

Bayesian Analysis of an Admixture Model With Mutations and Arbitrarily Linked Markers

Laurent Excoffier,^{*,†,1} Arnaud Estoup^{*} and Jean-Marie Cornuet^{*}

^{*}Institut National de la Recherche Agronomique, Centre de Biologie et de Gestion des Populations (CBGP), Campus International de Baillarguet, 34988 Montferrier-sur-Lez Cedex, France and [†]Computational and Molecular Population Genetics Lab (CMPG), Zoological Institute, University of Bern, 3012 Bern, Switzerland

Manuscript received September 13, 2004
Accepted for publication December 1, 2004

ABSTRACT

We introduce here a Bayesian analysis of a classical admixture model in which all parameters are simultaneously estimated. Our approach follows the approximate Bayesian computation (ABC) framework, relying on massive simulations and a rejection-regression algorithm. Although computationally intensive, this approach can easily deal with complex mutation models and partially linked loci, and it can be thoroughly validated without much additional computation cost. Compared to a recent maximum-likelihood (ML) method, the ABC approach leads to similarly accurate estimates of admixture proportions in the case of recent admixture events, but it is found superior when the admixture is more ancient. All other parameters of the admixture model such as the divergence time between parental populations, the admixture time, and the population sizes are also well estimated, unlike the ML method. The use of partially linked markers does not introduce any particular bias in the estimation of admixture, but ML confidence intervals are found too narrow if linkage is not specifically accounted for. The application of our method to an artificially admixed domestic bee population from northwest Italy suggests that the admixture occurred in the last 10–40 generations and that the parental *Apis mellifera* and *A. ligustica* populations were completely separated since the last glacial maximum.

HYBRID populations have been central to theories on adaptation and speciation (BARTON 2001), and their study has encountered a new interest since it was shown that they could be ideal in detecting disease genes (CHAKRABORTY and WEISS 1988). The assessment of the degree of admixture of a given population has traditionally relied on the comparison of allele frequencies between two potential parental populations and a putative hybrid population (ROBERTS and HIORNS 1965; CHAKRABORTY and WEISS 1988; LONG 1991). Recently, these methods have been improved by incorporating information on the molecular diversity present in the admixed and in parental populations (BERTORELLE and EXCOFFIER 1998; DUPANLOUP and BERTORELLE 2001) or by explicitly taking into account the genetic drift of allele frequencies since the admixture event (CHIKHI *et al.* 2001; WANG 2003). However, the accuracy of the estimation of the contribution of the parental populations to the hybrid depends highly on the extent of differentiation between parental populations (BERTORELLE and EXCOFFIER 1998) and the time elapsed since the admixture event (CHIKHI *et al.* 2001; CHOISY *et al.* 2004). No single method was found to date superior to others in all circumstances (CHOISY *et al.* 2004). Recent

likelihood-based methods, including Bayesian (CHIKHI *et al.* 2001) and maximum-likelihood (WANG 2003) approaches, are computationally intensive but have been shown to produce estimates with smaller variances across independent replicates or simulations, especially when the estimate was based on a small number of loci (WANG 2003; CHOISY *et al.* 2004). A promising alternative to these methods has been the development of an approach using nongenetic information to more precisely define the contribution of sampled populations to the hybrid (GAGGIOTTI *et al.* 2002, 2004). Finally, recognizing that a major drawback of all these former approaches is to require an explicit definition of the source populations, some recent methods have attempted to identify admixed individuals without requiring the source parental populations to be defined (PRITCHARD *et al.* 2000; DAWSON and BELKHIR 2001; ANDERSON and THOMPSON 2002; FALUSH *et al.* 2003), but their statistical power remains to be assessed.

As stated previously, a common problem with most of the previous methods is their inability to explicitly handle mutations (but see BERTORELLE and EXCOFFIER 1998; DUPANLOUP *et al.* 2004), whereas this is likely to be particularly important when the admixture event is ancient. While ML methods have the potential to provide accurate estimations of demographic and mutational parameters, the calculation of likelihoods under models of nonrecent hybridization events for which mu-

¹Corresponding author: Computational and Molecular Population Genetics Lab, Zoological Institute, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland. E-mail: laurent.excoffier@zoo.unibe.ch

tations have to be taken into account at both independent and partially linked markers remains problematic. A powerful Bayesian alternative to likelihood computation for parameter estimation has been introduced recently (FU and LI 1997; TAVARÉ *et al.* 1997; PRITCHARD *et al.* 1999; ESTOUP *et al.* 2001), dubbed as approximate Bayesian computation (ABC; BEAUMONT *et al.* 2002; MARJORAM *et al.* 2003). This approach does not require the computation of likelihoods, but simply relies on the comparison of summary statistics computed on observed data with those computed on data simulated under a model for which the parameters of interest are known (BEAUMONT *et al.* 2002; MARJORAM *et al.* 2003). Although the ABC method relies on summary statistics and thus does not use all available data, it has been shown to provide very accurate results in the analysis of relatively simple evolutionary scenarios where full maximum-likelihood methods were available (BEAUMONT *et al.* 2002; MARJORAM *et al.* 2003). Hence, by construction, ABC methods have the potential to consider models of any complexity, provided only that data can be simulated under the model. Recent applications of the latest developments of ABC methods (BEAUMONT *et al.* 2002) illustrate their potential for the analysis of complex demographic scenarios (ESTOUP and CLEGG 2003; ESTOUP *et al.* 2004). Recent coalescent-based packages (*e.g.*, HUDSON 2002; LAVAL and EXCOFFIER 2004) provide an efficient tool for simulating genetic data under complex scenarios (including introgression or hybridization scenarios) and have the potential to generate data for independent or partially linked markers. Such versatile simulation packages make it possible, even for biologists unfamiliar with simulation algorithms, to perform parameter estimation under the ABC framework and consider various evolutionary scenarios.

In this article, we apply the ABC method to the estimation of all the parameters of an explicit admixture model (Figure 1) defined previously (BERTORELLE and EXCOFFIER 1998; WANG 2003) and described in METHODS. We use the SIMCOAL2 coalescent simulation program (LAVAL and EXCOFFIER 2004) to generate a large number of microsatellite data sets for random values of the admixture model parameters, on which several summary statistics are evaluated. These simulated summary statistics are used for parameter estimation in a series of test data sets, which allows us to validate our approach and to compare its performance with a previously published maximum-likelihood (ML) method (WANG 2003). The method is then applied to the case of an admixed population of honeybees from northwestern Italy.

METHODS

The demographic model: To compare the behavior and performances of our approach with previous methods, we used a classical admixture scenario described

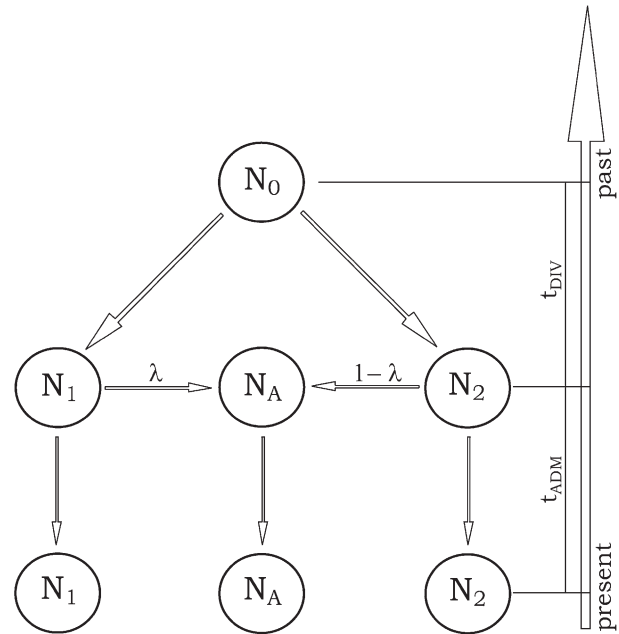


FIGURE 1.—Admixture model considered in this study. A hybrid population is created t_{ADM} generations ago from a mixture of two parental populations that diverged t_{DIV} generations before admixture time. Except for the admixture event itself, all populations are genetically isolated. The demographic model is characterized by seven parameters, which are the effective number of genes in the ancestral (N_0), parental (N_1 and N_2), and admixed (N_A) populations; the times of admixture (t_{ADM}) and divergence (t_{DIV}); as well as the admixture proportion (λ) taken as the relative contribution of parental population 1 to the admixed population.

in Figure 1 and similar to that used in previous studies (*e.g.*, LONG 1991; BERTORELLE and EXCOFFIER 1998; WANG 2003; CHOISY *et al.* 2004).

The genetic model: Unlike almost all methods considering that gene frequencies evolve only through genetic drift, our approach also takes mutations into account (as in BERTORELLE and EXCOFFIER 1998). This involves the choice of a mutation model and of its parameters. We restricted our study to microsatellite markers for which we used a multistep mutation model, sometimes called generalized stepwise mutation (GSM) model (ZHIVOTOVSKY *et al.* 1997; ESTOUP *et al.* 2002), requiring two parameters per locus: the mutation rate (μ_i) and the coefficient (P_i) of the geometric distribution of the length by which a new mutant allele differs from its ancestor. However, these two series of parameters are considered as nuisance parameters, and we will pay attention only to their average values across loci: $\bar{\mu}$ and \bar{P} .

Data thus consist here of multilocus genotypes of n individuals sampled from each of the three populations.

The ABC approach: The rationale and the full description of the ABC method are given in BEAUMONT *et al.* (2002). In short, the approach involves three successive steps detailed in Figure 2. The first step (simulation step) consists of simulating many (typically 1 mil-

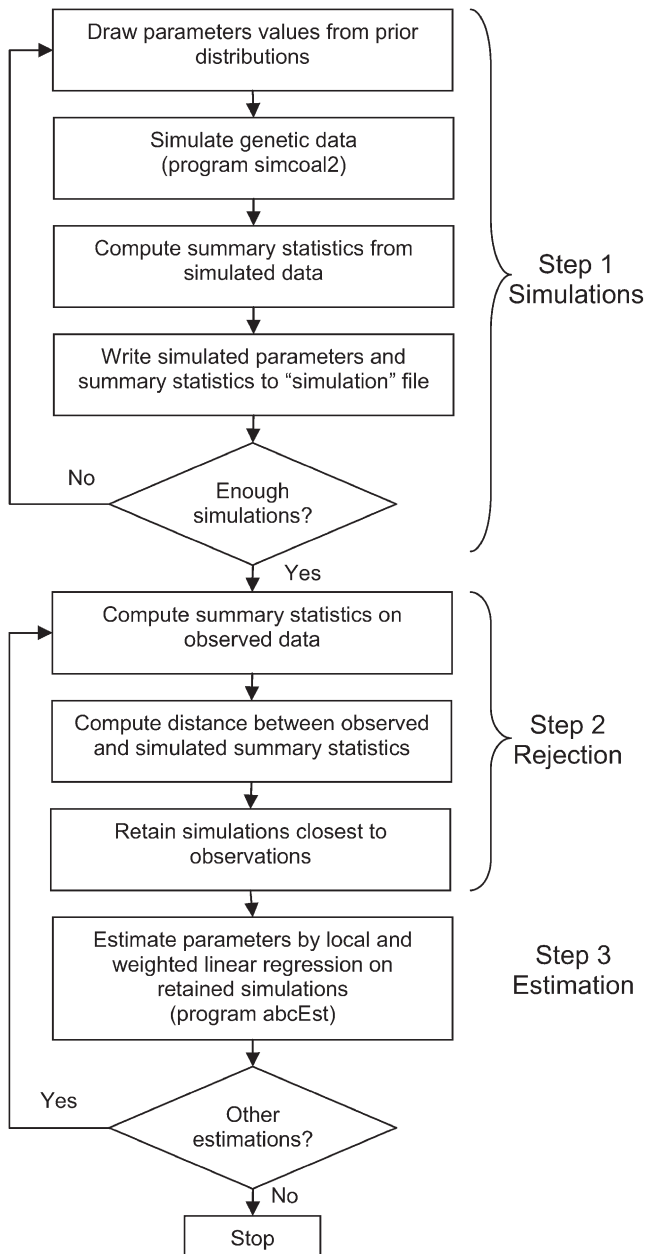


FIGURE 2.—Synopsis of the ABC parameter estimation approach. Step 1 usually includes a loop over hundreds of thousands to millions of simulations. It is the most time-consuming task, generally involving several days of computations. In contrast, the computations in steps 2 and 3 usually take seconds or minutes. To validate the ABC approach, steps 2 and 3 can be repeated hundreds of times on pseudo-observed data generated by step 1-type simulations based on fixed predefined parameter values.

lion) multilocus data sets with characteristics similar to the observed data set (same number of samples, same number of individuals per sample, same number of loci), using parameter values randomly drawn from some prior distributions (as defined in Table 1). The second step consists of comparing the simulated data set to an observed data set, by mean of a series of summary

statistics, retaining the simulations that are arbitrarily close to the observations, and rejecting the other simulations. Finally, the third step is the estimation of the parameters by performing a multiple and locally weighted linear regression on the summary statistics associated with the retained simulations. The set of simulations retained for parameter estimation was selected by strictly following BEAUMONT *et al.* (2002), by computing a Euclidean distance (δ) between simulated and observed summary statistics and retaining the 1000 simulations having the smallest δ distance (being closest) to the test data set.

The SIMCOAL2 program (LAVAL and EXCOFFIER 2004), freely available on <http://cmpg.unibe.ch/software/simcoal2>, has been used to generate microsatellite data sets in the first step, and a new program (abcEst) has been developed for parameter estimation (step 3 in Figure 2). The program abcEst (Windows or Linux version) is available from L. Excoffier upon request. Compared to the published version of the SIMCOAL2 program, two enhancements were added: the implementation of the generalized stepwise mutation model and the possibility of having different mutation rates at different loci. Microsatellite allele size constraints were included in our simulations by imposing reflecting boundaries at the edge of an allele size range of 30 continuous allelic states (FELDMAN *et al.* 1997; POLLOCK *et al.* 1998). This range is consistent with empirical data on repeat numbers at microsatellites in various species (*e.g.*, GARZA *et al.* 1995; GOLDSTEIN and POLLOCK 1997; ESTOUP *et al.* 2000).

Regarding mutation modeling, we draw for each simulation an average mutation rate across loci $\bar{\mu}$ from a log Uniform distribution, and individual locus mutation rates are then drawn from a Gamma distribution with mean equal to $\bar{\mu}$. A similar procedure is also used for the average and individual locus coefficients of the geometric distribution of step lengths \bar{P} and P_i (see Table 1 for details). Note that we have chosen to implement this hierarchy of parameters and did not draw locus-specific parameters μ_i and P_i from unique distributions, since the average parameters $\bar{\mu}$ and \bar{P} would have been virtually identical across simulations of a large number of loci and equal to the mean of the priors. Their estimation would thus have been meaningless. Note also that we have chosen a relatively broad prior for $\bar{\mu}$ compared to previous studies (*e.g.*, WILSON and BALDING 1998), such as to cover a wide range of possible mutation rates (see Table 1).

In addition to the 9 basic parameters of the admixture model (the admixture proportion λ , the four effective population sizes, the time of divergence t_{DIV} , the time of admixture t_{ADM} counted in generations, and the mutational parameters $\bar{\mu}$ and \bar{P}), 11 composite parameters were computed and recorded. They correspond, respectively, to the times of divergence and admixture scaled by the population sizes (t/N_i , with $t = t_{ADM}$ or t_{DIV} , and

TABLE 1
Prior distributions of simulated parameters

Parameters	Distribution	Mean	Mode	Quantiles		
				5%	50%	95%
N_0, N_1, N_2, N_A	Log Uniform[100, 40,000]	6,690	100	135	1,993	29,676
λ	Uniform[0, 1]	0.500	IR	0.05	0.500	0.950
t_{ADM}	Log Uniform[1, 1,000]	143	1	1.4	30.8	704
t_{DIV}	Log Uniform[100, 100,000]	14,540	100	141	3,170	70,942
$\bar{\mu}$	Log Uniform[10^{-4} , 5×10^{-3}]	1.3×10^{-3}	10^{-4}	1.2×10^{-4}	7.1×10^{-4}	4.1×10^{-3}
μ_i	Gamma(2, $2/\bar{\mu}$)	1.3×10^{-3}	1.7×10^{-4}	5.6×10^{-5}	5.5×10^{-4}	4.9×10^{-3}
P	Uniform[0, 0.5]	0.250	IR	0.025	0.250	0.475
P_i	Beta(a, b)*	0.250	IR	0.023	0.250	0.482

$N_0, N_1, N_2,$ and N_A , effective population size (number of gene copies) in ancestral (N_0), parental (N_1 and N_2), and admixed (N_A) populations, respectively; λ , contribution of parental population 1 to the admixed population; t_{ADM} , time since admixture; t_{DIV} , divergence time between parental populations before admixture; $\bar{\mu}$ and μ_i , average and individual-locus mutation rates, respectively; P and P_i , average and individual-locus parameters of the geometric distribution of the GSM, respectively; *Prior distribution for P_i is as follows: if $P \geq 0.001$ then $P = \text{Beta}(a, b)$ with $a = 0.5 + 199P$ and $b = a(1 - P)/P$; otherwise $P = 0$. IR, irrelevant.

$i = 0, 1, 2,$ or A), to the population sizes scaled by the mutation rate ($\theta_i = 2N_i\mu_i$, with $i = 0, 1, 2,$ and A), and to the times of divergence and admixture scaled by the mutation rate ($\tau = 2t\mu$, with $t = t_{ADM}$ or t_{DIV}). The estimation procedure was thus carried out separately on the 9 basic parameters as well as on the 11 composite parameters.

Summary statistics: The following 15 summary statistics were computed on all the simulated microsatellite data sets: the average number of alleles over loci for each of the two parental and the admixed population samples, the average heterozygosity over loci and average modified M statistics (GARZA and WILLIAMSON 2001) over loci for the same three samples, the $(\delta\mu)^2$ genetic distance (GOLDSTEIN *et al.* 1995) between the two parental population samples, the measure of differentiation F_{ST} (WEIR and COCKERHAM 1984) between all three pairs of population samples, the average extent of linkage disequilibrium D' between independent markers in the admixed population, and the m_y admixture coefficient estimator (BERTORELLE and EXCOFFIER 1998). The formula of the modified M statistics is $\sum_{l=1}^L k_l / \sum_{l=1}^L (1 + r_l)$, where k_l is the number of alleles at the l th locus, r_l is the difference in number of repeats between the largest and the smallest allele at locus l (*i.e.*, the range of allele sizes), and L is the number of loci. Compared to its original definition (GARZA and WILLIAMSON 2001), it just avoids a division by zero when a gene sample is fixed for a single allele. Note that the summary statistics were chosen such as to capture different features of the data, both at the within- and at the between-population level. This choice is partially arbitrary, since there is currently no objective way to define an optimal set of statistics (BEAUMONT *et al.* 2002), but we have tried to use statistics thought to be informative for some of the parameters of our model. For instance, one would ex-

pect heterozygosity to be informative for the estimation of population size, but it should also depend on the admixture proportion in the hybrid population. Also, pairwise F_{ST} 's are expected to bring information about divergence times between parental populations and about admixture proportions. The m_y admixture coefficient should obviously bring information on admixture proportion, while D' in the admixed population should decay with admixture time, but also depend on the absolute sizes of the populations (drift). However, we did not attempt here to define an optimal set of statistics or to study the effect of removing or adding summary statistics, which could be the subject of a later study.

Simulated data sets: A first series of 10^6 data sets was simulated and consisted of 50 diploid individuals (100 genes) typed at 50 independent microsatellite loci. This large data set was fractioned into subsets to study the effect of sample size and number of loci on parameter estimation, and thus data sets consisting of 5, 10, 20, and 50 loci studied in samples of 20 and 100 genes were obtained. A second series of 10^6 data sets, consisting of 50 diploid individuals typed at a mixture of 20 independent and partially linked loci, was simulated. The 20 loci consisted of two unlinked groups of 10 partially linked loci. Each group of 10 partially linked loci was itself divided into two subsets of 5 completely linked loci (genetic distance of 0 cM), 1 cM distant from each other. The 190 pairs of loci thus fell into three linkage categories: unlinked (100 pairs of loci), partially linked at 1 cM (50), and totally linked (40). The coefficient of linkage disequilibrium D' was computed separately in the three categories of markers, thus adding two summary statistics to these simulated data sets with recombination. Note that our choice of three categories of linkage is somewhat arbitrary. While the "completely linked" and independent sets of markers are easy to

TABLE 2

Effect of the number of independent loci on the estimation of the admixture rate λ by the ABC and WANG's (2003) methods

Sample size	No. of loci	ABC											
		100,000 simulations				1,000,000 simulations				WANG03			
		Bias	RMSE	Coverage 95%	Factor 2	Bias	RMSE	Coverage 95%	Factor 2	Bias	RMSE	Coverage 95%	Factor 2
20	5	-0.057	0.232	0.99	0.98	-0.029	0.228	0.99	0.99	0.003	0.212	0.95	0.98
	10	-0.055	0.181	1.00	0.99	-0.046	0.171	1.00	1.00	0.016	0.157	0.89	1.00
	20	-0.044	0.125	1.00	1.00	-0.054	0.118	0.99	1.00	-0.010	0.114	0.93	1.00
	50	-0.012	0.071	1.00	1.00	-0.010	0.074	0.99	1.00	0.003	0.074	0.93	1.00
100	5	-0.089	0.175	1.00	1.00	-0.078	0.162	1.00	1.00	-0.026	0.134	0.89	1.00
	10	-0.051	0.125	1.00	1.00	-0.043	0.112	1.00	1.00	-0.009	0.091	0.86	1.00
	20	-0.035	0.091	1.00	1.00	-0.025	0.079	0.99	1.00	-0.009	0.070	0.87	1.00
	50	-0.003	0.058	1.00	1.00	-0.010	0.042	0.99	1.00	-0.004	0.039	0.89	1.00

Simulated conditions are independent loci. $\lambda = 0.3$, $t_{ADM} = 5$, $t_{DIV} = 5000$, $N_0 = N_1 = N_2 = 300$. Bias and root mean square error (RMSE) are expressed in relative units. Coverage 95% represents the number of times among 100 that the true value of λ (0.3) lies within the estimated 95% confidence interval. Factor 2 represents the number of times that the true value of λ lies within an interval limited by 50 and 200% of the estimated λ value.

justify and are commonly found in many data sets, the spacing of 1 cM was chosen such as to have a different amount of loss of potential disequilibrium created by the admixture process over the time periods studied below. Indeed, one would expect that markers 1 cM apart would lose ~ 5 , 63.4, and 98.2% of the original disequilibrium caused by the admixture after 5, 100, and 400 generations, respectively, thus allowing one to potentially use linkage disequilibrium (LD) to estimate admixture time.

Performance evaluation and test data sets: The performances of our ABC approach were evaluated in a series of samples having fixed values of the admixture model. For each combination of parameters, the SIMCOAL2 program was used to generate 100 data sets, on which summary statistics were computed and then used as pseudo-observed summary statistics. The same data set was also used as input to a recent ML method (WANG 2003) denoted hereafter WANG03. The latter method has been chosen for a comparison with our approach, because it has been shown to produce good estimates of admixture coefficients, and because it estimates other parameters of the admixture model that can be also compared with those of our ABC method. Moreover, compared to the method of CHIKHI *et al.* (2001), Wang's ML method was notably faster, allowing us to get 100 estimates for fixed simulated parameter values in a reasonable amount of time.

It is worth noting that while the simulation of 1 million data sets and the computation of their associated summary statistics for our ABC approach is time consuming (~ 12 hr on 15 computer nodes), the ABC estimation of the parameters on a given test data set takes only seconds to minutes, so that the evaluation of the performance of our estimation procedure can be easily achieved without much additional computing cost. This

evaluation was thus performed in seven situations. Due to the huge amount of computations needed for the comparisons presented here, a few parameters were fixed across the simulations. The population sizes (numbers of genes) were set to 300, the average mutation rate to 0.0005 (reviewed in ELLEGREN 2004), and the geometric coefficients P to 0.3 (*e.g.*, ESTOUP *et al.* 2002). The first situation modeled a recent admixture ($t_{ADM} = 5$ generations, *i.e.*, $t_{ADM}/N_e = 0.0167$), an ancient divergence ($t_{DIV} = 5000$ generations, *i.e.*, $t_{DIV}/N_i = 16.7$), and an admixture proportion of 0.3. This situation was used to evaluate the effects of different numbers of loci and of different sample sizes (Table 2). The other six situations were chosen to evaluate the effects of increasing the time of admixture for two different admixture proportions and of having partially linked markers. The performance of our ABC method and of WANG03 was characterized by the *relative bias* (average difference between the estimate and the true value divided by the true value), the *relative root mean square error* (RMSE—square root of the mean square error divided by the true value), the *95% coverage* (proportion of times in which the true value is within the equal-tailed 95% confidence or credible interval around the estimate), and the *factor 2* (proportion of times in which the estimated value is in an interval bounded by values equal to 50 and 200% that of the true value). All measurements of bias, RMSE, and factor 2 were computed by taking the mode of the posterior distribution as a point estimate. The factor 2 parameter is intuitively appealing and brings qualitatively different information than the 95% coverage. It indeed tells users how often the estimator is arbitrarily close (factor 2 here) to the true value, while the inclusion of the true value within a confidence interval does not imply that the estimated parameter is “close” to its true value, as this depends on the width of this interval.

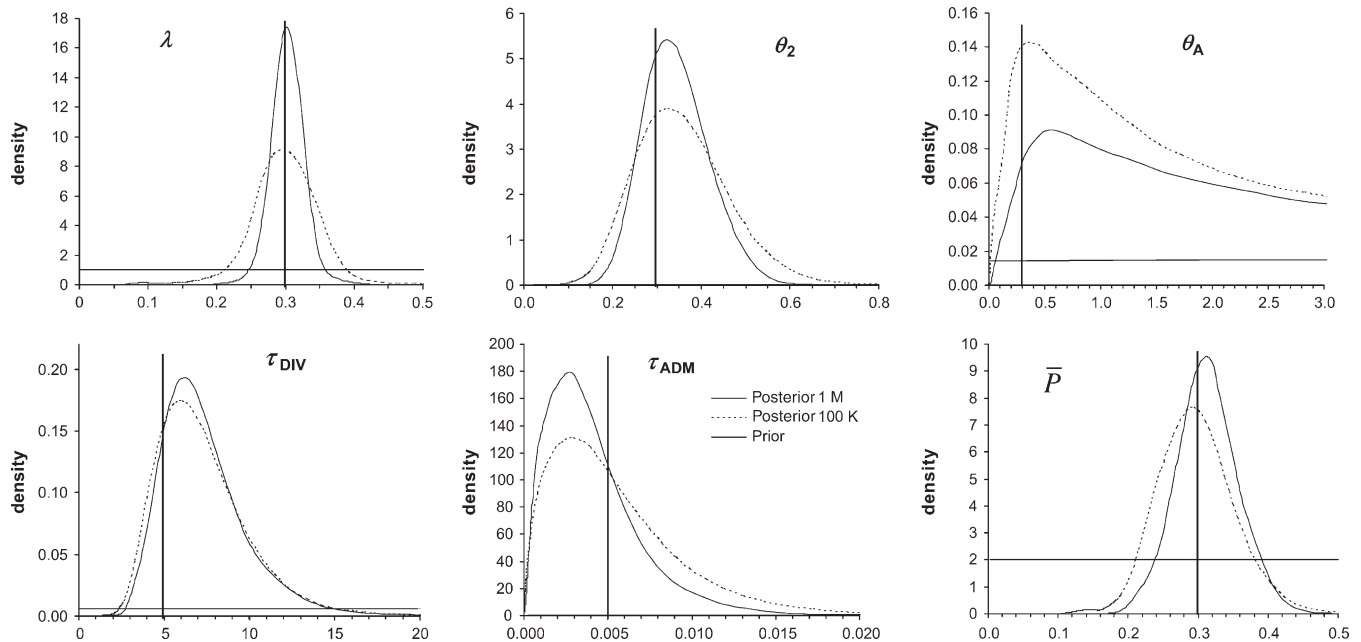


FIGURE 3.—Posterior distributions of some parameters of the admixture model. We contrast here posterior distributions obtained from an analysis performed on a set of 1 million (solid line) or 100,000 (dashed line) simulated summary statistics. In both cases, the estimation and the posterior distribution were obtained by a local weighted regression (BEAUMONT *et al.* 2002) on the 1000 simulations closest to the test data set. True parameter values are shown as vertical boldface lines: $N = 300$ for all population sizes; admixture rate, $\lambda = 0.3$; divergence time between populations, $t_{DIV} = 5000$ generations; admixture time, $t_{ADM} = 5$; mutation rate, $\bar{\mu} = 5 \times 10^{-4}$; and parameter of the geometric distribution of mutation steps, $\bar{P} = 0.3$. Note that the posterior distributions shown here are the output of a single (randomly chosen) analysis, and that they are not averaged over 100 replicates as reported in Tables 2–4.

All measures of performance were estimated over 100 simulated test data sets. Note that 100 replicates may not be enough to get very accurate estimates of relative RMSEs, so that the numbers for this measure should be considered as indicative only.

RESULTS

Recent admixture events: The performance of the ABC method on the recovery of admixture proportions λ for different numbers of loci and different samples sizes is reported in Table 2 and compared to the ML method of WANG (2003). This comparison is based on a scenario that can be considered as advantageous for admixture estimation, because it involves a small admixture time (5 generations) and a long divergence time (5000 generations) relative to the population size (300 genes). In that case, when ABC estimation is performed on 1 million simulated samples, its performance is very similar to Wang’s ML method, as attested by the relative RMSE, especially when the number of loci is high (20 or more). As expected, estimations obtained with 1 million simulations are more accurate than those obtained with 100,000 simulations. However, the latter are already quite good with virtually identical negative relative bias and only slightly larger relative RMSE. Note, however, that the same trend is visible in Figure 3, where we report the posterior distributions obtained from the

analysis of a single (randomly chosen) case from 10^6 or 10^5 simulations. While the modes of the distributions (taken as a point estimate) obtained from the analysis of 10^6 or 10^5 simulations are very similar, the distributions obtained from 10^6 simulations are usually narrower and would lead to smaller credible intervals. We note here that the ABC method generally produces a small negative bias consisting of underestimating the contribution of the source population contributing the least to the admixed population, but that this bias becomes negligible with a larger number of loci.

The ABC and Wang’s ML methods are found consistent as their accuracy increases with larger samples sizes and larger numbers of loci. They both produce estimates that are almost always closer than a factor 2 from the true value. The only notable difference between the two methods is in the coverage of the 95% confidence intervals around the estimated values: the ABC method tends to produce conservative (too broad) intervals, while Wang’s ML method gives too narrow intervals with larger samples where the true value is found only in $<90\%$ of the cases.

Old admixture events: In Table 3, we report the effect of older admixture times on the estimation of the admixture rate for 20 independent or 20 partially linked loci. While the ABC and Wang’s ML methods have very similar performance for short admixture time, the ABC method produces more accurate results when the ad-

TABLE 3

Effect of admixture time and partial linkage on the estimation of the admixture rate λ by the ABC and WANG's (2003) methods

	t_{ADM}	ABC				WANG03			
		Bias	RMSE	Coverage 95%	Factor 2	Bias	RMSE	Coverage 95%	Factor 2
20 unlinked loci									
$\lambda = 0.1$	5	-0.107	0.193	1.00	0.99	-0.015	0.126	0.90	1.00
	100	0.029	0.575	0.91	0.80	-0.470	0.643	0.52	0.16
	400	0.574	0.875	0.99	0.71	0.568	2.294	0.23	0.13
$\lambda = 0.3$	5	-0.025	0.079	0.99	1.00	-0.009	0.070	0.87	1.00
	100	-0.045	0.253	0.94	0.97	-0.123	0.412	0.52	0.84
	400	-0.032	0.384	0.99	0.92	0.348	1.078	0.12	0.43
20 partially linked loci									
$\lambda = 0.1$	5	-0.060	0.311	0.97	0.94	-0.031	0.323	0.38	0.95
	100	0.326	1.013	0.85	0.55	-0.077	1.059	0.37	0.37
	400	0.970	1.314	0.98	0.54	0.351	2.218	0.13	0.05
$\lambda = 0.3$	5	-0.050	0.194	0.93	1.00	-0.013	0.172	0.33	1.00
	100	-0.040	0.436	0.89	0.83	-0.049	0.570	0.32	0.69
	400	0.161	0.554	0.99	0.84	0.359	1.244	0.04	0.25

Simulated conditions are 10^6 simulations; sample size, 100 genes. $\lambda = 0.3$, $t_{DIV} = 5000$, $N_0 = N_1 = N_2 = N_A = 300$.

mixture event occurred >100 generations ago, as shown by much smaller relative RMSE values, higher factor 2 scores, and much better coverage properties for the ABC than for the ML method. For both unlinked and partially linked loci, it is important to note that the coverage of the ABC 95% confidence intervals is always very good. On the other hand, confidence intervals provided by the ML method become poorer with longer admixture time for unlinked loci and are already much too low in the case of a recent admixture studied with partially linked loci. The latter effect is certainly due to the fact that the ML method assumes that the loci are unlinked. As a consequence, loci that are correlated provide similar information and tend to generate thinner distributions because they overestimate the amount of information in the data. This is not the case for the ABC method since we explicitly model the correlation between partially linked markers in our simulations.

While 20 independent loci provide accurate estimation of admixture rates, there is a serious drop in the quality of the ABC estimates based on partially linked markers, especially for very unequal contribution of the parental population to the admixed population (*i.e.*, $\lambda = 0.1$). The decrease in ABC accuracy between linked and unlinked loci is especially marked for older admixture events. Curiously, the ML method is less affected than the ABC method by partial linkage, in the sense that its performance evaluated by the relative bias and RMSE does not degrade much when partially linked markers are used instead of independent markers. However, although the ABC method somewhat suffers from the use of nonindependent loci, its relative RMSE remains two

to three times lower than that obtained from the ML method for the oldest admixture times (400 generations).

Estimation of divergence and admixture times: Wang's ML method provides estimates of composite parameters such as divergence and admixture times scaled by population sizes; we report in Table 4 the corresponding parameters obtained from the ABC method. Because this ML method assumes that no mutation occurred since the divergence of the two parental populations, and thus that genetic differences between populations are due to a pure drift process, it leads to grossly underestimated divergence and admixture times and presents poor coverage property, even for recent admixtures. By contrast, the divergence time scaled by parental population size N_2 (t_{DIV}/N_2) is only slightly overestimated with the ABC method from both linked and unlinked markers, with good coverage and factor 2 scores. The admixture time scaled by parental population size N_2 (t_{ADM}/N_2) is very well estimated by the ABC method when it is relatively ancient and is underestimated only by 12 and 48% on average when it is recent (five generations) for unlinked and linked markers, respectively. This parameter is also, to a lesser extent, well estimated by the ML method when admixture is recent. However, it is increasingly underestimated for older admixture times, resulting in a virtual absence of coverage by the ML confidence intervals for admixture times ≥ 100 generations. Finally, the admixture time scaled by the admixed population size N_A (t_{ADM}/N_A) is only relatively well estimated by the ML method for recent admixtures. Its estimation follows a more complex pattern for the ABC

TABLE 4
Effect of admixture time and partial linkage on the estimation of various composite parameters by the ABC and WANG's (2003) methods

Parameters	t_{ADM}	ABC				WANG03			
		Bias	RMSE	Coverage 95%	Factor 2	Bias	RMSE	Coverage 95%	Factor 2
t_{DIV}/N_2									
Unlinked loci	5	0.134	0.427	0.99	0.97	-0.976	0.976	0.00	0.00
	100	0.226	0.633	0.96	0.85	-0.982	0.982	0.00	0.00
	400	0.232	0.636	1.00	0.83	-0.978	0.978	0.00	0.00
Linked loci	5	0.343	0.727	0.97	0.84	-0.975	0.975	0.00	0.00
	100	0.315	0.673	0.98	0.89	-0.980	0.980	0.00	0.00
	400	0.226	0.636	0.99	0.80	-0.978	0.978	0.00	0.00
t_{ADM}/N_2									
Unlinked loci	5	-0.122	0.534	0.97	0.73	-0.720	0.759	0.20	0.19
	100	0.007	0.581	0.99	0.79	-0.841	0.849	0.00	0.00
	400	0.039	0.460	0.97	0.88	-0.969	0.970	0.00	0.00
Linked loci	5	-0.478	0.556	0.98	0.45	-0.690	0.734	0.22	0.23
	100	-0.019	0.561	0.97	0.69	-0.883	0.889	0.00	0.00
	400	-0.048	0.536	0.98	0.78	-0.964	0.965	0.00	0.00
t_{ADM}/N_A									
Unlinked loci	5	-0.531	0.575	0.85	0.37	-0.625	0.639	0.02	0.19
	100	0.101	1.489	1.00	0.68	-0.764	0.766	0.00	0.00
	400	0.014	0.337	0.98	0.96	-0.904	0.904	0.00	0.00
Linked loci	5	-0.565	0.740	0.83	0.27	-0.773	0.798	0.05	0.09
	100	0.059	1.148	1.00	0.65	-0.773	0.776	0.00	0.00
	400	0.018	0.442	1.00	0.88	-0.911	0.912	0.00	0.00

Simulated conditions are 10^6 simulations, 20 loci, sample size 100 genes, $\lambda = 0.3$, $t_{DIV} = 5000$, $N_0 = N_1 = N_2 = N_A = 300$.

method. The bias is large and negative for recent admixture events, and it becomes positive and associated with a large RMSE for $t_{ADM} = 100$; for older admixture times ($t_{ADM} = 400$), the bias becomes very low and the relative RMSE drops considerably. This pattern is probably due to the poor estimation of the admixed population size N_A for short admixture times, since small or large population sizes will not create very contrasting patterns of diversity in a few generations, while they should lead to more contrasted patterns for longer evolutionary periods such as a few hundred generations.

ABC estimation of mutation-scaled parameters: In Table 5, we present results on the estimation of composite parameters depending on mutations. These parameters are computed only in the ABC method so that comparison with Wang's ML method is not possible. The scaled divergence time τ_{DIV} is relatively well estimated for short admixture time (17% of positive bias) and its relative RMSE is only slightly increased with longer admixture times, resulting in a small drop (96–89%) for the factor 2 score. The scaled admixture time τ_{ADM} is increasingly better estimated with older admixture events, in keeping with results obtained for the scale parameter t_{ADM}/N_2 . The relatively poor recovery of this parameter for recent admixture is also visible in Figure 3, where the posterior distribution of τ_{ADM} is not centered at all around the true value in that case. The scaled population sizes θ_A and θ_2 (θ_1 is not shown in

Table 5, but follows the same pattern as θ_2) are very well estimated even for old admixture times, while the scaled size of the admixed population (θ_A) is better estimated with increasing admixture times. For $t_{ADM} = 400$, θ_A estimation shows virtually no relative bias (-0.4%), a relative RMSE (31%) becoming very similar to that of θ_2 (26%), and an excellent factor 2 score (98%). The relatively flat posterior distribution of θ_A for recent admixtures (five generations) underlines the absence of information in the data for such recent events (Figure 3). On the other hand, the mean parameter of the geometric distribution of the GSM model \bar{P} is well estimated with 20 loci and does not seem much affected by the age of admixture. Finally, we note that the coverage of the 95% confidence intervals is very good for all parameters and tends to be too conservative except for \bar{P} .

Application to a honeybee data set: This honeybee data set has been previously described and analyzed in CHOISY *et al.* (2004). The population under study is located in Courmayeur at the extreme north of the Aosta valley (northwestern Italy) and represented by a sample of 33 worker bees (one per colony). It is considered an artificially admixed population between two different subspecies of *Apis mellifera*, the West-European black honeybee (*A. m. mellifera*) and the Italian yellow honeybee (*A. m. ligustica*). The two parental populations are represented in the analysis by a sample of *A. m. mellifera* from the sanctuary of Ouessant (French Brit-

TABLE 5

Effect of admixture time on the estimation of various composite parameters depending on the mutation rate, as well as the admixed population size (N_A) by the ABC method

Parameters	t_{ADM}	ABC			
		Bias	RMSE	Coverage 95%	Factor 2
$\tau_{DIV} = 2\mu t_{DIV}$	5	0.170	0.437	1.00	0.96
	100	0.128	0.447	1.00	0.96
	400	0.231	0.575	1.00	0.89
$\tau_{ADM} = 2\mu t_{ADM}$	5	-0.158	0.439	0.97	0.85
	100	-0.034	0.429	0.99	0.80
	400	-0.016	0.412	0.97	0.89
$\theta_2 = 2N_2\mu$	5	0.069	0.248	1.00	1.00
	100	-0.025	0.272	0.98	0.97
	400	0.032	0.257	0.98	0.98
$\theta_A = 2N_A\mu$	5	2.666	3.425	1.00	0.24
	100	-0.147	0.477	1.00	0.82
	400	-0.004	0.307	0.99	0.98
\bar{P}	5	-0.024	0.210	0.97	0.98
	100	-0.019	0.239	0.95	0.95
	400	-0.052	0.246	0.95	0.95
N_A	5	5.594	6.230	0.99	0.01
	100	0.675	0.924	1.00	0.75
	400	0.090	0.315	1.00	0.99

tany, $n = 49$) and a sample of *A. m. ligustica* from Forli (Emilia-Romania, $n = 19$), an area of intensive queen rearing for exportation. All sampled honeybees were characterized at eight microsatellite loci, and the admixture coefficient of the Courmayeur sample has already been estimated by six different methods (see CHOISY *et al.* 2004 for more details). Such estimates of the proportion of *A. m. mellifera* genes in the Courmayeur genetic pool ranged from 0.195 to 0.371 (CHOISY *et al.* 2004). Table 6 shows that our ABC estimate (0.259) is well within this range, as is Wang's ML estimate (0.287). These two methods also agree in their estimates of the time of admixture, which is ~ 0.01 – 0.02 in units of N . Considering that effective population sizes (in number of gene copies) in European honeybee subspecies are of the order of 1000–2000 [ESTOUP *et al.*'s (1995) Table 4], this implies a rather recent admixture of 10–40 generations, corresponding to 20–80 years (using an average generation time of 2 years for the queens). This is in good agreement with the development of the Italian queen selling industry in Europe in the middle of the twentieth century. As expected from our previous simulations (Table 4), the two methods provide very different estimates of the time of divergence of the two parental populations scaled by effective population sizes. Wang's ML estimates are ~ 0.15 – 0.25 , whereas the ABC esti-

mates reach 7.2–8.3. *A. m. ligustica* and *A. m. mellifera* have long been considered as two very distinct subspecies of honeybees. At the end of the 1980s (*e.g.*, RUTTNER 1988), the current theory based on paleogeography and morphometry was that the Quaternary ice ages were responsible for the separation of the two subspecies, so the divergence time was estimated at $\sim 50,000$ years before present (BP). However, mitochondrial studies showed that these two subspecies belonged to two highly divergent lineages having probably diverged ~ 1 million years ago (GARNERY *et al.* 1992). Quite recently, FRANCK *et al.* (2000) showed that the subspecies *ligustica* had actually a hybrid origin using a much larger sample of colonies, and that its genetic pool was a mixture of two lineages: the M lineage constituted mainly by the *mellifera* subspecies and the C lineage encompassing the South-European subspecies *carnica* and *cecropia*, as well as the Asian *caucasica*. According to FRANCK *et al.* (2000), the admixture might have taken place any time after the Riss period (in the last 130,000 years), and it is probably rather ancient. The estimated divergence time could thus not correspond to the separation of the C and M lineages, but rather to the time when the admixed *ligustica* and the *mellifera* subspecies were last separated. If we admit the timing given by FRANCK *et al.* (2000), a sensible estimate would be some time during the last ice age (which at maximum occurred 22,000–14,000 years BP), when honeybee populations were restricted to southern Mediterranean refuges (namely the Iberian and Italian peninsulas, respectively). Taking population sizes as above, we get divergence time estimate intervals of 150–500 years with Wang's ML estimates and 14,400–33,200 years with our ABC approach. Wang's ML estimates for the time of divergence of the two subspecies hence appear clearly underestimated, while the ABC method gives estimates much more compatible with our current knowledge of the evolutionary history of European honeybee populations.

The ABC approach also allows the estimation of several other parameters not estimated by Wang's ML method (Table 6), such as the mutation scaled population sizes (θ 's) and the times of divergence τ_{DIV} or admixture τ_{ADM} . Using the mode of the posterior distribution of the average mutation rate (1.85×10^{-4}), we obtain an estimate of 23,665 generations (47,330 years) for the divergence time and 26 generations (52 years) for the time of admixture. Both values are in excellent agreement with those mentioned above and with other studies (RUTTNER 1988; FRANCK *et al.* 2000). The average geometric coefficient \bar{P} of the GSM mutation model is very high (0.446) and very close to the upper bound of our prior distribution (Table 1). This extreme value implies a surprisingly large proportion of mutations leading to non-single-step mutations (precisely 0.446; ESTOUP *et al.* 2002). This probably results from the fact that this data set does not fit well to the modeled scenario. More specifically, the potential hybrid nature of one parental

TABLE 6
Estimated parameters of the admixture model for the Courmayeur honeybee sample

Parameters	ABC		WANG03	
	Mode	95% CI	Mode	95% CI
λ	0.26	0.11–0.38	0.29	0.15–0.40
Parameters scaled by population sizes				
t_{DIV}/N_1	8.25	3.11–97.39	0.25	0.15–0.38
t_{DIV}/N_2	7.17	2.00–64.56	0.15	0.076–0.242
t_{ADM}/N_1	0.018	0.0003–0.071	0.0053	<0.0005–0.019
t_{ADM}/N_2	0.007	0.0002–0.044	0.0176	<0.0005–0.035
t_{ADM}/N_A	0.027	0.0003–0.062	0.0170	0.009–0.025
Parameters scaled by the mutation rate				
θ_0	2.27	0.04–12.34		
θ_1	0.73	0.29–1.62		
θ_2	1.15	0.57–2.50		
θ_A	0.61	0.02–6.06		
τ_{DIV}	8.76	2.20–74.67		
τ_{ADM}	0.001	0.0002–0.053		
Parameters of the mutation model				
$\frac{\bar{\mu}}{P}$	1.85×10^{-4}	3.9×10^{-5} – 8.8×10^{-4}		
P	0.45	0.32–0.51		

Simulated conditions are 10^6 simulations, prior distributions are as in Table 1. CI, credibility interval.

population (*ligustica*) may have widened the distribution of allele lengths in the corresponding sample, forcing the analysis to increase the average length of the mutation steps to cope with this widened allelic distribution.

DISCUSSION

This study shows that the ABC framework allows a fine analysis of an admixture model, providing very satisfactory estimates of admixture rate (λ), mutation-scaled parental population sizes (θ_1 and θ_2), and divergence time τ_{DIV} , as well as those of the mutation model. Estimates of scaled ancestral population size (θ_0) are usually poor, and those of the admixed population size (θ_A) are good only when the admixture time is large. The mutation-scaled admixture time (τ_{ADM}) is itself very well estimated when the admixture event is relatively old (100 or more generations), while it leads to reasonable point estimates but large credible intervals when it is very recent. Unscaled parameters, such as raw population sizes and raw divergence and admixture times, were usually not estimated as well as the scaled parameters (results not shown), as they do not have independent and contrasting effects on genetic diversity. However, it is worth noting that the size of the admixed population N_A was very well estimated in the case of old admixture events (*i.e.*, 400 generations). As shown in Table 5, the relative bias on N_A is indeed <10% when the admixture time is 400 generations, while it was $\sim 560\%$ for an admixture event only 5 generations old. This result sug-

gests that the absolute size of old admixed populations could be well estimated under our framework. This is probably because our method implicitly attempts to reconstruct the genetic composition of the admixed population at the time of admixture, which puts us into a framework very similar to a temporal spacing of samples, which is the ideal situation for estimating population sizes independently from mutation rates (*e.g.*, WILLIAMSON and SLATKIN 1999; ANDERSON *et al.* 2000; BERTHIER *et al.* 2002).

Compared to Wang's ML method, our ABC approach shows comparable performance for the estimation of the admixture coefficient when admixture is recent, but leads to increasingly better relative results when the admixture time is older. We attribute this better performance to the specific handling of mutations, which cannot be neglected when admixture time is ancient. However, to estimate admixture coefficients, methods based on a pure drift process are not handicapped by mutations having occurred before the admixture, as they merely result in larger diversity in parental populations. Drift-based (like current likelihood-based) methods seem also to better deal with short divergence time between parental populations (*e.g.*, 200 generations instead of 5000) than does our ABC procedure when the admixture is recent (results not shown). However, this advantage is valid only for recent admixtures (*e.g.*, <50 generations). Another advantage of the present ABC approach is its ability to correctly estimate other parameters of the admixture model, such as divergence and admixture

times. These parameters are often as important as the admixture coefficient itself. The better performance of our approach is probably linked to the fact that we are using information not specifically handled by Wang's ML method, such as information on patterns of LD and mutations, as well as range of allele size. Moreover, our ABC approach allows us to explicitly include information on partial linkage between markers, so that, in contrast to Wang's ML method, accurate confidence intervals are also obtained in this case.

While the admixture model analyzed here (with a hybrid population and two isolated parental populations at mutation-drift equilibrium) corresponds to the standard model assumed by most methods of estimation of admixture coefficients (*e.g.*, LONG 1991; BERTORELLE and EXCOFFIER 1998; WANG 2003; CHOISY *et al.* 2004), real models of admixture may be much more complex. They may indeed involve: (i) more than two source populations (DUPANLOUP and BERTORELLE 2001); (ii) some regular (and thus not instantaneous) admixture events over relatively long periods (reviewed in CHAKRABORTY 1986); (iii) subdivided source populations, so that the actual parental population is only partially sampled; and (iv) parental population(s) that are not at mutation-drift equilibrium, due to population size fluctuations or introgression event(s) in a more or less recent past. The ABC approach has the potential to assess the effect of such deviations from the standard admixture model on parameter estimations since the ratio of acceptance under two alternative models approximates the Bayes factor (*e.g.*, ESTOUP *et al.* 2004; PRITCHARD *et al.* 1999). Such quantitative model comparisons could be particularly useful in the present evolutionary context to assess the likelihood of different deviations from the standard admixture model and hence learn more about the admixture process that produced the observed data set and potentially consider more realistic models for parameter estimation.

A general feature of the ABC methods that should be underlined here is their ability to assess their performance at almost no extra computation cost. Other estimation methods generally require a validation step, which includes the time-consuming analysis of independently produced simulated data sets (*e.g.*, CHOISY *et al.* 2004), whereas this is intrinsic in the ABC approach (*cf.* Figure 2). As a matter of fact, the same ABC process used to build the reference table can be derived to produce test data sets with known values of parameters. The same rejection and regression steps can then be applied to these data sets to produce estimates of parameters that can be compared to their known true values. It is therefore relatively quick and easy to evaluate the performance of the method for any subset of the parameter space under a given model. The applicability of the ABC method to particular cases should, however, depend on available computer power, as a few days of computing time are necessary to obtain a large number

of simulated summary statistics from which the estimation procedure proceeds (*e.g.*, 10^6 iterations). However, reasonable point estimates can be obtained using much fewer simulations and hence shorter computation times (*e.g.*, 10^5 iterations). It seems reasonable to anticipate that progress in simulation algorithms and higher computing power will be available in future years, promoting the ABC method as the method of choice for analyzing complex evolutionary scenarios and, more specifically in the context of the present study, for old admixture models in which mutation cannot be neglected or when nonindependent markers are available.

We are grateful to Lounès Chikhi and Mark Beaumont for their comments on the manuscript. L.E. was supported by Swiss National Science Foundation grant 3100A0-100800, as well as a grant from the Institut de la Recherche Agronomique during his 2004 sabbatical visit at the Centre de Biologie et de Gestion des Populations. This study was also partially supported by a grant from the French Bureau des Ressources Génétiques.

LITERATURE CITED

- ANDERSON, E. C., and E. A. THOMPSON, 2002 A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* **160**: 1217–1229.
- ANDERSON, E. C., E. G. WILLIAMSON and E. A. THOMPSON, 2000 Monte Carlo evaluation of the likelihood for N_e from temporally spaced samples. *Genetics* **156**: 2109–2118.
- BARTON, N. H., 2001 The role of hybridization in evolution. *Mol. Ecol.* **10**: 551–568.
- BEAUMONT, M. A., W. ZHANG and D. J. BALDING, 2002 Approximate Bayesian computation in population genetics. *Genetics* **162**: 2025–2035.
- BERTHIER, P., M. A. BEAUMONT, J. M. CORNUET and G. LUIKART, 2002 Likelihood-based estimation of the effective population size using temporal changes in allele frequencies: a genealogical approach. *Genetics* **160**: 741–751.
- BERTORELLE, G., and L. EXCOFFIER, 1998 Inferring admixture proportions from molecular data. *Mol. Biol. Evol.* **15**: 1298–1311.
- CHAKRABORTY, R., 1986 Gene admixture in human populations: models and predictions. *Yearb. Phys. Anthropol.* **29**: 1–43.
- CHAKRABORTY, R., and K. M. WEISS, 1988 Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc. Natl. Acad. Sci. USA* **85**: 9119–9123.
- CHIKHI, L., M. W. BRUFORD and M. A. BEAUMONT, 2001 Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics* **158**: 1347–1362.
- CHOISY, M., P. FRANCK and J. M. CORNUET, 2004 Estimating admixture proportions with microsatellites: comparison of methods based on simulated data. *Mol. Ecol.* **13**: 955–968.
- DAWSON, K. J., and K. BELKHIR, 2001 A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genet. Res.* **78**: 59–77.
- DUPANLOUP, I., and G. BERTORELLE, 2001 Inferring admixture proportions from molecular data: extension to any number of parental populations. *Mol. Biol. Evol.* **18**: 672–675.
- DUPANLOUP, I., G. BERTORELLE, L. CHIKHI and G. BARBUJANI, 2004 Estimating the impact of prehistoric admixture on the genome of Europeans. *Mol. Biol. Evol.* **21**: 1361–1372.
- ELLEGREN, H., 2004 Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* **5**: 435–445.
- ESTOUP, A., M. BEAUMONT, F. SENNETOT, 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.
- ESTOUP, A., and S. M. CLEGG, 2003 Bayesian inferences on the recent island colonization history by the bird *Zosterops lateralis lateralis*. *Mol. Ecol.* **12**: 657–674.

- ESTOUP, A., L. GARNERY, M. SOLIGNAC and J. M. CORNUET, 1995 Microsatellite variation in honey bee (*Apis mellifera* L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models. *Genetics* **140**: 679–695.
- ESTOUP, A., C. R. LARGIADER, J. M. CORNUET, K. GHARBI, P. PRESA *et al.*, 2000 Juxtaposed microsatellite systems as diagnostic markers for admixture: an empirical evaluation with brown trout (*Salmo trutta*) as model organism. *Mol. Ecol.* **9**: 1873–1886.
- ESTOUP, A., I. J. WILSON, C. SULLIVAN, J. M. CORNUET and C. MORITZ, 2001 Inferring population history from microsatellite and enzyme data in serially introduced cane toads, *Bufo marinus*. *Genetics* **159**: 1671–1687.
- ESTOUP, A., P. JARNE and J. M. CORNUET, 2002 Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Mol. Ecol.* **11**: 1591–1604.
- FALUSH, D., M. STEPHENS and J. K. PRITCHARD, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.
- FELDMAN, M. W., A. BERGMAN, D. D. POLLOCK and D. B. GOLDSTEIN, 1997 Microsatellite genetic distances with range constraints: analytic description and problems of estimation. *Genetics* **145**: 207–216.
- FRANCK, P., L. GARNERY, G. CELEBRANO, M. SOLIGNAC and J. M. CORNUET, 2000 Hybrid origins of honeybees from Italy (*Apis mellifera ligustica*) and Sicily (*A. m. sicula*). *Mol. Ecol.* **9**: 907–921.
- FU, Y. X., and W. H. LI, 1997 Estimating the age of the common ancestor of a sample of DNA sequences. *Mol. Biol. Evol.* **14**: 195–199.
- GAGGIOTTI, O. E., F. JONES, W. M. LEE, W. AMOS, J. HARWOOD *et al.*, 2002 Patterns of colonization in a metapopulation of grey seals. *Nature* **416**: 424–427.
- GAGGIOTTI, O. E., S. P. BROOKS, W. AMOS and J. HARWOOD, 2004 Combining demographic, environmental and genetic data to test hypotheses about colonization events in metapopulations. *Mol. Ecol.* **13**: 811–825.
- GARNERY, L., J. M. CORNUET and M. SOLIGNAC, 1992 Evolutionary history of the honey bee *Apis mellifera* inferred from mitochondrial DNA analysis. *Mol. Ecol.* **1**: 145–154.
- GARZA, J. C., and E. G. WILLIAMSON, 2001 Detection of reduction in population size using data from microsatellite loci. *Mol. Ecol.* **10**: 305–318.
- GARZA, J. C., M. SLATKIN and N. B. FREIMER, 1995 Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. *Mol. Biol. Evol.* **12**: 594–603.
- GOLDSTEIN, D. B., and D. D. POLLOCK, 1997 Launching microsatellites: a review of mutation processes and method of phylogenetic inference. *J. Hered.* **88**: 335–342.
- GOLDSTEIN, D. B., A. RUIZ LINARES, L. L. CAVALLI-SFORZA and M. W. FELDMAN, 1995 Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc. Natl. Acad. Sci. USA* **92**: 6723–6727.
- HUDSON, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**: 337–338.
- 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.
- LONG, J. C., 1991 The genetic structure of admixed populations. *Genetics* **127**: 417–428.
- MARJORAM, P., J. MOLITOR, V. PLAGNOL and S. TAVARE, 2003 Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **100**: 15324–15328.
- POLLOCK, D. D., A. BERGMAN, M. W. FELDMAN and D. B. GOLDSTEIN, 1999 Microsatellite behavior with range constraints: parameter estimation and improved distances for use in phylogenetic reconstruction. *Theor. Popul. Biol.* **53**: 256–271.
- PRITCHARD, J., M. SEIELSTAD, A. PEREZ-LEZAUN and M. FELDMAN, 1999 Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.* **16**: 1791–1798.
- PRITCHARD, J. K., M. STEPHENS and P. DONNELLY, 2000 Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.
- ROBERTS, D., and R. HIORNS, 1965 Methods of analysis of the genetic composition of a hybrid population. *Hum. Biol.* **37**: 38–43.
- RUTTNER, F., 1988 *Biogeography and Taxonomy of Honeybees*. Springer-Verlag, Berlin.
- TAVARÉ, S., D. BALDING, R. C. GRIFFITHS and P. DONNELLY, 1997 Inferring coalescence times from DNA sequence data. *Genetics* **145**: 505–518.
- WANG, J., 2003 Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics* **164**: 747–765.
- WEIR, B. S., and C. C. COCKERHAM, 1984 Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.
- WILLIAMSON, E. G., and M. SLATKIN, 1999 Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics* **152**: 755–761.
- WILSON, I. J., and D. J. BALDING, 1998 Genealogical inference from microsatellite data. *Genetics* **150**: 499–510.
- ZHIVOTOVSKY, L. A., M. W. FELDMAN and S. A. GRISHECHKIN, 1997 Biased mutations and microsatellite variation. *Mol. Biol. Evol.* **14**: 926–933.

Communicating editor: M. VEUILLE