate Bayesian computation (ABC) techniques permit inferences in complex demographic models, but are computationally inefficient. A Markov chain Monte Carlo (MCMC) approach has been proposed (Marjoram *et al.* 2003), but it suffers from computational problems and poor mixing. We propose several methodological developments to overcome the shortcomings of this MCMC approach and hence realize substantial computational advances over standard ABC. The principal idea is to relax the tolerance within MCMC to permit good mixing, but retain a good approximation to the posterior by a combination of subsampling the output and regression adjustment. We also propose to use a partial least-squares (PLS) transformation to choose informative statistics. The accuracy of our approach is examined in the case of the divergence of two populations with and without migration. In that case, our ABC–MCMC approach needs considerably lower computation time to reach the same accuracy than conventional ABC. We then apply our method to a more complex case with the estimation of divergence times and migration rates between three African populations.

With the advent of large-scale genotyping techniques (*e.g.*, Green *et al.* 2006; Levy *et al.* 2007), genetic data can be produced at an unprecedented scale (Altshuler *et al.* 2005; Bustamante *et al.* 2005), and the genetic variability of individuals and populations can now routinely be examined at hundreds of loci across the genome (Rosenberg *et al.* 2002; Williamson *et al.* 2005; Becquet and Przeworski 2007; Frazer *et al.* 2007). These large data sets offer the hope to better understand the evolutionary forces that have shaped the diversity of many species, including humans, and to identify genome regions involved in past selective events (Anisimova and Liberles 2007; Nielsen *et al.* 2007). However, the demographic history of the populations needs to be accounted for to disentangle its effects from those of selection (Haddrill *et al.* 2005; Nielsen *et al.* 2005; Biswas and Akey 2006). It therefore seems important to be able to properly estimate this past demography from neutral genetic data or to estimate demography and selection simultaneously (*e.g.*, Williamson *et al.* 2005). The statistical estimation of mutation and demographic parameters has drastically improved in the last 10 years, particularly with the use of Bayesian and full-likelihood approaches (Beaumont *et al.* 2002; Marjoram and Tavare 2006). However, these methods are still restricted to relatively simple models whose likelihood can be computed or to small data sets that can be analyzed in a reasonable amount of time. The handling of large data sets and the estimation of demographic parameters under realistic models remain problematic, and goodness-of-fit methods have been often used in those cases (see, *e.g.*, Marth *et al.* 2004; Schaffner *et al.* 2005; Plagnol and Wall 2006).

The approximate Bayesian computation (ABC) framework (Tavare *et al.* 1997; Pritchard *et al.* 1999; Beaumont *et al.* 2002), which is based on a simple rejection algorithm, has been applied to the estimation of demographic parameters in a variety of evolutionary models in nonmodel organisms and in humans (Estoup *et al.* 2004; Tallmon *et al.* 2004; Excoffier *et al.* 2005a; Hickerson *et al.* 2006; Fagundes *et al.* 2007; Rosenblum *et al.* 2007). The generic principle first outlined in Tavare *et al.* (1997) is to simulate data (D') similar to observations (D) for sample size and number of loci under a given model, with parameters ($\boldsymbol{\theta}$) drawn from some prior distributions. If D' is identical to D, the parameters are stored, and discarded otherwise, and the retained parameters are used to estimate the posterior distribution. Since it is very unlikely to simulate D' identical to D for large data sets or complex models (Marjoram and Tavare 2006), it has been proposed (Pritchard *et al.* 1999) to replace data by a set of summary statistics $\mathbf{S}$ and to retain a

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.109.102509/DC1.

[1]*Corresponding author:* Computational and Molecular Population Genetics Laboratory, Institute of Ecology and Evolution, University of Bern, Baltzerstrasse 6, CH-3012 Bern, Switzerland.
E-mail: laurent.excoffier@iee.unibe.ch

particular simulation if the simulated summary statistics $\mathbf{S}'$ are sufficiently close to the observed summary statistics $\mathbf{S}$. To account for the difference between $\mathbf{S}$ and $\mathbf{S}'$, Beaumont *et al.* (2002) recently proposed to perform a locally weighted linear regression to compute the posterior distribution. This regression adjustment step was shown to lead to much improved estimations. However, even with this improvement the ABC methodology is computationally not very efficient, as it requires the simulation of millions of samples, a large majority of which, typically ≥99%, will be discarded for parameter estimation.

More recently, Marjoram *et al.* (2003) proposed another likelihood-free approach where simulations are directly embedded within a Markov chain Monte Carlo (MCMC) framework. They showed that a Markov chain where newly simulated data D' would be accepted if they were equal to the observed data D, and rejected otherwise, would converge to the right posterior distribution $\Pr(\boldsymbol{\theta} \mid D)$. For complex data sets where the acceptance rate is too small, they proposed to replace the full data by summary statistics and to accept new parameter values if sufficiently close to the data, like in conventional ABC. A problem linked with this approach is to define how close simulations need to be to accept them, which conditions (i) the acceptance rate and (ii) the mixing and the convergence of the chain, but also (iii) the burn-in period, since it may require a very large number of simulations to have the first accepted step if the starting point is in a region with low likelihood. Other likelihood-free approaches, like sequential Monte Carlo (Sisson *et al.* 2007) or population Monte Carlo (M. A. Beaumont, J.-M. Cornuet, J.-M. Marin and C. P. Robert, unpublished results), have been recently proposed to address the problem of slow convergence and to more efficiently explore multimodal posteriors. However, these likelihood-free approaches have only rarely been tested for complex evolutionary models (but see Ratmann *et al.* 2007 for a comparison of evolutionary dynamics of protein networks).

In this article, we present a new approach, borrowing the best features of both conventional ABC based on a rejection algorithm and MCMC without likelihood, which deals efficiently with all the problems mentioned above. We test our approach in the case of different models of population isolation and migration introduced by Nielsen and Wakeley (2001), but do not attempt to compare our methodology with full-likelihood approaches. Our new methodology is finally applied to a model of population isolation and migration for three African populations.

METHODS

We begin by describing Marjoram *et al.*'s (2003) likelihood-free MCMC algorithm based on summary statistics (SS), hereafter called SS–MCMC, showing its underlying problems. We then propose solutions borrowed from conventional ABC, which lead to a new algorithm called ABC–MCMC.

SS–MCMC algorithm: Given some observed data D generated under a given model M defined by a set of (unknown) parameters $\boldsymbol{\theta}$ with prior distribution $\pi(\boldsymbol{\theta})$, Marjoram *et al.* (2003) have shown that the posterior distribution $f(\boldsymbol{\theta} \mid D)$ could be obtained from samples of a Markov (M) chain without likelihood generated by the following algorithm:

M1. Propose a move from current state $\boldsymbol{\theta}$ to a new state $\boldsymbol{\theta}'$ according to a transition kernel $q(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}')$.
M2. Simulate data D' under model M, using the new parameters $\boldsymbol{\theta}'$.
M3. If $D' = D$, go to step M4; otherwise remain at state $\boldsymbol{\theta}$ and go back to step M1.
M4. Accept state $\boldsymbol{\theta}'$ with probability $h(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}') = \min(1, \pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}' \rightarrow \boldsymbol{\theta})/\pi(\boldsymbol{\theta})q(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}'))$; otherwise remain at state $\boldsymbol{\theta}$. Go to step M1.

Since it is very unlikely to simulate D' identical to D for large data sets or complex models (Marjoram and Tavare 2006), Marjoram *et al.* (2003) proposed to replace the data D by a set of sufficient summary statistics $\mathbf{S}$ and the condition $D' = D$ by the less stringent condition $\delta = \|\mathbf{S}' - \mathbf{S}\| \leq \delta_\varepsilon$ in step M3, where δ_ε is an arbitrarily small distance between $\mathbf{S}'$ and $\mathbf{S}$. Note that if the summary statistics are not sufficient, that is, the statistics do not capture the full information contained in the data for the parameters $\boldsymbol{\theta}$, the resulting posterior will be only an approximation of the true posterior distribution (Marjoram and Tavare 2006). With this new step M3, Marjoram *et al.* (2003) claimed that the stationary distribution of this SS–MCMC chain is $f(\boldsymbol{\theta} \mid \delta \leq \delta_\varepsilon)$, which should be a good approximation of the true posterior $f(\boldsymbol{\theta} \mid D)$ if δ_ε is small (Marjoram *et al.* 2003).

Problems with the SS–MCMC algorithm: The use of summary statistics and of a threshold value δ_ε in the modified step M3 introduces some issues that we address below.

1. Since the Markov chain can begin only if $\delta \leq \delta_\varepsilon$, the number of steps necessary to first satisfy this condition is undefined and it depends on the (unknown) distribution of δ, $\pi(\delta \mid \mathbf{S}, \boldsymbol{\theta})$ in the following.
2. The choice of the threshold value δ_ε is important as a too large tolerance interval results in a chain being dominated by the prior $\pi(\boldsymbol{\theta})$, as unlikely $\boldsymbol{\theta}$ will be often accepted. On the other hand, a too small value leads to a very small acceptance rate. Additionally, the acceptance rate of the chain is proportional to the likelihood of $\boldsymbol{\theta}$ (Sisson *et al.* 2007), making it very sticky in regions of low likelihood, preventing the use of a too small threshold value δ_ε.
3. Finally, the accuracy of the estimation depends on the choice of the summary statistics and of the

distance function $\|\cdot\|$ (Hamilton *et al.* 2005). In complex models where the likelihood of the model cannot be computed, it is difficult to find sufficient statistics for all parameters, and therefore the definition of an optimal set of summary statistics is an important and still unresolved issue.

Modifications to the SS–MCMC algorithm: *Calibration step:* We propose here to address the first problem by performing a series of n simulations (we typically used $n = 10{,}000$), where parameters are each time randomly drawn from their prior to obtain $p_n(\delta \mid \mathbf{S}, \boldsymbol{\theta})$, an empirical approximation of $\pi(\delta \mid \mathbf{S}, \boldsymbol{\theta})$. With this calibration step, we can conveniently define a tolerance level ε and a threshold distance δ_ε such that $P(\delta \le \delta_\varepsilon) = \varepsilon$. For instance, by setting $\varepsilon = 0.01$, we define a value $\delta_{0.01}$ below which the condition $\|\mathbf{S}' - \mathbf{S}\| \le \delta_\varepsilon$ will be true for 1% of all randomly simulated data sets. We should then be able to use any simulation for which $\delta \le \delta_\varepsilon$ as a starting point for the chain. These n simulations are also used to adjust the proposal range and therefore the transition kernel q. The proposal range for new parameter values in the Markov chain is set as a uniform range of width φ expressed in units of standard deviations, computed independently for each parameter within the $n\varepsilon$ retained simulations. If the proposal range is too large, the chain may often jump to a state for which $\delta > \delta_\varepsilon$, which results in a low acceptance rate and consequently a smaller number of effective samples. If the proposal range is too small, the chain may never explore the whole parameter space and therefore may lead to a large variance in estimated parameters among replicates. Note that standard MCMC techniques aiming at increasing mixing, such as simulated tempering (Bortot *et al.* 2007), can easily be included in the framework presented here.

Combining SS–MCMC with ABC: We propose to address the second issue (small acceptance rate) by launching an SS–MCMC chain of length s with a relatively large tolerance (*i.e.*, $\varepsilon = 0.01$). As shown in the appendix, the stationary distribution $f(\boldsymbol{\theta} \mid \delta \le \delta_\varepsilon)$ of this SS–MCMC chain is identical to that obtained by a simple rejection algorithm (as in conventional ABC), using the same tolerance level ε. We therefore propose to estimate parameters on a subsample of size t consisting of the simulations associated with the smallest distances δ generated by the Markov chain, as is commonly done in the ABC framework (Pritchard *et al.* 1999; Beaumont *et al.* 2002), and to perform a local regression adjustment where samples are weighted by their associated δ_i-values (Beaumont *et al.* 2002), which greatly improves the quality of the estimation. This approach allows the chain to have a large acceptance rate, while making the final estimation not too sensitive to the prior due to the regression adjustment step (Beaumont *et al.* 2002). This two-step approach should therefore lead to the same stationary distribution as a simple rejection algorithm having tolerance $\varepsilon' = \varepsilon(t/s)$. The main gain of SS–MCMC compared to simpler rejection samplers is thus to require many fewer simulations to get t samples at tolerance ε': $s + n$ simulations under SS–MCMC compared to s/ε simulations under ABC, implying a theoretical reduction in computing cost by a factor $s[\varepsilon(s+n)]^{-1}$. For instance, with a calibration step of $n = 10{,}000$, a chain of $s = 90{,}000$ steps, and a tolerance of $\varepsilon = 1\%$, the expected gain is a 90-fold computational cost reduction. Therefore, we would expect that a total of 100,000 simulations under the improved SS–MCMC framework would correspond to 9 million simulations under the conventional ABC framework, but there are practical limits to this gain that are inherent to the difficulty in running chains for small ε-values (see below). In the following, we use the term ABC–MCMC to designate the procedure where an ABC regression adjustment is performed on a given fraction of the output of an SS–MCMC chain.

Partial least-squares transformation of summary statistics: We propose to address the problem of the choice of informative summary statistics by using a partial least-squares (PLS) regression approach (see, *e.g.*, Tenehaus *et al.* 1995; Boulesteix and Strimmer 2007). Like principal component analysis (PCA), PLS extracts orthogonal components from a high-dimensional data set $\mathbf{X}$ of predictor variables, but in addition, these components are chosen to appropriately explain the variability of the response variables by maximizing the covariance matrix of predictor and response variables (see, *e.g.*, Boulesteix and Strimmer 2007). In the present ABC context, the predictor variables are raw summary statistics and the response variables are model parameters. The choice of the number of PLS components to include is usually based on a leave-one-out validation procedure (Mevik and Wehrens 2007), examining the root mean square error (RMSE) of the parameters predicted by the regression. As a result of a PLS analysis, one should thus get a much reduced number of independent components, as compared to a large initial set of potentially correlated summary statistics, some of them being little correlated with any parameters and thus only adding noise to the Euclidean distance. Another advantage of the PLS transformation is that it guarantees that the matrix of PLS-transformed summary statistics is nonsingular, which is required when performing the final locally weighted linear regression for estimating parameters (Beaumont *et al.* 2002). In practice, we propose to compute a relatively large set of summary statistics (assumed to be informative about the parameters) on each simulated and observed data set. The n random simulations done in the calibration step are used to compute the PLS components, which are then used to transform the summary statistics computed on the simulated data sets

generated during the Markov chain. We used the routine "plsr" from the freely available R package "PLS" to compute the PLS components and to select an optimal set of k components according to a leave-one-out validation procedure (Mevik and Wehrens 2007). Since the PLS transformation assumes a linear relationship between parameters and statistics, we first applied a multivariate Box–Cox transformation (Box and Cox 1964) on each statistic separately before defining PLS components. Note that other ways to choose appropriate summary statistics could be imagined, like scoring them according to whether their inclusion substantially improves the quality of the inference, as recently proposed (Joyce and Marjoram 2008).

ABC–MCMC algorithm: We describe here the ABC–MCMC algorithm (AM), incorporating the proposed improvements compared to the plain MCMC approach without likelihood:

AM1. Perform n simulations with parameters $\boldsymbol{\theta}'$ randomly drawn from their priors, and each time compute their associated set of summary statistics $\mathbf{S}'$.
AM2. Compute PLS components from the n $\boldsymbol{\theta}'$ and $\mathbf{S}'$ vectors after a Box–Cox transformation of the statistics.
AM3. For all n simulations, transform the summary statistics $\mathbf{S}'$ into k retained PLS components, as $\mathbf{S}'_{\text{PLS}}$. Transform the observed summary statistics $\mathbf{S}$ as $\mathbf{S}_{\text{PLS}}$ and compute $p_n(\delta \mid \mathbf{S}_{\text{PLS}}, \boldsymbol{\theta})$.
AM4. Fix ε, estimate δ_ε from $p_n(\delta \mid \mathbf{S}_{\text{PLS}}, \boldsymbol{\theta})$, and set the proposal range of the parameters for the transition kernel $q(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}')$ on the basis of φ and the variability of the parameters among the $n\varepsilon$ retained simulations.
AM5. Start an MCMC chain of total length s from a position $\boldsymbol{\theta}$ randomly chosen from the $n\varepsilon$ simulations closest to D. Set $i = 0$.
AM6. If now at $\boldsymbol{\theta}$, propose a move to $\boldsymbol{\theta}'$ according to a transition kernel $q(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}')$. Increment i.
AM7. Simulate D' on the basis of $\boldsymbol{\theta}'$. Compute the summary statistics $\mathbf{S}'$ and transform them into $\mathbf{S}'_{\text{PLS}}$.
AM8. If $\delta_i = \|\mathbf{S}'_{\text{PLS}} - \mathbf{S}_{\text{PLS}}\| \geq \delta_\varepsilon$, stay at $\boldsymbol{\theta}$ and go to AM6.
AM9. Accept $\boldsymbol{\theta}'$ with probability $\min(1, \pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}' \rightarrow \boldsymbol{\theta})/\pi(\boldsymbol{\theta})q(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}'))$; otherwise stay at $\boldsymbol{\theta}$.
AM10. If $i < s$, go to AM6.
AM11. From the s samples of the chain, retain a subsample of size t consisting of simulations with smallest associated δ_i-values and discard the other s-t samples.
AM12. Perform an ABC regression adjustment (Beaumont *et al.* 2002) on the t retained samples to estimate parameters.

To prevent the chain from remaining stuck at a given starting position after step AM5, we choose a new initial $\boldsymbol{\theta}$-value if the chain does not move to a new value $\boldsymbol{\theta}'$ after 20 proposals. Note also that we update all parameters at the same time in step AM6. As mentioned above, the width of the proposal range is adjusted independently for each parameter, and it is set to a fraction φ of the standard deviation of the parameters computed from the $n\varepsilon$ simulations retained in step AM4. Following Beaumont *et al.* (2002), we used the Euclidean norm as a distance function $\|\cdot\|$, but without standardization of $\mathbf{S}_{\text{PLS}}$.

Parallelizing ABC–MCMC: Simulations performed under the conventional ABC approach can be easily distributed among many CPUs. We thus implemented a parallelized version of the SS–MCMC algorithm as follows: after an initial calibration phase of 10,000 random simulations (which can easily be parallelized), we run 10 independent chains of 9000 simulations (including startup) with identical ε- and φ-values. The 10 chains are then concatenated and used to estimate posterior densities on the 5000 "best" simulations, using a local regression adjustment (Beaumont *et al.* 2002) to complete the ABC–MCMC algorithm. With this approach, 90% of the 100,000 simulations can thus be distributed on 10 different CPUs, and larger numbers of simulations could be distributed on more CPUs.

Illustration and application: We have tested and applied the ABC–MCMC approach to a model of population divergence with migration (Nielsen and Wakeley 2001). In this model, one assumes that some T generations in the past, an ancestral population of size N_A splits into two populations of size N_1 and N_2 and that migrants are then exchanged between the two populations at rates (looking forward in time) m_{12} and m_{21}. This methodology has been applied to several different data sets (*e.g.*, Becquet and Przeworski 2007; Hoelzel *et al.* 2007). It must be noted that it is one of the most complex population genetic models for which a full-likelihood implementation is available (Hey and Nielsen 2007; Kuhner 2009). However, we shall not attempt here to compare our methodology with the full-likelihood approach, since the purpose of this article is to improve over conventional ABC and not to compete with full-likelihood methods.

We first compared the conventional ABC and the new ABC–MCMC approaches in a simple case without migration ($m_{12} = m_{21} = 0$). All parameters were drawn from uniform priors: number of gene copies per population $N \sim U[0, 30{,}000]$ and divergence time $T \sim U[0, 16{,}000]$ generations. We simulated genetic data with the program SIMCOAL 2.0 (Laval and Excoffier 2004) for 25 diploid individuals per population genotyped at 50 unlinked microsatellites each. Microsatellite data were simulated under a pure stepwise mutation model. While we used a fixed mutation rate $\mu = 5 \times 10^{-4}$ in this case, the mutation rate was allowed to vary in the applied case, as explained below. A total of 100 test data sets with parameter values drawn randomly from the prior distributions were used to compare ABC and ABC–MCMC. We then compared the conventional ABC and the new ABC–MCMC approaches in a more complex case with migration rates chosen in $U[0, 0.003]$. Again,

100 test data sets were randomly simulated by drawing parameter values from prior distributions.

Summary statistics: Using the software package Arlequin 3.1 (Excoffier *et al.* 2005b), we computed in each population the average and standard deviation (over loci) of the number of alleles (K), the range of the allele size (R), the expected heterozygosity (H), the Garza–Williamson statistic (Garza and Williamson 2001) modified as $GW = K/(R + 1)$ (Excoffier *et al.* 2005a), and another modification of GW computed as $GW^* = K/(R_{Tot} + 1)$, where R_{Tot} is the range in allele size computed over all sampled populations. The idea behind the use of the GW* statistic is that it should reflect population-specific drift effects, since the denominator is the same in all populations. The same statistics and their standard deviation over populations were also computed over the two pooled populations, except GW* since GW = GW* in that case. We additionally computed the differentiation index F_{ST} and the genetic distance $(\delta\mu)^2$ (Goldstein *et al.* 1995) between the two populations. We thus computed a total of 31 summary statistics over all observed and simulated data sets. PLS components were extracted from summary statistics on the basis of the 10,000 simulations performed in the calibration step. We used the R package PLS (Mevik and Wehrens 2007) to find the appropriate number of PLS components to use (10 for both cases with and without migration).

Measuring the accuracy of the methods: Since simulations are computationally much more demanding than the estimation of the posterior distributions obtained by regressing summary statistics on parameters (Excoffier *et al.* 2005a), we compared ABC to ABC–MCMC samplers in terms of their accuracy. We measured the root mean integrated squared error (RMISE) of the whole posterior distribution, defined as $\text{RMISE} = \sqrt{\int(\theta_k - \mu_k)^2 f(\theta_k \mid s)\delta\theta_k}$, where μ_k is the true value of the kth parameter and $f(\theta_k \mid s)$ is the estimated posterior density. We also computed the error of three point estimates (mode, mean, and median) as the absolute difference between the point estimate and the true parameter value. To assess the overall quality of the estimation, we measured the geometric mean of these accuracy measures over all parameters, relative to a conventional ABC approach done with 100,000 simulations. More formally, if $x_{i_{ABC}}$ is the measure of accuracy for the ith parameter ($i = 1, \ldots, n$), then the relative accuracy RA of the ABC–MCMC approach is defined as $\text{RA} = \left(\sum_{i=1}^{n} x_{i_{\text{ABC-MCMC}}} / x_{i_{\text{ABC}}}\right)^{1/n}$.

The coverage property of the posterior distributions obtained by different methods is also worth checking. By coverage, we mean the proportion of times a true parameter value is present in a given credible interval. For instance, 80 and 95% credible intervals should include the true parameter with probabilities 0.8 and 0.95, respectively. In other words, the posterior quantiles of the true parameter values should be uniformly distributed in [0, 1] (Cook *et al.* 2006). To check the overall coverage property of the posterior distributions obtained with different approaches, we therefore computed the posterior quantiles of the parameters for 100 test data sets for which true parameter values were known. Their uniformity was then assessed with a classical Kolmogorov–Smirnov test.

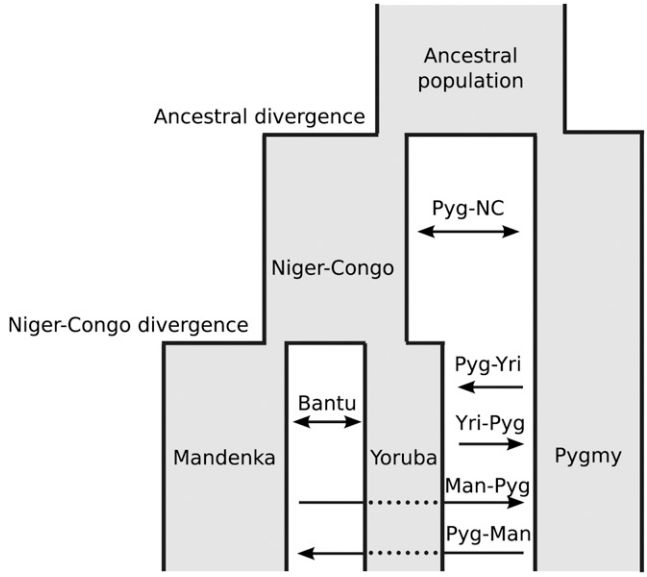


Figure 1.—Assumed evolutionary model describing the relationship between three African populations: Yoruba, Mandenka, and Mbuti Pygmies. See text for a description of the parameters and their priors.

Application of ABC–MCMC to human evolution: We studied the genetic relationships between three human African populations (the Yoruba and the Mandenka population belonging to the recently diverged Niger–Congo linguistic group and the Mbuti Pygmy population, hereafter simply called Pygmies) analyzed for ~800 microsatellite markers (Ramachandran *et al.* 2005). We assumed an evolutionary model where the Mandenka and the Yoruba diverged recently from an ancestral Niger–Congo population, itself having diverged earlier from the Pygmy population (see Figure 1). All current and ancestral population pairs were assumed to exchange migrants, but at different rates. We assumed symmetric migration between the two Niger–Congo populations and between the Pygmies and the ancestral Niger–Congo population, but migration rates between Pygmies and the two Niger–Congo populations were allowed to be asymmetrical. To ensure a good fit with the simulated stepwise mutation model, we considered only a subset of 331 tetramicrosatellites selected to have <5% missing data or imperfect repeat alleles coded as missing data.

The model parameters were drawn from the following uniform distributions: number of gene copies per population, $N \sim U[0, 20{,}000]$; number of migrant genes per generation, $Nm \sim 10^{U[0,2]}$; divergence time of the two Niger–Congo populations, $\text{TDIV}_{\text{NG}} \sim U[0, 1000]$ generations; and ancestral divergence time, $\text{TDIV}_{\text{A}} \sim U[0,$

TABLE 1

Relative accuracy of conventional ABC and ABC–MCMC for a model of population divergence without migration

Approach	φ	ε	RRMISE	RE mode	RE mean	RE median	Acceptance rate[a]	Coverage *P*-value[b]
ABC–5M[c]	—	—	0.86 (0.1)	0.90 (0.56)	0.90 (0.36)	0.91 (0.34)	—	0.469
ABC–MCMC	5	0.1	0.88 (0.1)	0.98 (0.83)	0.91 (0.4)	0.92 (0.4)	0.30 (0.12)	0.606
	5	0.01	0.85 (0.14)	0.95 (0.93)	0.89 (0.44)	0.94 (0.47)	0.11 (0.08)	0.773
	5	0.001	0.79 (0.15)	0.94 (1.33)	0.99 (0.58)	0.98 (1.2)	0.03 (0.03)	0.286
	2	0.1	0.89 (0.09)	0.94 (0.9)	0.89 (0.31)	0.92 (0.55)	0.52 (0.12)	0.418
	2	0.01	0.84 (0.13)	0.95 (0.65)	0.9 (0.33)	0.92 (0.56)	0.24 (0.11)	0.536
	2	0.001	0.78 (0.16)	0.98 (2.47)	0.85 (0.45)	0.8 (0.82)	0.08 (0.07)	0.685
	1	0.1	0.88 (0.09)	0.9 (0.78)	0.91 (0.35)	0.93 (0.43)	0.60 (0.12)	0.484
	1	0.01	0.83 (0.12)	0.9 (0.72)	0.89 (0.35)	0.87 (0.38)	0.29 (0.13)	0.228
	1	0.001	0.79 (0.14)	0.91 (1.03)	0.88 (0.41)	0.88 (0.6)	0.11 (0.08)	0.756
	0.5	0.1	0.89 (0.08)	0.88 (0.72)	0.96 (0.38)	0.92 (0.32)	0.63 (0.12)	0.656
	0.5	0.01	0.84 (0.12)	0.86 (0.53)	0.91 (0.46)	0.92 (0.42)	0.30 (0.14)	0.567
	0.5	0.001	0.81 (0.13)	0.91 (0.83)	0.89 (0.55)	0.91 (0.57)	0.11 (0.09)	0.539
	0.1	0.1	0.90 (0.11)	0.92 (0.61)	0.96 (0.4)	0.97 (0.44)	0.66 (0.18)	0.917
	0.1	0.01	0.82 (0.11)	1.03 (0.81)	0.91 (0.44)	0.86 (0.41)	0.33 (0.17)	0.59
	0.1	0.001	0.76 (0.15)	1.08 (0.94)	0.94 (0.63)	1.05 (0.67)	0.13 (0.10)	0.013
	0.01	0.1	0.92 (0.26)	1.25 (2.2)	1.37 (1.1)	1.3 (1.08)	0.65 (0.26)	<0.001
	0.01	0.01	0.87 (0.2)	1.14 (1.13)	1.04 (0.71)	1.02 (0.88)	0.35 (0.20)	0.007
	0.01	0.001	0.79 (0.18)	1.37 (1.88)	1.06 (0.9)	1.21 (1.03)	0.14 (0.11)	<0.001
	0.001	0.1	1.02 (0.27)	1.52 (2.22)	1.59 (1.48)	1.73 (2.04)	0.67 (0.27)	<0.001
	0.001	0.01	0.9 (0.19)	1.33 (2.22)	0.92 (0.82)	1.09 (1.4)	0.37 (0.21)	0.018
	0.001	0.001	0.81 (0.22)	1.56 (1.15)	1.26 (1)	1.34 (0.9)	0.15 (0.11)	<0.001
	0.0001	0.1	1.02 (0.31)	1.78 (3.94)	1.46 (1.65)	1.63 (2.64)	0.69 (0.26)	<0.001
	0.0001	0.01	0.98 (0.25)	1.65 (2.89)	1.21 (0.7)	1.46 (1.46)	0.36 (0.21)	0.005
	0.0001	0.001	0.88 (0.18)	1.79 (1.72)	1.19 (0.67)	1.24 (0.85)	0.15 (0.12)	0.001

Mean and standard deviation (within parentheses) of accuracy measures were computed over 100 independent estimations, relative to those performed under a conventional ABC approach based on 100,000 simulations (ABC–100K). We report ABC–MCMC accuracy for different tolerance values ε and different proposal ranges φ. In all ABC cases, parameter estimation was done after a locally weighted regression adjustment on the 5000 simulations closest to observations (Beaumont *et al.* 2002).

[a] Fraction of ABC–MCMC steps for which new parameter values were accepted.

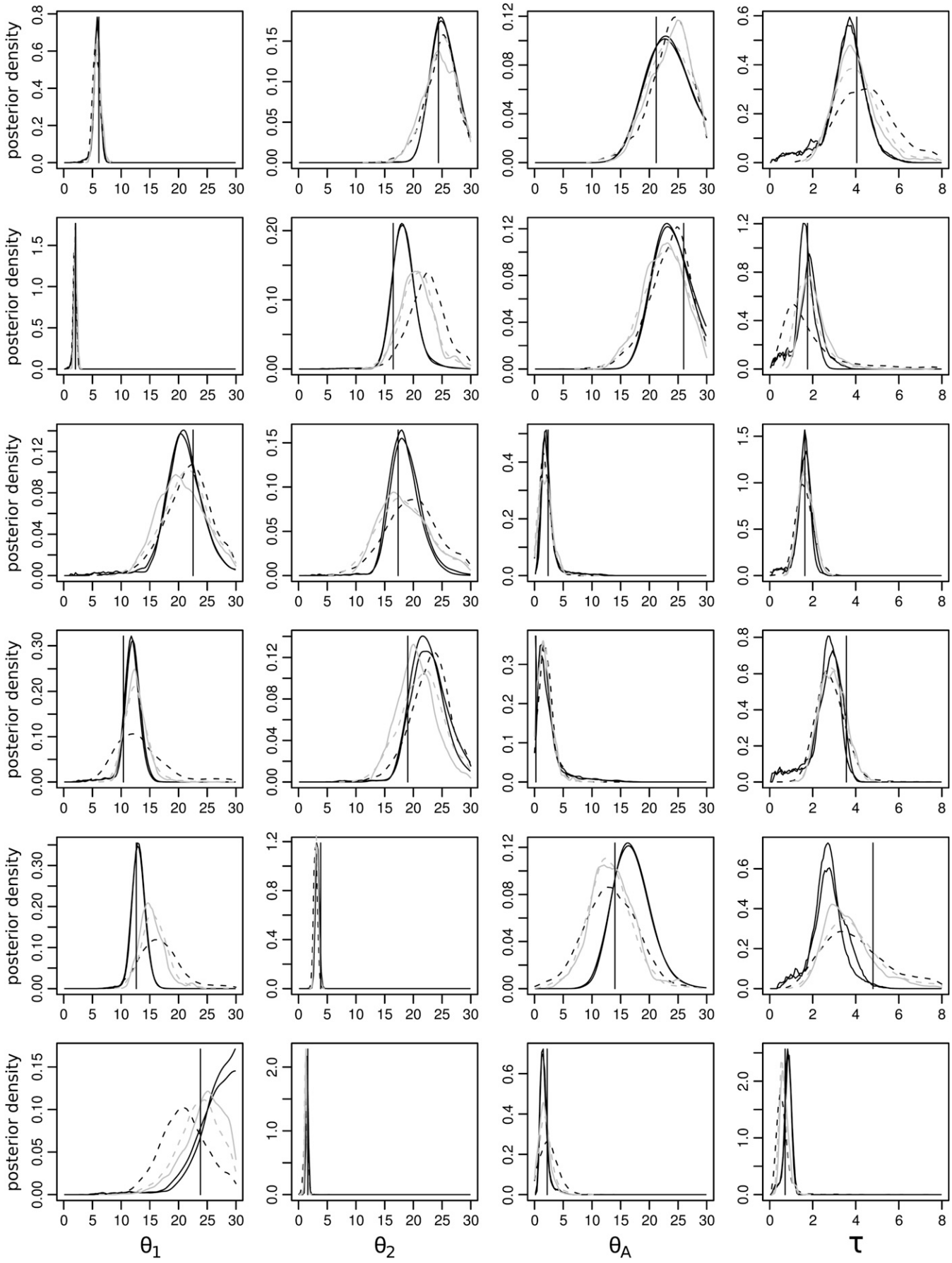
[b] For each data set, we estimated the posterior quantile of the true value of each parameter from its posterior distribution. We then tested the uniformity of the quantile values by a Kolmogorov–Smirnov test, which gives an indication of the quality of the coverage of the posterior distributions.

[c] ABC estimation performed on the basis of 5 million simulations.

4000] generations. We have imposed an additional constraint on the divergence times, such that $TDIV_A > TDIV_{NG}$, resulting in a nonuniform prior on $TDIV_A$. All simulations were performed with the program SIMCOAL 2.0, assuming a pure stepwise mutation model with a per generation average mutation rate $\overline{\mu}$ drawn in $U[2 \times 10^{-4}, 7 \times 10^{-4}]$. We implemented locus-specific mutation rates distributed as a Gamma $(\alpha, \alpha/\overline{\mu})$ with the shape parameter α drawn in $U[8, 16]$. The same set of statistics as that defined above was computed and PLS transformed. We retained the first 11 PLS components for the MCMC chain and parameter estimations.

RESULTS AND DISCUSSION

Performance of various ABC approaches: We have studied the performance of different samplers relative to the ABC approach for a model of population divergence without migration (Table 1). As expected, conventional ABC estimations based on 5 million simulations (ABC–5M) do improve on those done on only 100,000 simulations (ABC–100K), and their inferred posteriors have a very good coverage (Kolmogorov–Smirnov test of uniformity of posterior quantile values, $p = 0.469$). As expected, the performance of the ABC–MCMC approach depends strongly on the choice of the proposal range (φ) and tolerance (ε) values used to run the MCMC chains. Overall, the accuracy of ABC–MCMC increases with lower tolerance ε for all φ-values. It is found to be better than that of ABC–100K, except for very small proposal ranges ($\varphi \leq 0.01$). In that case, the relative errors of the parameters can be very large, and posterior distributions do not pass the test of posterior quantiles uniformity, suggesting that the parameter space is not adequately explored in those cases. Overall, ABC–MCMC performs best for a proposal range of $\varphi = 1$ and tolerance $\varepsilon = 0.01$, where accuracy measures are found to be even better than those obtained with ABC–5M. In that case, the acceptance rate of the chain is ~30% and the coverage of the


posterior density
θ1
θ2
θA
τ

posteriors is adequate ($p = 0.23$). We note that the acceptance rate of the ABC–MCMC sampler drops sharply with tolerance level ε. For instance, with $\varepsilon = 0.0001$ we were able to run ABC–MCMC chains only for a very few pseudo-observed data sets (six at most with $\varphi = 0.001$). Additionally, the acceptance rate is also strongly influenced by the proposal range φ, being two to three times higher with $\varphi = 0.001$ than with $\varphi = 5$. As also shown in Figure 2 for six random datasets, the ABC–MCMC approach generally leads to posterior distributions close to those obtained with ABC–5M. Posteriors obtained with ABC–100K are in most cases wider, in agreement with the accuracy measures reported in Table 1.

We then compared the different ABC samplers under a model of population divergence with migration. The results presented in supporting information, Table S1 are qualitatively very similar to the case without migration. The values of $\varepsilon = 0.01$ and $\varphi = 1$ again lead to the best overall accuracy for the ABC–MCMC approach, which shows correct coverage properties ($p = 0.29$). The accuracy of ABC–MCMC is again comparable to that obtained with the ABC–5M approach, confirming the strong reduction in computation time provided by this approach compared to conventional ABC.

Usefulness of PLS components: In Figure 3, we compare the distribution of the posterior quantiles of the true parameters for different models and different sets of statistics. While the distributions of the posterior quantiles obtained under ABC–5M based on PLS components are uniformly distributed (Kolmogorov–Smirnov test, $p = 0.469$ and $p = 0.193$ for divergence models with and without migration, respectively), this is not the case when raw statistics are used. The posterior quantiles obtained by considering all statistics in ABC–5M are indeed not uniformly distributed and tend to be too large ($p = 0.002$ and $p < 10^{-12}$ for divergence models with and without migration, respectively), which implies that the true parameters are globally underestimated when using all statistics. These results show that the use of PLS generally improves the coverage properties of the credible intervals. This was expected since many summary statistics may carry only very little information about the parameters, which makes it very difficult to calculate meaningful distances. Of course, a small set of carefully chosen summary statistics based on their theoretical relation with parameters or on some scoring procedure (*e.g.*, Joyce and Marjoram 2008) may equally well lead to unbiased posterior distributions. However, the present PLS approach seems to provide an objective way to reduce the dimensionality of the summary statistics space while retaining as much of the information about the parameters as possible.

Application of ABC–MCMC to African evolution: We estimated the parameters of a model of divergence and migration between three African populations (Figure 1) on the basis of data from 331 microsatellites using our parallelized ABC–MCMC approach with 1000 independent chains of 10,000 simulations (including startup) with proposal range $\varphi = 1$ and tolerance level $\varepsilon = 0.01$. The posterior distributions of the parameters are reported in Figure 4. We find evidence for a very recent divergence between the two Niger–Congo populations (142 generations or ~3550 years ago, based on a generation time of 25 years), which is in very good agreement with the age of the expansion of farming in western Africa and the diversification of the Niger–Congo language family (Wood *et al.* 2005). The divergence between the Pygmies and the Niger–Congo populations is found to be much more ancient, with $T_{\text{mode}} = 2135$ generations ago (~53,400 years), in broad agreement with previous studies (Quintana-Murci *et al.* 2008; Verdu *et al.* 2009), but we note that the 95% highest posterior density credible interval for this time is quite large (1075–3712 generations). While the observed data are compatible with high rates of gene flow between the Yoruba and the Mandenka populations, the pygmies exchange overall many fewer migrants with the two Niger–Congo populations. We find some evidence for higher levels of gene flow from the Pygmies to the Yoruba than in the other direction, an asymmetry already observed in previous analyses of gene flow between Pygmy and neighboring populations (Quintana-Murci *et al.* 2008; Verdu *et al.* 2009). Contrastingly, we find no migration asymmetry between the Pygmies and the Mandenka, which might be expected given the large geographic distance between these two populations. Our results further suggest even lower levels of gene flow between the Pygmies and the ancestral Niger–Congo population, suggesting that the African population was more subdivided at that time. The posterior distributions for the population sizes are all quite wide and point toward relatively large values, even for the Pygmies, with $N_{\text{mode}} = 10{,}876$ gene copies. Interestingly, we found the average mutation rate for tetramicrosatellites ($\overline{\mu} \cong 2.45 \times 10^{-4}$) is much lower than that previously estimated ($\overline{\mu} \sim 6.4 \times 10^{-4}$, Zhivotovsky *et al.* 2003), and its variability across loci is relatively low ($\alpha \cong 14$).

Conclusions: We have shown here how likelihood-free MCMC approaches can be used to produce approximate posterior distributions. Since δ_ε has to be

Figure 2.—Comparison of posterior distributions under a model of population divergence. In each of the six rows, we report the marginal posterior distributions of the four parameters $\theta_1 = 2N_1\mu$, $\theta_2 = 2N_2\mu$, $\theta_A = 2N_A\mu$, and $\tau = T\mu$ of a random test data set. The vertical lines represent the true value of the parameters. The black lines are posterior estimates obtained under ABC–100K (dashed) and ABC–MCMC with 10^5 simulations and tolerance $\varepsilon = 0.01$ (solid). Posterior distributions obtained under ABC–5M are shown as a solid gray line.

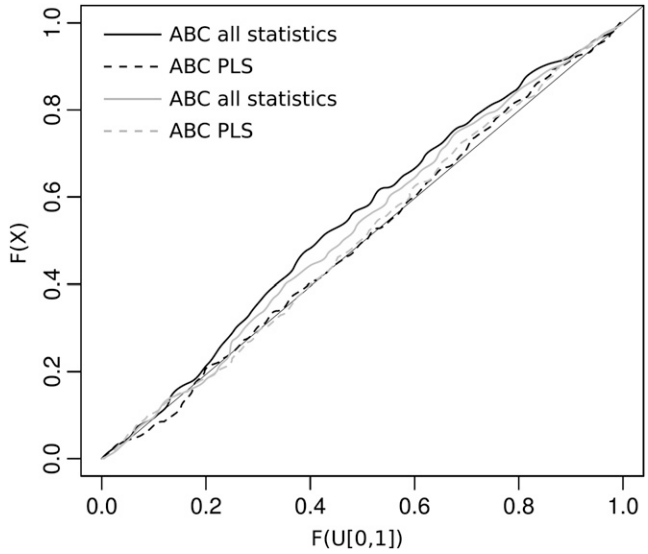


FIGURE 3.—Effect of PLS transformation on posterior distribution coverage properties. We show a *QQ*-plot of the posterior quantiles (X) of true parameter values against those of random variables drawn in $U[0, 1]$. The cumulative distribution function $F(X)$ should be uniformly distributed for ideal posterior distributions. We distinguish the results for a population divergence model without (solid curves) and with migration (shaded curves). Solid lines represent distributions obtained for posteriors computed under ABC–5M by considering all available summary statistics, and dashed lines are for distributions obtained for posteriors computed under ABC–5M by considering only the first 10 PLS components. If all statistics are used to compute posterior distributions, the uniformity of $F(X)$ is rejected by a Kolmogorov–Smirnov test for the model with ($p = 0.011$) and without migration ($p = 0.002$), while uniformity is accepted when PLS components are used ($p = 0.193$ and $p = 0.469$, respectively).

chosen relatively large to ensure adequate mixing, the stationary distribution of an SS–MCMC chain $f(\boldsymbol{\theta} \mid \delta \leq \delta_\varepsilon)$ may still be a relatively crude estimation of the posterior distribution $f(\boldsymbol{\theta} \mid \mathbf{S})$, and the ABC regression adjustment aiming at obtaining $f(\boldsymbol{\theta} \mid \delta \to 0) \approx f(\boldsymbol{\theta} \mid \mathbf{S})$ needs to be performed on the output of the SS–MCMC chain. The transition mechanism of the chain needs to be fine tuned for it to mix properly and lead to posteriors with adequate coverage. The combination of a tolerance level $\varepsilon = 0.01$ and a proposal range $\varphi = 1$ (standard deviation) seems to provide the overall best results under two models of different complexity. Similar values φ and ε should ensure good mixing in other situations, since the widths of the proposal range φ and the initial tolerance interval ε are expressed in a generic fashion: φ is expressed in units of standard deviations of the parameter retained values, which should therefore scale up for different parameters and be adjusted to the observed data, and ε is just a proportion of simulations arbitrarily close to the observations, which depends neither on the choice of summary statistics nor on the parameterization of the model. Note finally that we did not perform any thinning on the output of the ABC–MCMC chain prior to the regression adjustment, as the rejection step is expected to remove a large fraction of the autocorrelation, which should not affect the estimates if the chain converged.

Note that our reported posterior distributions generally have a very similar modal value but are slightly wider than those obtained under a full-likelihood IMa (HEY and NIELSEN 2007) approach (results not shown). While we would not recommend using an ABC approach if a full-likelihood method exists, our results suggest that complex scenarios for which no likelihood-based estimations are available can be relatively well studied with ABC–MCMC, and at a fraction of the computational cost than under conventional ABC. This and similar improvements (*e.g.*, M. A. BEAUMONT, J.-M. CORNUET, J.-M. MARIN and C. P. ROBERT, unpublished results) should be very useful given the growing use of

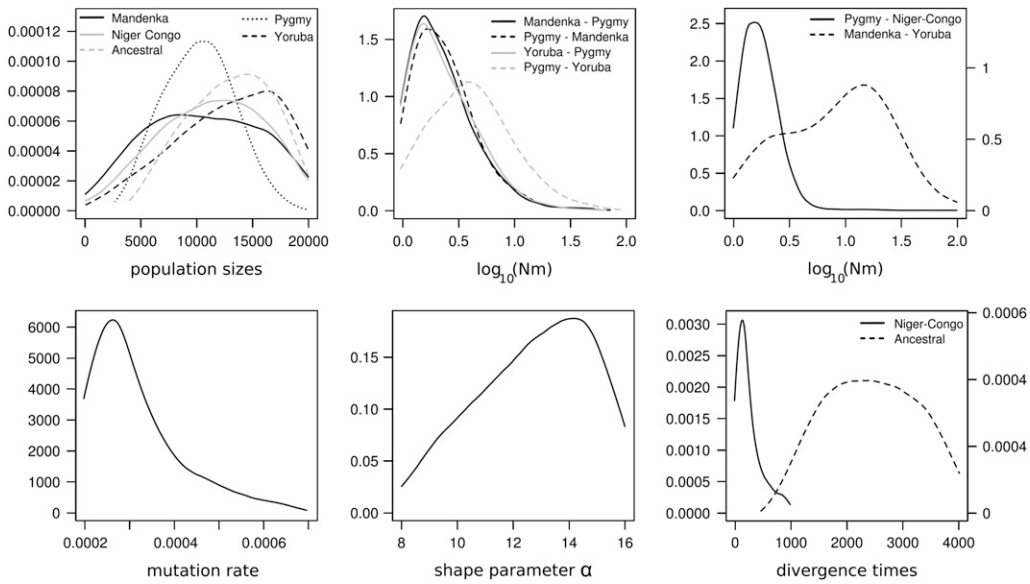


FIGURE 4.—Posterior distributions of the parameters of a model of isolation with migration among three African populations. We used an iterative parallelized ABC–MCMC approach with 1 million simulations, a proposal range $\varphi = 1$, and a final tolerance level $\varepsilon = 0.01$. Posteriors were then estimated from the 5000 best simulations with a locally weighted regression adjustment (BEAUMONT *et al.* 2002). Note that on the two leftmost graphs, the scale of the dashed posteriors is shown on the right *y*-axis.

ABC techniques in demo-genetics studies (see, *e.g.*, Tallmon *et al.* 2004; Excoffier *et al.* 2005a; Hamilton *et al.* 2005; Chan *et al.* 2006; Hickerson *et al.* 2006; Shriner *et al.* 2006; Pascual *et al.* 2007; Legras *et al.* 2007; Rosenblum *et al.* 2007; Cornuet *et al.* 2008; Neuenschwander *et al.* 2008). Indeed, setting up simulation files for an arbitrarily complex model can be done in a few hours using existing simulation programs (*e.g.*, Hudson 2002; Laval and Excoffier 2004; Cornuet *et al.* 2008), allowing one to focus on realistic evolutionary models rather than restricting oneself only to models for which specific programs have been developed. PLS transformation should also allow one to extract as much information as possible from a large set of summary statistics, while keeping the dimensionality of the problem relatively low. While parameter estimation under an ABC framework still requires extensive computing times, it should allow evolutionary geneticists to reasonably estimate the parameters they are really interested in, rather than require them to shift their interest to problems for which full-likelihood solutions are available.

We are grateful to Samuel Neuenschwander, Nicolas Ray, Olivier François, and Gerald Heckel for helpful discussions. We further thank Matthieu Foll, David Balding, Jody Hey, and one anonymous reviewer for their comments on the manuscript. We also thank Nelson Fagundes for suggesting the use of the GW* statistics. This work was supported by a grant from the Swiss National Foundation (no. 3100A0-112072) to L.E.

LITERATURE CITED

Altshuler, D., L. D. Brooks, A. Chakravarti, F. S. Collins, M. J. Daly *et al.*, 2005 A haplotype map of the human genome. Nature **437:** 1299–1320.

Anisimova, M., and D. A. Liberles, 2007 The quest for natural selection in the age of comparative genomics. Heredity **99:** 567–579.

Beaumont, M. A., W. Zhang and D. J. Balding, 2002 Approximate Bayesian computation in population genetics. Genetics **162:** 2025–2035.

Becquet, C., and M. Przeworski, 2007 A new approach to estimate parameters of speciation models with application to apes. Genome Res. **17:** 1505–1519.

Biswas, S., and J. M. Akey, 2006 Genomic insights into positive selection. Trends Genet. **22:** 437–446.

Bortot, P., S. G. Coles and S. A. Sisson, 2007 Inference for stereological extremes. J. Am. Stat. Assoc. **102:** 84–92.

Boulesteix, A. L., and K. Strimmer, 2007 Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Brief. Bioinform. **8:** 32–44.

Box, G. E. P., and D. R. Cox, 1964 An analysis of transformations. J. R. Stat. Soc. Ser. B Stat. Methodol. **26:** 211–252.

Bustamante, C. D., A. Fledel-Alon, S. Williamson, R. Nielsen, M. T. Hubisz *et al.*, 2005 Natural selection on protein-coding genes in the human genome. Nature **437:** 1153–1157.

Chan, Y. L., C. N. Anderson and E. A. Hadly, 2006 Bayesian estimation of the timing and severity of a population bottleneck from ancient DNA. PLoS Genet. **2:** e59.

Cook, S. R., A. Gelman and D. B. Rubin, 2006 Validation of software for Bayesian models using posterior quantiles. J. Comput. Graph. Stat. **15:** 675–692.

Cornuet, J. M., F. Santos, M. A. Beaumont, C. P. Robert, J. M. Marin *et al.*, 2008 Inferring population history with DIY ABC: a user-friendly approach to Approximate Bayesian Computation. Bioinformatics **23:** 2713–2719.

Estoup, A., M. Beaumont, F. Sennedot, C. Moritz and J. M. Cornuet, 2004 Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, Bufo marinus. Evolution **58:** 2021–2036.

Excoffier, L., A. Estoup and J. M. Cornuet, 2005a Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. Genetics **169:** 1727–1738.

Excoffier, L., G. Laval and S. Schneider, 2005b Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol. Bioinform. Online **1:** 47–50.

Fagundes, N. J., N. Ray, M. Beaumont, S. Neuenschwander, F. M. Salzano *et al.*, 2007 Statistical evaluation of alternative models of human evolution. Proc. Natl. Acad. Sci. USA **104:** 17614–17619.

Frazer, K. A., D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve *et al.*, 2007 A second generation human haplotype map of over 3.1 million SNPs. Nature **449:** 851–861.

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

Goldstein, D. B., A. Ruiz-Linares, L. L. Cavalli-Sforza and M. W. Feldman, 1995 Microsatellite loci, genetic distances, and human evolution. Proc. Natl. Acad. Sci. USA **92:** 6723–6727.

Green, R. E., J. Krause, S. E. Ptak, A. W. Briggs, M. T. Ronan *et al.*, 2006 Analysis of one million base pairs of Neanderthal DNA. Nature **444:** 330–336.

Haddrill, P. R., K. R. Thornton, B. Charlesworth and P. Andolfatto, 2005 Multilocus patterns of nucleotide variability and the demographic and selection history of Drosophila melanogaster populations. Genome Res. **15:** 790–799.

Hamilton, G., M. Currat, N. Ray, G. Heckel, M. Beaumont *et al.*, 2005 Bayesian estimation of recent migration rates after a spatial expansion. Genetics **170:** 409–417.

Hey, J., and R. Nielsen, 2007 Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc. Natl. Acad. Sci. USA **104:** 2785–2790.

Hickerson, M. J., E. A. Stahl and H. A. Lessios, 2006 Test for simultaneous divergence using approximate Bayesian computation. Evolution **60:** 2435–2453.

Hoelzel, A. R., J. Hey, M. E. Dahlheim, C. Nicholson, V. Burkanov *et al.*, 2007 Evolution of population structure in a highly social top predator, the killer whale. Mol. Biol. Evol. **24:** 1407–1415.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18:** 337–338.

Joyce, P., and P. Marjoram, 2008 Approximately sufficient statistics and Bayesian computation. Stat. Appl. Genet. Mol. Biol. **7:** 18.

Kuhner, M. K., 2009 Coalescent genealogy samplers: windows into population history. Trends Ecol. Evol. **24:** 86–93.

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

Legras, J. L., D. Merdinoglu, J. M. Cornuet and F. Karst, 2007 Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. Mol. Ecol. **16:** 2091–2102.

Levy, S., G. Sutton, P. C. Ng, L. Feuk, A. L. Halpern *et al.*, 2007 The diploid genome sequence of an individual human. PLoS Biol. **5:** e254.

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

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

Marth, G. T., E. Czabarka, J. Murvai and S. T. Sherry, 2004 The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics **166:** 351–372.

Mevik, B. H., and R. Wehrens, 2007 The pls package: principal component and partial least squares regression in R. J. Stat. Softw. **18:** 1–28.

Neuenschwander, S., C. R. Largiader, N. Ray, M. Currat, P. Vonlanthen *et al.*, 2008 Colonization history of the Swiss Rhine basin by the bullhead (Cottus gobio): inference under a Bayesian spatially explicit framework. Mol. Ecol. **17:** 757–772.

Nielsen, R., 2005 Molecular signatures of natural selection. Annu. Rev. Genet. **39:** 197–218.
Nielsen, R., and J. Wakeley, 2001 Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics **158:** 885–896.
Nielsen, R., I. Hellmann, M. Hubisz, C. Bustamante and A. G. Clark, 2007 Recent and ongoing selection in the human genome. Nat. Rev. Genet. **8:** 857–868.
Pascual, M., M. P. Chapuis, F. Mestres, J. Balanya, R. B. Huey *et al.*, 2007 Introduction history of *Drosophila subobscura* in the New World: a microsatellite-based survey using ABC methods. Mol. Ecol. **16:** 3069–3083.
Plagnol, V., and J. D. Wall, 2006 Possible ancestral structure in human populations. PLoS Genet. **2:** e105.
Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun and M. W. Feldman, 1999 Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol. Biol. Evol. **16:** 1791–1798.
Quintana-Murci, L., H. Quach, C. Harmant, F. Luca, B. Massonnet *et al.*, 2008 Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. Proc. Natl. Acad. Sci. USA **105:** 1596–1601.
Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman *et al.*, 2005 Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc. Natl. Acad. Sci. USA **102:** 15942–15947.
Ratmann, O., O. Jorgensen, T. Hinkley, M. Stumpf, S. Richardson *et al.*, 2007 Using likelihood-free inference to compare evolutionary dynamics of the protein networks of H. pylori and P. falciparum. PLoS Comput. Biol. **3:** 2266–2278.
Rosenblum, E. B., M. J. Hickerson and C. Moritz, 2007 A multilocus perspective on colonization accompanied by selection and gene flow. Evol. Int. J. Org. Evol. **61:** 2971–2985.
Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd *et al.*, 2002 Genetic structure of human populations. Science **298:** 2381–2385.
Schaffner, S. F., C. Foo, S. Gabriel, D. Reich, M. J. Daly *et al.*, 2005 Calibrating a coalescent simulation of human genome sequence variation. Genome Res. **15:** 1576–1583.
Shriner, D., Y. Liu, D. C. Nickle and J. I. Mullins, 2006 Evolution of intrahost HIV-1 genetic diversity during chronic infection. Evolution **60:** 1165–1176.
Sisson, S. A., Y. Fan and M. M. Tanaka, 2007 Sequential Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. USA **104:** 1760–1765.
Tallmon, D. A., G. Luikart and M. A. Beaumont, 2004 Comparative evaluation of a new effective population size estimator based on approximate Bayesian computation. Genetics **167:** 977–988.
Tavare, S., D. J. Balding, R. C. Griffiths and P. Donnelly, 1997 Inferring coalescence times from DNA sequence data. Genetics **145:** 505–518.
Tenehaus, M., J.-P. Gauchi and C. Menardo, 1995 Régression PLS et applications. Rev. Stat. Appl. **43:** 57.
Verdu, P., F. Austerlitz, A. Estoup, R. Vitalis, M. Georges *et al.*, 2009 Origins and genetic diversity of pygmy hunter-gatherers from western central Africa. Curr. Biol. **19:** 312–318.
Williamson, S. H., R. Hernandez, A. Fledel-Alon, L. Zhu, R. Nielsen *et al.*, 2005 Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc. Natl. Acad. Sci. USA **102:** 7882–7887.
Wood, E. T., D. A. Stover, C. Ehret, G. Destro-Bisol, G. Spedini *et al.*, 2005 Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. Eur. J. Hum. Genet. **13:** 867–876.
Zhivotovsky, L. A., N. A. Rosenberg and M. W. Feldman, 2003 Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. Am. J. Hum. Genet. **72:** 1171–1186.

Communicating editor: R. Nielsen

APPENDIX

Marjoram *et al.* (2003) have formally shown that the posterior distribution of the parameters $\boldsymbol{\theta}$ of a given model $f(\boldsymbol{\theta} \mid D)$ could be obtained from the stationary distribution of a MCMC without likelihood, where data D' simulated from the current parameter values $\boldsymbol{\theta}'$ are accepted with probability $h(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}') = \min(1, \pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}' \rightarrow \boldsymbol{\theta})/\pi(\boldsymbol{\theta})q(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}'))$ if D' is equal to the observed data D and rejected otherwise. When data are replaced by a set of summary statistics $\mathbf{S}$, Marjoram *et al.* (2003) have proposed to accept parameters with probability $h(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}')$ if summary statistics $\mathbf{S}'$ simulated from $\boldsymbol{\theta}'$ are arbitrarily close to the observed summary statistics (if $\|\mathbf{S}' - \mathbf{S}\| = \delta \leq \delta_\varepsilon$). In that case they suggest that the stationary distribution of such a SS–MCMC chain is $f(\boldsymbol{\theta} \mid \delta \leq \delta_\varepsilon)$. We show hereafter that this is the case.

Posterior distribution of an SS–MCMC: Let us first assume that the data are summarized by a single continuous statistic, say S. The probability to generate S arbitrarily close to the observed statistics S_O is

$$\Pr(S, \delta < \delta_\varepsilon) = \int_{s=S_O-\delta_\varepsilon}^{S_O+\delta_\varepsilon} f(s)\,ds, \tag{A1}$$

where $f(s)$ is the prior distribution of the statistic. For notational convenience, we note $\Pr(S, \delta < \delta_\varepsilon)$ as $\Pr(\delta < \delta_\varepsilon)$ hereafter. $f(\boldsymbol{\theta} \mid \delta \leq \delta_\varepsilon)$ can therefore be defined as

$$f(\boldsymbol{\theta} \mid \delta < \delta_\varepsilon) = \frac{\Pr(\delta < \delta_\varepsilon \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\Pr(\delta < \delta_\varepsilon)} = \frac{\int_{S_O-\delta_\varepsilon}^{S_O+\delta_\varepsilon} f(s \mid \boldsymbol{\theta})\,ds\,\pi(\boldsymbol{\theta})}{\int_{S_O-\delta_\varepsilon}^{S_O+\delta_\varepsilon} f(s)\,ds}, \tag{A2}$$

where $\pi(\boldsymbol{\theta})$ is the prior of the parameters, and $f(s \mid \boldsymbol{\theta})$ is the conditional density of the statistic.

Following Marjoram *et al.* (2003), if $r(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}')$ is the transition mechanism of the chain to move from state $\boldsymbol{\theta}$ to state $\boldsymbol{\theta}'$, and if we assume that $h(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}') \leq 1$, then

$$
\begin{aligned}
f(\boldsymbol{\theta} \mid \delta \le \delta_\varepsilon) r(\boldsymbol{\theta} \to \boldsymbol{\theta}') &= \frac{\Pr(\delta < \delta_\varepsilon \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta})}{\Pr(\delta < \delta_\varepsilon)} q(\boldsymbol{\theta} \to \boldsymbol{\theta}') \Pr(\delta < \delta_\varepsilon \mid \boldsymbol{\theta}') h(\boldsymbol{\theta}, \boldsymbol{\theta}') \\
&= \frac{\Pr(\delta < \delta_\varepsilon \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta})}{\Pr(\delta < \delta_\varepsilon)} q(\boldsymbol{\theta} \to \boldsymbol{\theta}') \Pr(\delta < \delta_\varepsilon \mid \boldsymbol{\theta}') \frac{\pi(\boldsymbol{\theta}') q(\boldsymbol{\theta}' \to \boldsymbol{\theta})}{\pi(\boldsymbol{\theta}) q(\boldsymbol{\theta} \to \boldsymbol{\theta}')} \\
&= \frac{\Pr(\delta < \delta_\varepsilon \mid \boldsymbol{\theta}') \pi(\boldsymbol{\theta}')}{\Pr(\delta < \delta_\varepsilon)} q(\boldsymbol{\theta}' \to \boldsymbol{\theta}) \Pr(\delta < \delta_\varepsilon \mid \boldsymbol{\theta}) \\
&= f(\boldsymbol{\theta}' \mid \delta \le \delta_\varepsilon) q(\boldsymbol{\theta}' \to \boldsymbol{\theta}) \Pr(\delta < \delta_\varepsilon \mid \boldsymbol{\theta}) h(\boldsymbol{\theta}', \boldsymbol{\theta}) \\
&= f(\boldsymbol{\theta}' \mid \delta \le \delta_\varepsilon) r(\boldsymbol{\theta}' \to \boldsymbol{\theta}),
\end{aligned} \tag{A3}
$$

showing that the chain is fully reversible and that $f(\boldsymbol{\theta} \mid \delta \le \delta_\varepsilon)$ is the stationary distribution of the chain. Equations A1 and A2 can easily be extended to more than one summary statistics. If data are summarized by multivariate statistics $\mathbf{S} = (S_1, \ldots, S_n)$, then $\Pr(\delta < \delta_\varepsilon) = \Pr(\mathbf{S}, \delta < \delta_\varepsilon)$ becomes the multiple integral

$$
\Pr(\delta < \delta_s) = \int_{B(s_0, \delta_s)} f(s)\, ds,
$$

where $B(S_0, \delta_s) = \{s \in \mathbb{R}^n : \|s - S_0\| < \delta_s\}$ is a sphere in the Euclidean n-dimensional space of radius δ_s around S_0 with respect to the chosen norm $\|\cdot\|$.

Note that $f(\boldsymbol{\theta} \mid \delta \le \delta_\varepsilon)$ is also, by definition, the distribution of the retained parameters under a simple ABC algorithm, where randomly generated parameters are accepted if they lead to summary statistics for which $\delta \le \delta_\varepsilon$ and rejected otherwise. It implies that the SS–MCMC approach has the same stationary distribution as an ABC algorithm with similar tolerance level ε. In Figure S1, we empirically show this is the case by reporting the distribution of the distances $\Pr(\delta < \delta_\varepsilon)$ generated under the conventional ABC rejection algorithm, as well as under the SS–MCMC.

GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.109.102509/DC1

Efficient Approximate Bayesian Computation Coupled With Markov Chain Monte Carlo Without Likelihood

Daniel Wegmann, Christoph Leuenberger and Laurent Excoffier

DOI: 10.1534/genetics.109.102509

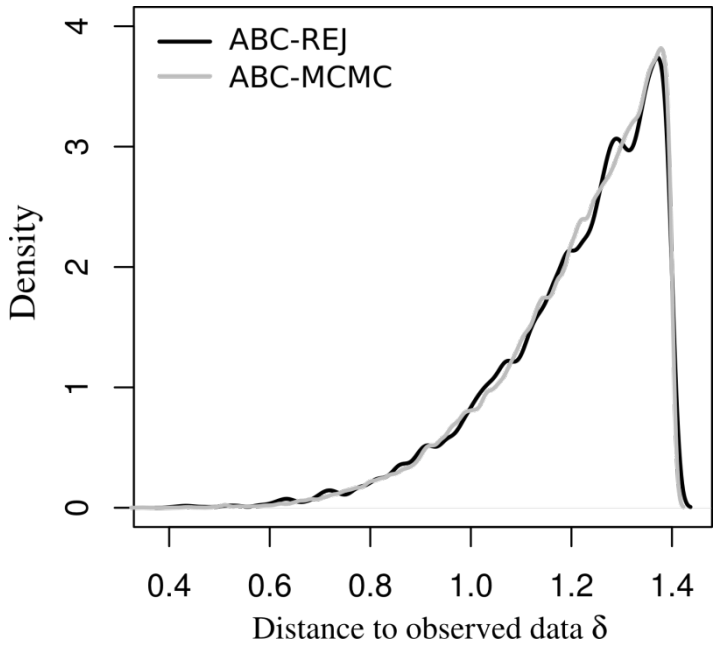


FIGURE S1.—Density distributions of the Euclidean distances between the observed and simulated statistics. We report here Epanechnikov kernel density estimations for distances $\delta < \delta_\varepsilon$ = 1.4 generated under the two sampling schemes ABC and ABC-MCMC with ε =0.005 for the model of population divergence without migration. Due to the tolerance level chosen, the proportion of distances δ <1.4 is expected to be 100 times higher for the ABC-MCMC than conventional ABC. We used 1 million steps for each sampler except ABC, for which we used ten million steps. The densities are estimated on 6434 and 66178 accepted simulations under the ABC and ABC-MCMC samplers respectively. The figure shows that the distributions of Euclidean distances obtained under ABC-MCMC approaches correspond exactly to those of a truncated prior $p_n(\delta \mid \mathbf{S}\boldsymbol{\theta}, \delta \leq \delta_\varepsilon)$.

TABLE S1

Relative accuracy of ABC and ABC-MCMC approaches for a model of population divergence with migration

Approach	φ	ε	*RRMISE*	*RE* mode	*RE* mean	*RE* median	Acceptance Rate [a]	Coverage p-value [b]
ABC-5M[c]	-	-	0.84 (0.14)	0.85 (0.53)	0.85 (0.32)	0.92 (0.36)	-	0.193
ABC-MCMC	5	0.1	0.85 (0.13)	0.84 (0.44)	0.93 (0.29)	0.85 (0.34)	0.18 (0.05)	0.587
	5	0.01	0.83 (0.15)	1.07 (0.56)	0.94 (0.4)	0.97 (0.41)	0.05 (0.04)	0.879
	5	0.001	0.66 (0.26)	1.47 (1.15)	1.11 (1.03)	1.27 (0.79)	0.01 (0.01)	<0.001
	2	0.1	0.86 (0.13)	0.96 (0.45)	0.94 (0.3)	0.98 (0.32)	0.33 (0.08)	0.747
	2	0.01	0.81 (0.15)	0.88 (0.55)	0.9 (0.35)	0.92 (0.39)	0.12 (0.08)	0.506
	2	0.001	0.72 (0.16)	1.15 (0.64)	0.98 (0.45)	0.89 (0.65)	0.03 (0.03)	0.131
	1	0.1	0.86 (0.12)	0.81 (0.45)	0.92 (0.29)	0.90 (0.33)	0.42 (0.09)	0.836
	1	0.01	0.84 (0.15)	0.89 (0.66)	0.85 (0.34)	0.89 (0.4)	0.18 (0.11)	0.286
	1	0.001	0.71 (0.17)	1.03 (0.89)	0.83 (0.55)	0.89 (0.64)	0.05 (0.05)	0.393
	0.5	0.1	0.86 (0.13)	0.94 (0.57)	0.94 (0.26)	0.95 (0.34)	0.48 (0.12)	0.941
	0.5	0.01	0.85 (0.14)	0.96 (0.74)	0.9 (0.37)	0.89 (0.5)	0.20 (0.13)	0.609
	0.5	0.001	0.76 (0.16)	0.97 (0.63)	0.94 (0.44)	0.87 (0.5)	0.07 (0.06)	0.418
	0.1	0.1	0.86 (0.15)	0.99 (0.64)	0.98 (0.43)	0.89 (0.5)	0.52 (0.18)	0.845
	0.1	0.01	0.80 (0.16)	1.03 (0.78)	0.86 (0.45)	0.89 (0.53)	0.24 (0.16)	0.991
	0.1	0.001	0.70 (0.19)	1.22 (0.84)	1.04 (0.51)	1.11 (0.59)	0.08 (0.08)	0.023
	0.01	0.1	0.84 (0.30)	1.53 (2.42)	1.1 (1.12)	1.25 (1.14)	0.54 (0.24)	0.007
	0.01	0.01	0.75 (0.22)	1.27 (1.26)	0.99 (0.72)	1.11 (0.83)	0.27 (0.19)	0.066
	0.01	0.001	0.74 (0.20)	1.4 (1.28)	1.24 (0.65)	1.14 (0.79)	0.09 (0.10)	<0.001
	0.001	0.1	0.85 (0.31)	1.58 (3.31)	1.11 (1.87)	1.36 (2.08)	0.54 (0.26)	<0.001
	0.001	0.01	0.80 (0.22)	1.35 (1.63)	0.99 (0.54)	1.06 (0.92)	0.27 (0.19)	0.049
	0.001	0.001	0.69 (0.34)	1.22 (2.67)	1.02 (0.97)	0.99 (1.31)	0.09 (0.11)	<0.001
	0.0001	0.1	0.88 (0.3)	1.67 (3.08)	1.23 (1.28)	1.27 (1.74)	0.53 (0.25)	<0.001
	0.0001	0.01	0.85 (0.32)	1.5 (1.78)	0.91 (0.94)	1.14 (1.48)	0.27 (0.21)	0.010
	0.0001	0.001	0.74 (0.23)	1.47 (1.67)	1.1 (0.96)	1.08 (1.27)	0.09 (0.11)	<0.001

Mean and standard deviation (within parenthesis) of accuracy measures computed over 100 independent estimations, relative to those performed under a conventional ABC-100K approach. We report ABC-MCMC accuracy for different tolerance values ε and different proposal ranges φ. In all ABC cases, parameter estimation was done after a locally-weighted regression adjustment on the 5000 simulations closest to observations (BEAUMONT *et al.* 2002).

[a] Fraction of ABC-MCMC steps for which new parameter values were accepted.

[b] For each data set, we estimated the posterior quantile of the true value of each parameter from its posterior distribution. We then tested the uniformity of the quantiles values by a Kolmogorov-Smirnov test, which gives an indication on the quality of the coverage of the posterior distributions.

[c] ABC estimation performed on the basis of 5 million simulations