Molecular Ecology (2004) **13**, 955–968 doi: 10.1111/j.1365-294X.2004.02107.x

Estimating admixture proportions with microsatellites: comparison of methods based on simulated data

M. CHOISY,* P. FRANCK† and J.-M. CORNUET‡
Centre d'Etude sur le Polymorphisme des Micro-organismes, UMR CNRS-IRD 9926, Montpellier, France, †Unité d'Ecologie des invertébrés, UMR INRA-UAPV 406, Avignon, France, ‡Centre de Biologie et de Gestion des Populations, UMR INRA-IRD-CIRAD-ENSAM 1062, Campus International de Baillarguet, CS 30016 Montferrier-sur-Lez, 34988 Saint-Gély du-Fesc Cedex, France

Abstract

Several methods have been developed to estimate the parental contributions in the genetic pool of an admixed population. Some pair-comparisons have been performed on real data but, to date, no systematic comparison of a large number of methods has been attempted. In this study, we performed a simulated data-based comparison of six of the most cited methods in the literature of the last 20 years. Five of these methods use allele frequencies and differ in the statistical treatment of the data. The last one also considers the degree of molecular divergence by estimating the coalescence times. Comparisons are based on the frequency at which the method can be applied, the bias and the mean square error of the estimation, and the frequency at which the true value is within the confidence interval. Eventually, each method was applied to a real data set of variously introgressed honeybee populations. In optimal conditions (highly differentiated parental populations, recent hybridization event), all methods perform equally well. When conditions are not optimal, the methods perform differently, but no method is always better or worse than all others. Some guidelines are given for the choice of the method.

Keywords: admixture, *Apis*, coalescent, simulation

Received 31 July 2003; revision received 2 December 2003; accepted 2 December 2003

Introduction

Populations can evolve more or less independently. This depends on the number of individual migrations occurring among them and determines how genetically distinct the populations are. In some instances, populations which have evolved separately for a long time come into contact producing a hybrid population, the genes of which are a mixture of the parental populations (Bernstein 1931). More generally, we call admixture or introgression the incorporation of genes from one genetically distinct population into another (Futuyama 1998) and this process appears to have been quite common in the evolutionary history of most species (Szymura & Barton 1986; Abernethy 1994; Paetkau & Strobeck 1994; Paetkau *et al*. 1995; De Wayne Shoemaker *et al*. 1996; Goostrey *et al*. 1998; Poteaux *et al*. 1998; Goodman *et al*. 1999; Reich *et al*. 1999). Human populations have been no exception and it is worth noting that, historically, the pioneering work on introgression was developed on our species, and particularly to estimate admixture in black North Americans (Bernstein 1931; Glass & Li 1953; Glass 1955; Roberts 1955; Reed 1969).

Even if the first interest in admixture was mainly fundamental, many applied issues have recently taken advantage of the knowledge of the introgressive process. For example, in medicine, studying admixture can provide information on the inheritance of complex genetic diseases (Chakraborty & Weiss 1988; McKeigue 1997). In biological conservation, studying introgression can help to supervise the efficiency of restocking (Giuffra *et al*. 1996) or, by contrast, evaluate the risk of genetic 'pollution' of native populations by the accidental introduction of alien individuals (Cornuet *et al*. 1986; Wayne & Jenks 1991; Gotteli *et al*. 1994; Ellstrand *et al*. 1999).

The study of introgression can be limited to the identification of parental population(s) (Glass 1955; Haig *et al*.

Correspondence: Jean-Marie Cornuet.
E-mail: jmcornuet@ensam.inra.fr

1997; Nielsen *et al*. 1997; Reed *et al*. 1997). It can also be extended to the evaluation of (i) the number and duration of the admixture events (Glass & Li 1953; Roberts 1955; Roberts & Hiorns 1962) and/or (ii) the different contributions of parental genomes to the admixed population (Roberts & Hiorns 1965; Szathmary & Reed 1978; Parra *et al*. 1998). The last point is most informative for characterization of the current state of the hybrid population (Chakraborty 1986) and is the one we are interested in here. Estimation of the genetic admixture proportion can be performed at the population level (Hanis *et al*. 1986) or at the individual level (Reed 1969; McLean & Workman 1973). In the former case, we are interested in the fraction of genes in the admixed population that come from one or other of the parental populations. In the latter case, it is the proportion of loci in the genome of a single individual that come from a parental population which is under scrutiny. The most commonly used level is the population one (Chakraborty 1986) and this is the one considered here.

Different methods have been developed to estimate the genetic admixture proportion from genotypic data. Most are based on the analysis of frequency data but differ in the way allelic frequencies are used in computations. For instance, one method (Chakraborty *et al*. 1991) considers only private alleles, i.e. those that are found in one parental population and not in the other. Another method (Chakraborty 1975) is based on average identity coefficients (within and between populations). Recently, a method based on allele coalescence times rather than allelic frequencies has been published (Bertorelle & Excoffier 1998). In any case, most of these methods have been applied to experimental data sets. Some have been compared with one or other previously published method also based on real data. However, no systematic comparison of these published methods has been attempted so far. To be systematic, the comparison must be made on a wide set of well-defined situations and the only way to achieve this is to use simulated data. Here we propose to compare six methods that seem to be the most frequently used in the literature from the last 20 years. All the data concern the idealized case in which two parental populations, having diverged for a certain amount of time, gave birth instantaneously to a single hybrid population some time ago. In the simulations, we explored 648 situations that combine the variation of six parameters. These parameters concerned either the admixture scenario (amount of divergence of parental populations, time since the admixture event, proportions of parental genetic pools in the hybrid population) or the genotyping effort (sample sizes, number of loci). All simulations were based on the coalescent process and were designed essentially for microsatellite loci. Eventually, each method was applied to a microsatellite data set collected on the introgression zone between the French (*Apis mellifera mellifera*) and Italian (*A. m. ligustica*) subspecies of honeybees.

Materials and methods

The hybridization model

We considered the simple hybridization model presented in Fig. 1. An ancestral population P_0 splits into two parental populations P_1 and P_2, which evolve separately for g_0 generations and give birth instantaneously to a single hybrid population H. A proportion λ_H of the genetic pool of the hybrid population comes from population P_1 and a proportion $(1 - \lambda_H)$ from population P_2. The parameter λ_H is the admixture proportion at the time of the hybridization event. The three populations then evolve separately for g_1 generations until the present time. Evolutionary forces such as genetic drift or mutation can modify allele frequencies (selection is not considered here). We thus call λ_P the admixture proportion at the present time, i.e. λ_P is the current proportion of genes descending from P_1 in the hybrid population sample.

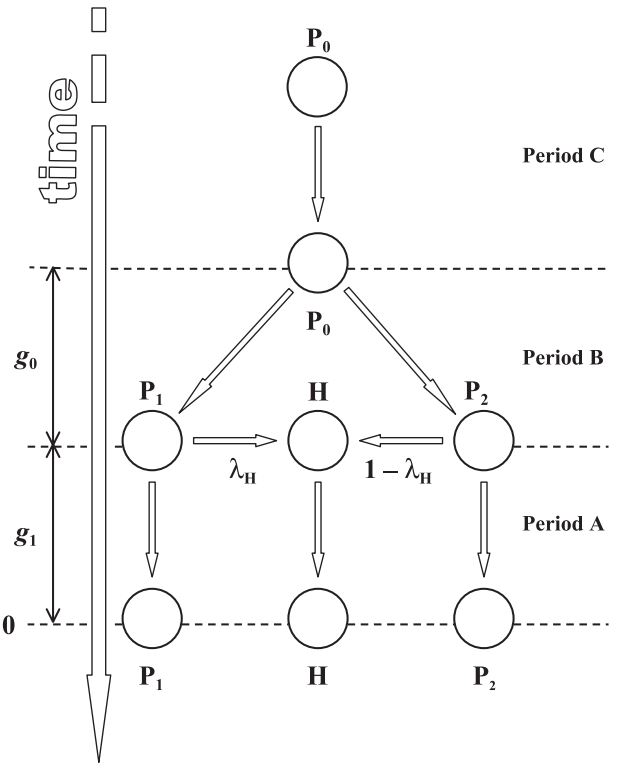

Fig. 1 The admixture model. An ancestral population P_0 splits into two parental populations P_1 and P_2, which evolve separately for g_0 generations (period B). These two populations then mixed instantaneously producing a hybrid population H. The three populations evolve separately for g_1 generations until the present time (period A). The introgressive rate λ_H is the proportion of genes coming from population P_1.

Data simulation

The data were simulated in two steps. The first consisted of building a coalescent tree of genes using the coalescent process in isolated populations following Simonsen *et al*. (1995). As usual, the coalescent tree was built backward in time from the present to the time of the most recent common ancestor (MRCA). At the generation at which hybridization occurred, the ancestral lines of the hybrid population sample of genes were added to those of population P_1 with probability λ_H (and to those of population P_2 with probability $1 - \lambda_H$, respectively). Note that we kept track of the origin (population P_1 or P_2) of sampled genes in the hybrid population, in order to compute the value of λ_P. Once the tree had been built, the second step consisted in adding the mutation events. The number of mutations was simulated along each branch of the tree according to a Poisson distribution with parameter $L\mu$ (L = length of the branch in generations and μ = mutation rate). The mutation rate was allowed to vary among loci following a gamma distribution with an average of 6.2×10^{-4} and a variance of 1.922×10^{-7} (Estoup *et al*. 2001). At each mutation event, a bounded (41 states) and symmetric generalized stepwise mutation model (GSM) was applied. In this model, the step size followed a geometric distribution the variance of which was drawn from an exponential with average equal to 0.36 (Estoup *et al*. 2001).

Genotype data were simulated independently for each locus. Monomorphic loci were immediately discarded so that, in the following, the number of loci always refers to the number of polymorphic (i.e. useful) loci.

Description of methods

The first four methods use allele frequencies and are based on Bernstein (1931)'s equation which expresses the frequency $p_i^{(h)}$ of allele i in the hybrid population as a linear combination of the allele frequencies $p_i^{(1)}$ and $p_i^{(2)}$ in the two parental populations:

$$p_i^{(h)} = \lambda p_i^{(1)} + (1 - \lambda) p_i^{(2)}. \quad (1)$$

Note that this equation is true at the time of the hybridization event, i.e. with $\lambda = \lambda_H$. More or less implicitly, the four methods assume that allelic frequencies are stable (no drift, no mutation) such that the admixture proportions (λ_P) estimated on current allele frequencies are representative of the admixture (λ_H) that occurred at hybridization time. In describing the method principles, we will then drop the subscripts P or H and keep only the symbol λ.

Method of gene identity (GI). Starting from eqn 1, it is easy to deduce a similar relationship with gene identities (Chakraborty 1975, 1986). Let J_{11}, J_{12} and J_{1h} be the arithmetic means over all loci of the probabilities of gene identity within population P_1, and between populations P_1 and P_2 and populations P_1 and H. Then eqn 1 can be written in the following form:

$$J_{1h} = \lambda J_{11} + (1 - \lambda) J_{12}$$

from which we get $\lambda = (J_{1h} - J_{12})/(J_{11} - J_{12})$

Thus, estimations of gene identity coefficients lead to a straightforward estimation of the genetic admixture proportions λ. Note that a symmetric formula can be obtained by permuting populations: $J_{2h} = \lambda J_{12} + (1 - \lambda) J_{22}$, leading to $\lambda = (J_{2h} - J_{22})/(J_{12} - J_{22})$.

When possible, both estimates were combined by taking the arithmetic average. The original article (Chakraborty 1975) provided only a standard error of the estimate. Although we could have built a confidence interval from this standard error, we preferred to compute it through a bootstrap procedure over loci (1000 repetitions).

Method of least square regression (LS). For any allele, eqn 1 can also be written as:

$$p_i^{(h)} = \lambda (p_i^{(1)} - p_i^{(2)}) + p_i^{(2)} \quad (2)$$

in which the admixture proportion λ can be regarded as the slope of a linear regression. The use of a least square regression to estimate the admixture proportions was first suggested by Roberts & Hiorns (1962, 1965) and successively improved by Elston (1971), Long & Smouse (1983) and Long (1991).

Over I independent alleles (i.e. total number of alleles minus total number of loci), the slope λ can be estimated by the classical least square regression formula:

$$\hat{\lambda} = (\boldsymbol{x}^T \boldsymbol{V}^{-1} \boldsymbol{x})^{-1} \boldsymbol{x}^T \boldsymbol{V}^{-1} \boldsymbol{y} \quad (3)$$

in which $\boldsymbol{x}^T$ is the transposed of the vector x of elements $\{x_i\} = p_i^{(1)} - p_i^{(2)}$, $\boldsymbol{y}$ is the vector of elements $\{y_i\} = p_i^{(h)} - p_i^{(2)}$, $\boldsymbol{V}$ is a variance–covariance matrix with diagonal elements $\{V_{ii}\} = \pi_i(1 - \pi_i)/n^{(h)}$ and off-diagonal elements $\{V_{ij}\} = -\pi_i \pi_j / n^{(h)}$ or 0 whether alleles i and j belong to the same locus or not, π_i being the expectation of $p_i^{(h)}$ approximated by $\lambda(p_i^{(1)} - p_i^{(2)}) + p_i^{(2)}$ (eqn 2) and $n^{(h)}$ the number of sampled genes in the hybrid population at the locus to which belong alleles i and j. Because λ appears in both sides of eqn 2, the solution is obtained through successive iterations (Long 1991). A confidence interval was built using the standard error of the estimate (eqn 4 in Long 1991). The correction for small samples in the computation of the matrix $\boldsymbol{V}$ (eqn 16 in Long 1991) was systematically applied (Cavalli-Sforza & Bodmer 1971; Long & Smouse 1983).

Method of Madansky's regression on private alleles (PA). Neel (1973) introduced the term 'private allele' to name

alleles which are present in only one group of individuals. The interest of private alleles for estimating λ is essentially to simplify eqn 1 which reduces to $p_i^{(h)} = \lambda p_i^{(1)}$. Various computational approaches have been used (Szathmary & Reed 1978; Byard *et al*. 1985; Chakraborty 1986; Williams *et al*. 1986). We focused on Chakraborty *et al*.'s (1991) approach based on Madansky's (1959) estimator which would have a better behaviour that the least square regression when covariates (private allele frequencies in parental and hybrid populations) are both estimated with potentially substantial errors. A confidence interval was built according to eqn 13 of Chakraborty *et al*. (1991). Whenever possible, estimates were obtained by considering successively the alleles private to P_1 and the alleles private to P_2. Both estimates were combined, weighted by the inverse of their respective variances:

$$\hat{\lambda} = \frac{\dfrac{\hat{\lambda}_1}{\sigma^2(\hat{\lambda}_1)} + \dfrac{1-\hat{\lambda}_2}{\sigma^2(\hat{\lambda}_2)}}{\dfrac{1}{\sigma^2(\hat{\lambda}_1)} + \dfrac{1}{\sigma^2(\hat{\lambda}_2)}} \quad \text{and} \quad \hat{\sigma}(\hat{\lambda}) = \frac{\dfrac{1}{\sigma(\hat{\lambda}_1)} + \dfrac{1}{\sigma(\hat{\lambda}_2)}}{\dfrac{1}{\sigma^2(\hat{\lambda}_1)} + \dfrac{1}{\sigma^2(\hat{\lambda}_2)}}$$

where $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are the estimates of λ when considering alleles private to P_1 and P_2, respectively.

Method of the maximum likelihood of hybrid genotypes (ML). Like the least square regression, this method was one of the very first used to estimate the genetic admixture proportions (Krieger *et al*. 1965; Roberts & Hiorns 1965; Elston 1971). Although several methodological improvements have been proposed (see the reviews of Korey 1978 and Chakraborty 1986), the basic idea of this method is quite simple. The likelihood of the hybrid sample individual genotypes is expressed as a function of the parental population allele frequencies and the admixture proportion. The estimate of the admixture proportion is then the parameter value, which maximizes the likelihood. For instance, consider two alleles *i* and *j* of a given locus *k* of an individual *l* drawn from the hybrid population sample. Eqn 1 gives $p_i^{(h)} = \lambda p_i^{(1)} + (1-\lambda)p_i^{(2)}$ and $p_j^{(h)} = \lambda p_j^{(1)} + (1-\lambda)p_j^{(2)}$. Assuming Hardy–Weinberg equilibrium, the likelihood of genotype g_{kl} is equal to $[\lambda p_i^{(1)} + (1-\lambda)p_i^{(2)}]^2$ if g_{kl} is homozygous for allele *i* and equal to $2[\lambda p_i^{(1)} + (1-\lambda)p_i^{(2)}][\lambda p_j^{(1)} + (1-\lambda)p_j^{(2)}]$ if g_{kl} is heterozygous for alleles *i* and *j*. Considering only genetically independent loci, the likelihood of a multilocus genotype is the product of the likelihood at each locus and the likelihood of the total hybrid population sample is obtained by multiplying likelihoods over individuals. The maximum likelihood estimation can be solved analytically (Krieger *et al*. 1965; Roberts & Hiorns 1965; Elston 1971) or numerically when the analytical computation becomes too complex (Destro-Bisol *et al*. 1999). The second option was taken here. A confidence interval of the value of $\hat{\lambda}$ was computed using a bootstrap procedure over loci (1000 repetitions).

Method of coalescence times (CT). More recently, Bertorelle & Excoffier (1998) developed a new method based on a coalescent approach which explicitly takes into account molecular information as well as gene frequencies. The idea is to consider the mean coalescence time for a pair of genes, one gene sampled in the hybrid population and the other in one parental population. For any such pair, ancestral lines of the two genes can coalesce in two ways. Either the 'hybrid' gene ancestral line switches to the same parental population as the second gene of the pair and both can coalesce from the hybridization event (looking backward in time). Or it switches to the other parental population and they will coalesce only in the ancestral (P_0) population (Fig. 1). Depending on which parental population is considered, the probability of the first event is simply λ or 1 – λ, the reverse occurring with probabilities 1 – λ and λ, respectively. Combining these relationships, Bertorelle & Excoffier (1998) provide several estimators among which one performs better (m_Y, eqn 2 in Bertorelle & Excoffier 1998). Only the latter was used in our computations. We also followed Bertorelle and Excoffier's suggestion of considering the average squared difference in allele size as an estimate of coalescence time for microsatellite data. Note that this assumes a strict single-step stepwise mutation model (SMM). A confidence interval of the value of $\hat{\lambda}$ was computed by a *bootstrap* procedure over loci (1000 repetitions).

Monte Carlo Markov chain method (MC). Chikhi *et al*. (2001) proposed a different approach also based on coalescence theory. It combines importance sampling and Monte Carlo Markov chain methodologies. Given the evolutionary scenario of Fig. 1 and fixed values of relevant parameters (effective population sizes, admixture coefficient and time of hybridization), an importance sampling (IS) scheme (inspired from Griffith & Tavaré 1994) is designed to estimate the combined probability of the three observed gene samples. It performs a backward simulation of the coalescence tree of lineages starting from the sampled genes and ending just before (forward in time) the hybridization step. The IS scheme computation is then completed by considering that the vectors of allelic occurrences in the parental populations are draws from a multinomial–Dirichlet. Initially, a uniform multinomial–Dirichlet with coefficients all equal to one was applied (as it is in the available software). However, we abandoned this prior distribution parameterization because, at marginal value of λ, it produced a significant bias in conditions whereas all other methods were practically unbiaised: recent admixture, highly differentiated parental populations. To reduce the bias, we decided to take the Dirichlet coefficients equal to the inverse of the total number of alleles, as in Rannala & Mountain (1997). In the process of simulating backward in time the genealogy of sampled genes, lineages are

randomly distributed between the two parental populations in the IS scheme used in the current version of the available software (LEA at http://www.pge.cnrs-gif.fr/bioinfo/lea/). This can be inefficient when parental populations are highly differentiated. In our study, we introduced an IS step, which largely improves the efficiency of the algorithm (manuscript in preparation). The IS scheme is embedded in a MCMC setting allowing to obtain samples from posterior distributions of the relevant demographic parameters (time of admixture scaled by effective populations sizes and admixture proportion). The aforementioned improvement of the IS scheme resulted in a better mixing, thus reducing the duration of the analysis. All runs were performed with a burn-in period of 1000 iterations and 1000 values of each parameter were sampled with a thinning of 10 iterations. We chose the median as a point estimate and use 5 and 95% quantiles as limits of the confidence interval.

Comparing the methods

Simulated data were produced under different values of four biological and two experimental parameters. The biological parameters concern the effective size (N_e) in number of diploid individuals of the three populations (for the sake of simplicity, we chose to take the same N_e for the three populations), the amount of divergence in number of generations of parental populations (g_0), the time in number of generations since the introgression event (g_1) and the admixture proportions at the time of the introgression event (λ_H). The experimental parameters (sample sizes and number of loci) convey the genotyping effort. The values taken by the six parameters in our study are presented in Table 1. Eventually, 648 combinations of parameters were used and for each of them 100 data sets were simulated and analysed using the six methods.

Each method was evaluated according to five criteria: (i) the applicability, expressed as the percentage of times the method successfully produced a complete and adequate result. By complete we mean that the result includes an estimate of λ and a confidence interval of this estimate and by adequate, we mean that the estimate of λ is expected to lie between 0 and 1. As we see later, methods are not always applicable; (ii) the mean bias computed as the mean difference (over the 100 simulated data sets) between the estimate and the real value; (iii) the mean square error (MSE) of λ estimates, a usual criterion for precision; (iv) and (v) the confidence interval (CI) success expressed as the proportion of times the confidence interval contained the true value of λ_H and λ_P, respectively. Because a 90% confidence interval was computed for all methods, the expected value of the last two criteria was 90%.

Table 1 Values taken by the parameters in the simulations. Biological parameters include the effective size (N_e) in number of diploid individuals, divergence time (in number of generations) of parental populations (g_0), the time (in number of generations) since the introgression event (g_1) and the admixture proportion at the time of the introgression event (λ_H). Experimental parameters are the number of loci (n_{loc}) and the sample size (ss). All 648 combinations of the six values were simulated 100 times

N_e	g_0	g_1	λ_H	n_{loc}	ss
10	100	1	0.1	5	10
100	10 000	10	0.5	10	30
1000	1 000 000	100		20	90
		1000			

Results

The results of the analysis of the 64 800 data sets are tentatively summarized in Figs 2 and 3. Figure 2 provides performances of the methods as functions of the biological parameters, averaged over the number of loci, sample sizes and repetitions (each dot is the average of 900 data sets). In Fig. 3, the performances are presented as functions of the genotyping effort, averaged over the biological parameters and repetitions (each dot representing the average of 3600 data sets).

Effects of biological parameters

A general and positive outcome is that, under optimal conditions (recent hybridization and highly differentiated parental populations), all five methods were applicable and provided rather unbiased and precise estimates of admixture proportions. However, when departing from these conditions, substantial bias and losses of precision sometimes occurred and consequently, confidence intervals partly lost their credibility. We also observed that the method PA was often inapplicable, especially when the effective size was low to medium and when the hybridization event was not recent (Fig. 2A and B, first column). We analyse below the main effects of the various biological parameters.

Parameter λ_H had a visible impact on all criteria, the least affected being the applicability. Figure 2 illustrates this by summarizing information for two 'extreme' values of λ_H ($\lambda_H = 0.1$ and $\lambda_H = 0.5$, hereafter referred to marginal and central values, respectively), in Fig. 2A and B, respectively. The mean bias was essentially positive for $\lambda_H = 0.1$, whereas it was negligible for $\lambda_H = 0.5$. At marginal values of λ_H (i.e. at $\lambda_H = 0.1$), a low differentiation (e.g. $g_0 = 100$) led to a substantial positive bias that grew with time since hybridization (g_1) for all six methods (Fig. 2A). As a consequence of these large biases, there was an important decrease in precision (large MSE) and in λ CI successes (Fig. 2A). For example, with a 1000-generation-old

A

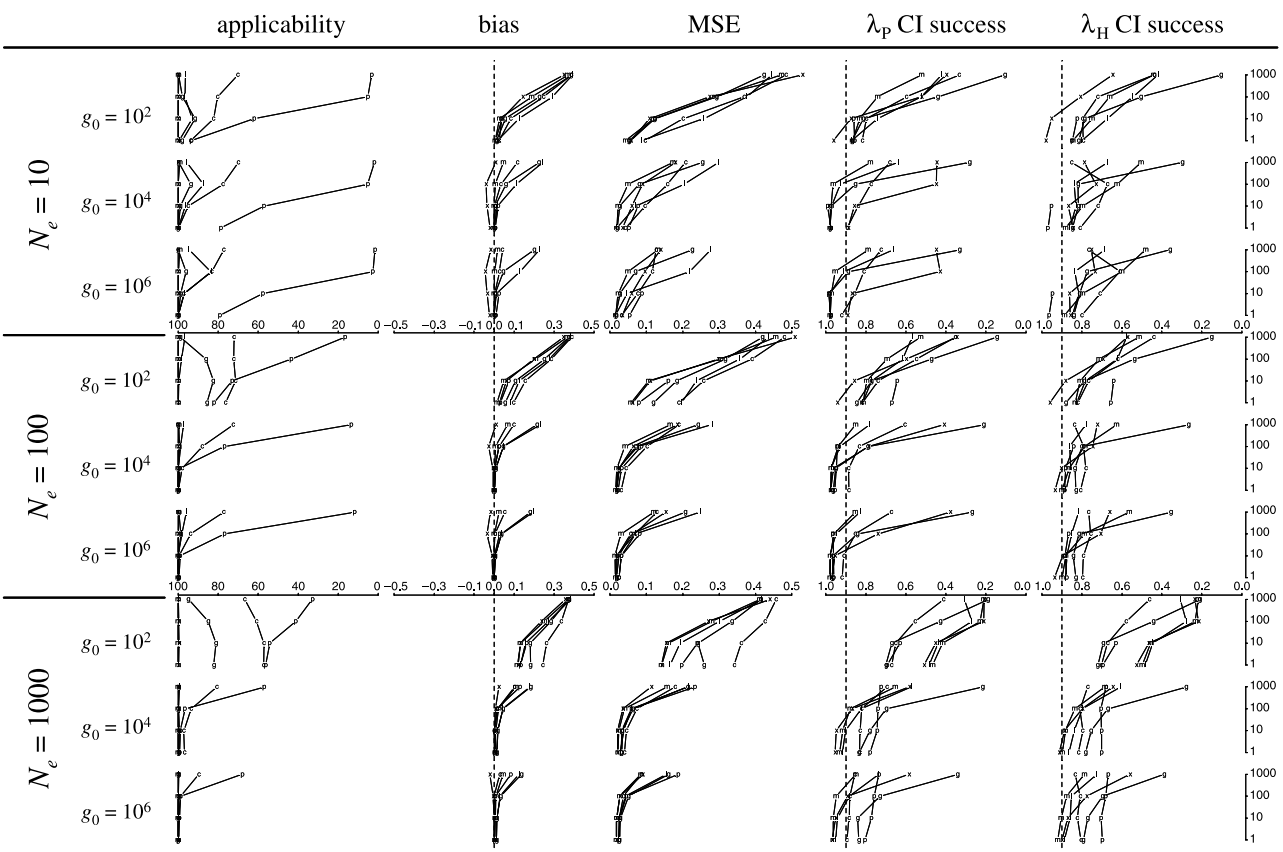

Fig. 2 Effects of biological conditions. The five criteria of appreciation for the five methods are presented for N_e = 10, 10^2 and 10^3 individuals, g_0 = 10^2, 10^4 and 10^6 generations and g_1 = 1, 10, 10^2 and 10^3 generations (from bottom to top on each curve). See text for details on each criteria of appreciation. Parameter λ_H is the admixture proportion at the time of the hybridization and λ_P is the current proportion of genes descending from P_1 in the hybrid population sample (modified from λ_H by genetic drift). Each point of the curves is a mean over the nine combinations of the three other parameters which are the introgressive rate and the numbers of loci and individuals studied. The methods are the gene identity (g), least square (l), private alleles (p), maximum likelihood (m), coalescence times (c), and Monte Carlo Markov chain (x). The dashed vertical line in the second column corresponds to a bias equal to 0 and the dashed vertical line in the last column corresponds to the 90% which is the expected value according to the chosen type-I error. Each point for which applicability was < 50% was not represented on the graph. A: λ_H = 0.1, B: λ_H = 0.5. Coloured versions of Fig. 2A and B are available at www.montpellier.inra/CBGP/progrechCornuet.htm.

hybridization, the GI method confidence interval bracketed the true value on < 20% of occasions.

We used three levels of differentiation of parental populations (low: g_0 = 10^2 generations, medium: g_0 = 10^4 generations and high: g_0 = 10^6 generations). At low differentiation, performances were reduced whatever the criterion. This was especially true when λ_H was marginal. For instance, when N_e = 10^3, the MSE was much larger and the λ CI successes farther from the expected 0.9 than when differentiation was higher. However, performances were rather similar when comparing medium and high levels of differentiation. Performances increase with the differentiation of the parental population but there is a differentiation threshold over which performances reach a plateau. In our conditions, the threshold is reached in about 10^4 generations after divergence.

When the hybridization event got older, the precision and hence the λ CI successes generally decreased, but not necessarily the applicability or the relative bias (Fig. 2B). For some methods, the performances ceased to decrease or even re-improved when the hybridization age increased over 10^2 generations. Note that it is observed essentially when parental populations diverged recently (g_0 = 10^2 generations). Actually, in the latter case, the time spent since the hybridization (g_1) adds to the initial divergence time (g_0) for differentiating parental populations. In the extreme case (g_1 = 10^3 generations), parental populations had a divergence time 11 ([100 + 1000]/100) times longer

B

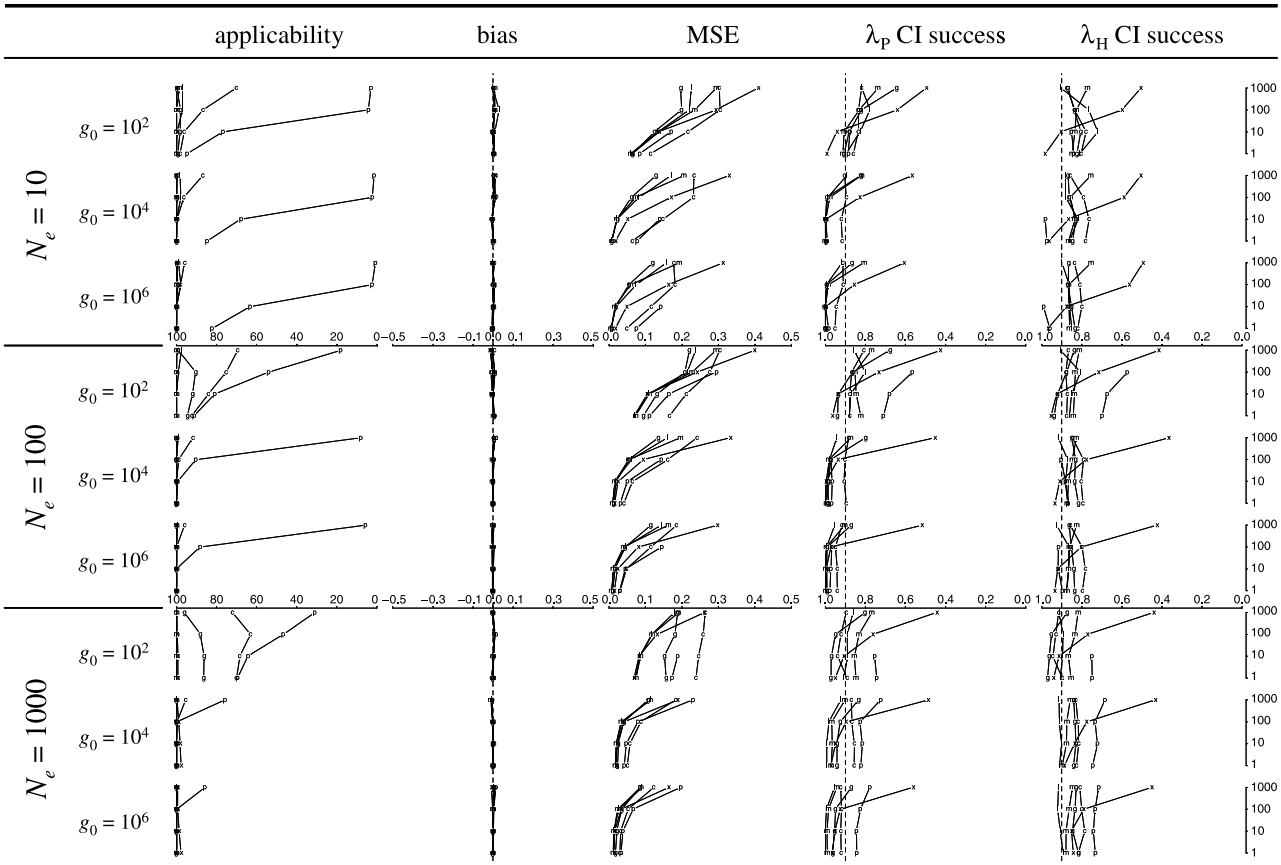

Fig. 2 *Continued*

than if they were sampled at the generation following hybridization. However, this effect was not identical for all methods and all criteria.

The effective population size had also some impact on the performance of the methods, especially when differentiation was low. At large effective sizes ($N_e = 10^3$), low differentiation ($g_0 = 10^2$ generations) and marginal λ_H, there was a substantial (positive) bias for all methods and the precision (MSE) and CI successes decreased dramatically. This was observed even for most recent hybridizations (one generation old).

Effects of the genotyping effort

Increasing the genotyping effort generally improved the performances of the methods (Fig. 3A and B for $\lambda_H = 0.1$ and 0.5, respectively). However, the improvement is very limited when genotyping more than 30 individuals per population. Relatively unbiased and precise estimates can even be obtained with only 10 individuals per population and 10 loci. Note that some methods performed worst on average with higher numbers of loci. This is clearly a consequence of the systematic bias observed at marginal admixture values and low differentiation of parental populations. In these conditions, adding loci will not reduce the bias at all, but will produce the opposite effect of reducing the λ CI around a false (biased) estimate.

Application to real data set: Honeybee populations in the Alps

European honeybee populations are profoundly differentiated in two evolutionary branches named C and M that diverged around 1 Ma (Garnery *et al*. 1992) corresponding to $\sim 5 \times 10^5$ generations. However, natural hybridizations occur along the Alpine arc between the Italian subspecies *Apis mellifera ligustica* (branch C) and the West European subspecies *A. m. mellifera* (branch M) (Franck *et al*. 2000). Furthermore, the recurrent introduction of *A. m. ligustica* queens for > 50 years by beekeepers in France has substantially modified the genetic composition of French *A. m. mellifera* populations (Garnery *et al*. 1998).

Admixture proportions were estimated in four populations collected along a transect across the Alps (St.

A

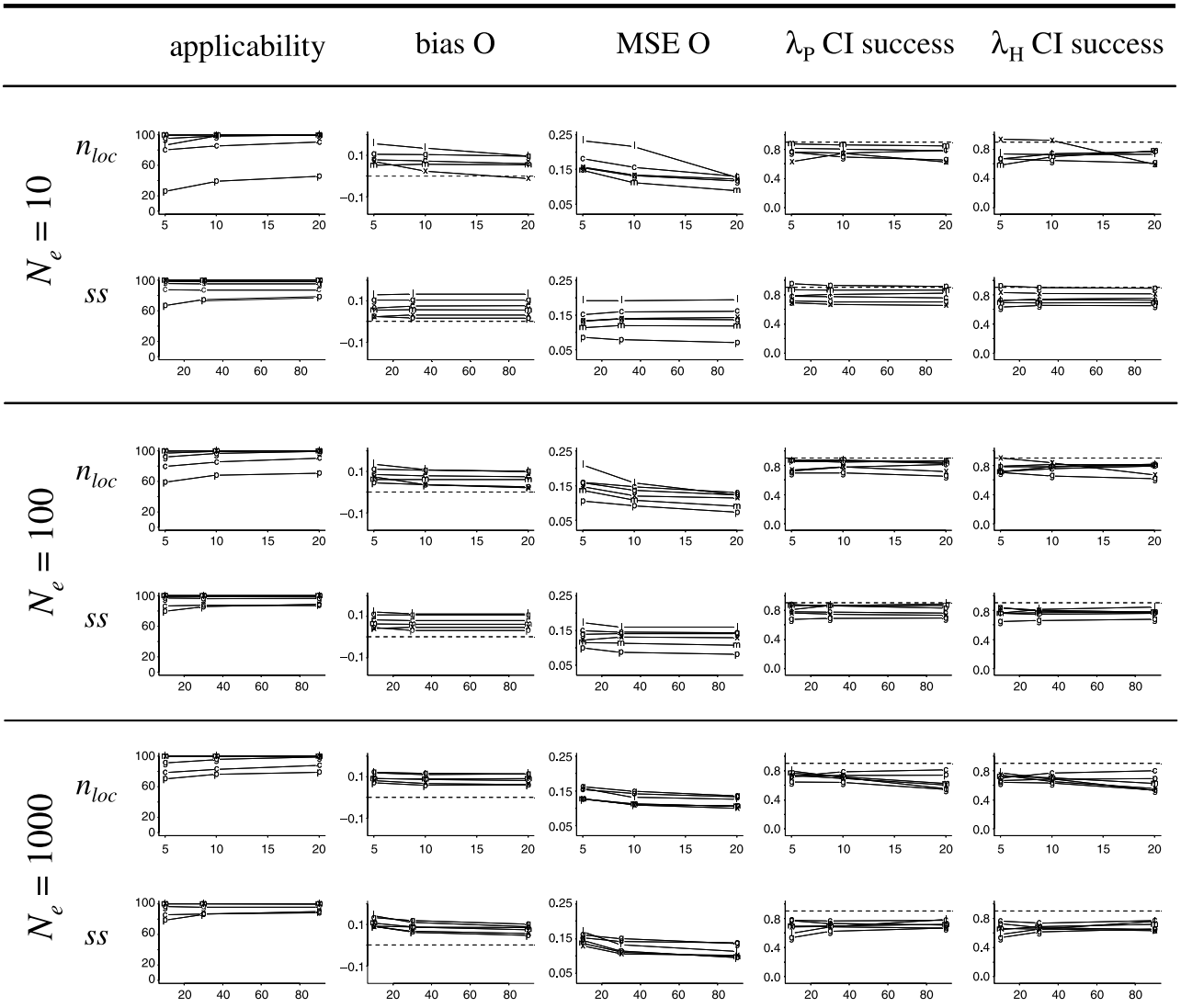

Fig. 3 Effects of experimental conditions. Each value of N_e corresponds to a pair of lines. For each pair of lines the upper graphs represent one appreciation criterion (y axis) as a function of the number (n_{loc}) of loci studied (x axis) and the lower graphs represent the same appreciation criteria (y axis) as a function of the number (ss) of sampled individuals (x axis). Parameter λ_H is the admixture proportion at the time of the hybridization and λ_P is the current proportion of genes descending from P_1 in the hybrid population sample (modified from λ_H by genetic drift). Each point of the curves is a mean value over the 12 combinations of the values of the parameters g_0, g_1. The methods are the gene identity (g), least square (l), private alleles (p), maximum likelihood (m), coalescence times (c), and Monte Carlo Markov chain (x). Each point for which applicability was < 50% was not represented on the graph. A: $\lambda_H = 0.1$, B: $\lambda_H = 0.5$. Coloured versions of Fig. 3A and B are available at www.montpellier.inra/CBGP/progrechCornuet.htm.

Vincent and Courmayeur in the Aosta Valley and Bourg St. Maurice and Annecy in Savoy) and in one population from southwestern France (Sabre). A pure *A. m. mellifera* population was collected in a sanctuary of the black honeybee in French Brittany (Ouessant). A pure *A. m. ligustica* population was sampled in Emily Romagne (Forlì) which is the main region that rear and export Italian honeybee queens. These last two samples were considered as the parental references. The sampling populations were genotyped at eight microsatellite loci (*A113*, *A28*, *A43*, *A8*, *A88*, *Ap43*, *B124* and *A24*) previously described by Estoup *et al*. (1995) and Franck *et al*. (1998).

The methods gave similar but different estimations of admixture proportions in each locality, the difference between extreme values reaching 0.2 in most cases (Fig. 4). However, each method ranks population estimates in the

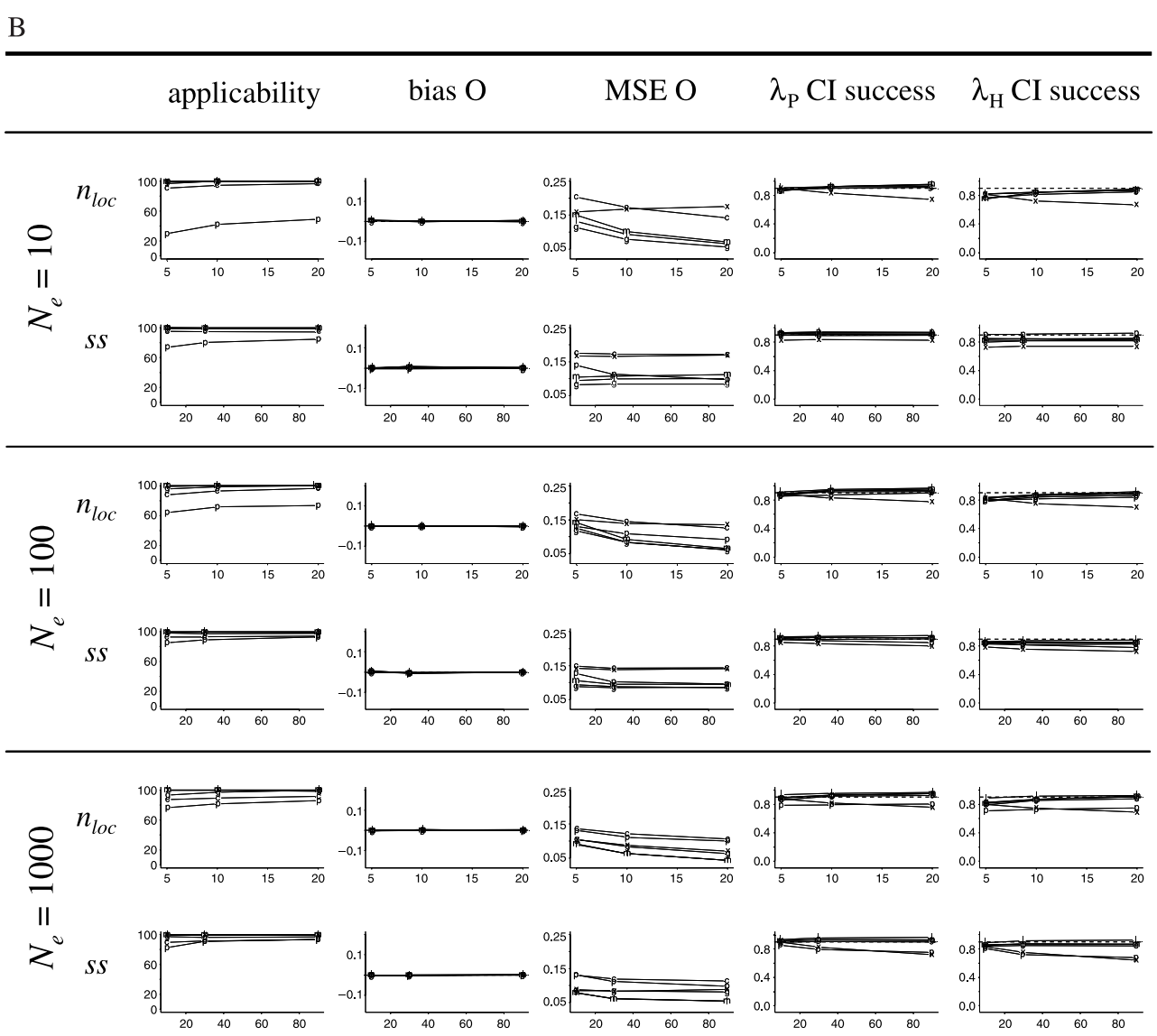

Fig. 3 *Continued*

same order. St. Vincent and Courmayeur are on the natural hybridization zone between the two subspecies. It is thus no surprise to find the highest proportion of *ligustica* genes in these populations. Moreover, in St. Vincent, which is located 50 km downstream of Courmayeur, estimations are clearly higher than in Courmayeur (~0.90 vs. ~0.70). Admixture estimations in Bourg St. Maurice and Annecy are quite similar (between 0.10 and 0.20). In Sabres introgression seems to be very low (< 0.10) which confirms the fact that in this place the French honeybee is considered to be very little admixed (Cornuet *et al*. 1982). Note that no value was obtained for Sabres when using the private allele method. Note also that in four of five populations, Chikhi *et al*.'s (2001) method provided more extreme admixture values. This is consistent with the observation that the MC method always shows less bias towards central values than any other method (cf. Fig. 2A).

Discussion

Applicability

When estimation methods are compared, it is generally assumed that computation is achievable. In our case, and for any of the five methods, there were situations in which computations aborted. This is why we introduced the criterion of applicability, which included two features of different significance. One is a measure of how often computations did not fail in providing an estimate of λ. The second is the frequency at which the estimate of λ

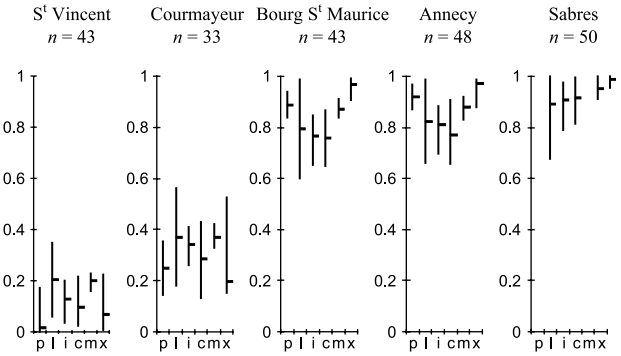

Fig. 4 Estimates of the introgressive rate of *Apis mellifera ligustica* within *A. m. mellifera* for five honeybee populations based on microsatellite data. As parental populations we used one from Ouessant French Brittany (49 individuals) for *A. m. mellifera* and one from Forlì in Central Italy (19 individuals) for *A. m. ligustica*. The number (*n*) of individuals sampled in each locality is written under the name of the city. Estimations were performed on a dataset of height microsatellite loci (see text). The small horizontal bar is the estimation and the vertical line represent the confidence interval ($\alpha = 0.10$). Methods used are, from left to right: gene identity (g), least square (l), private alleles (p), maximum likelihood (m), coalescence times (c) and Monte Carlo Markov chain method (x).

decreased to outside the expected [0; 1] interval. A special analysis was performed on a restricted sample of simulated data sets including all combinations of parameters but with only 10 repetitions instead of 100. This showed that all 'failures' of methods GI and CT were because estimates of λ were not in the [0; 1] interval. An estimate of λ of, for example, –0.1 or 1.05 would not be a real problem on actual data. So why use such a 'severe' criterion and reject outliers? When inspecting these outliers, we saw that they took values on a very large interval ([–20; +26.5] for GI and [–103; +36] for CT). If we kept all these values, the most erratic outliers would have had the largest impact on average biases, potentially altering our appreciation of the real performances of the methods. Another possibility would have been to set to 0 negative estimates and to 1 those > 1. However, we think that it is better to be aware that these methods can lead to values that are far from the [0; 1] segment and to have an estimation of how often they do so. In our conditions, the GI and CT methods provided an estimated value outside [0; 1] in 11.5 and 12.6% of simulated data sets, respectively. With both methods, this occurred more often with a larger N_e, a smaller differentiation, an older hybridization time, smaller number of loci and smaller sample size. Concerning method CT, a possible explanation of values outside the [0; 1] segment might be the use of a different mutation model (bounded GSM) in the data simulation than the one assumed in the m_γ formula (strict SMM). Only under the strict SMM is the time of coalescence proportional to the squared difference of allele sizes. Multi-step changes due to GSM will provide overestimated times of coalescence and possibly set the estimator m_γ outside the expected [0; 1] segment. Method ML does not suffer from the same problem because the estimate of λ is constrained to the [0; 1] interval in the computations. The only situations where computations failed were when all the alleles sampled in the hybrid population were private to this population (the likelihood of genotypes were then equal to 0 for any value of λ). The occurrence of such a situation was very low (0.04% in our tests). It was highly favoured by a low number of loci, a low effective population size and an old hybridization event. Methods LS and PA mainly failed for computational reasons. Method LS involves the inversion of a matrix and it essentially failed because the pivot happened to be null (in 3.14% of our tests). Note that this failure occurred more frequently at marginal values of λ (6.92% at $\lambda = 0.1$ vs. 1.26% at $\lambda = 0.5$). Eventually, method PA proved to be the most problematic, reaching almost 0% applicability in some situations (old hybridization event and low effective sizes). We observed 11 possible types of error among which the absence of private alleles was only a minor reason as it led to failure in only 0.04% of our tests. Far more common reasons were failures in the computations themselves such as null denominators. This occurred for instance when, in the reference parental population, all loci were monomorphic for their own private alleles or when private allele frequencies were all equal either in the reference parental population or in the hybrid population. These situations occurred particularly with low effective population sizes which increase the occurrence of monomorphic loci. As expected, increasing the number of loci decreased the probability of such situations: the relative frequency of such errors in our tests decreased from 60% with 5 loci to 1% with 40 loci. In addition to these computation failures, method PA also suffered from frequent outliers in the estimation of λ (in ~30% of our tests). As a consequence of its low applicability, the performances of method PA had a less stable behaviour in a few situations because performances were estimated on a largely reduced sample of data sets. In line with results on simulated data sets, this method failed in one population over five (the one with the more extreme admixture value according to all other methods) in the honeybee data set.

At the other end of the spectrum, method MC always provided an estimate of admixture proportion between 0 and 1. However, a different problem arose. In ~10% cases, the Markov chains did not mix well and became stuck at a likelihood maximum, resulting in a poor estimation of the quantiles (e.g. the 5 and 50% quantiles were identical). This poor mixing was more frequent with a higher number of loci and/or a higher sample size, as expected because the likelihood decreases sharply when the number of loci × individuals increases. Less evidently, the poor mixing

frequency also increased with the effective population size, with the level of differentiation of parental populations and with the time since admixture. Note, however, that all these conditions imply a lower likelihood mainly because they favour a higher total number of alleles.

Present (λ_P) vs. hybridization time (λ_H) admixture proportions

For clarity's sake, we distinguished two admixture proportions, one at the time of hybridization (λ_H) and one in the present-day hybrid sample (λ_P). To be exhaustive, we should have also considered the admixture proportion in the present-day population. However, the five methods do not make clear distinctions among these possible definitions. Because they use present-day data (allele frequencies for five of them and also allele states for method CT), they actually estimate λ_P. With neutral markers, allelic frequencies can fluctuate at random, but they keep constant expectations across generations and so does parameter λ. Because of these fluctuations, the precision of the estimation of λ_H from present-day data is expected (i) to be lower than that of λ_P and (ii) to decrease with the number of generations since hybridization. In our tests, we computed the relative bias and MSE considering both definitions and found negligible differences between them when averaging like in Figs 2 and 3 (remember that points represent averages of 900 and 3600 data sets in Figs 2 and 3, respectively). However perceptible differences exist on each individual result and this is why CI successes were generally higher for λ_P than for λ_H. It is interesting to notice that in optimal and suboptimal conditions and for several methods (e.g. LS, ML, GI and MC), confidence intervals include almost 100% the true value of λ_P, whereas they include generally less than the expected 90% the true value of λ_H.

Bias

When used under optimum conditions, no method is intrinsically biased. However, when the real value is marginal, a systematic bias towards central values of the admixture proportion (Figs 2A and 3A) occurs as the admixture gets older. This is logical because when the available information decreases, the admixture proportion tends to the average (0.5) of a uniform distribution over [0; 1]. For the same reason, there is virtually no bias when the admixture proportion is close to 0.5 (Figs 2B and 3B). Initially, we found a negative bias in several methods (GI, PA and LS). To avoid this bias, we had to modify these methods slightly by performing the computations in the two possible directions (taking each parental population as a reference) and averaging the results. For method MC, the coefficients of the prior distributions (Dirichlet) of parental allele frequencies were modified. The initial coefficients (= 1) produced a negative average bias of the admixture proportion which could reach 0.07 in the worst conditions ($N_e = 10^3$ and $g_0 = 10^6$). Coefficients equal to one correspond to the situation in which all prior configurations of allelic states of ancestral lines are equally frequent. Although such a prior distribution is often proposed for allelic frequencies, it does not seem well suited. Some authors (e.g. Rannala & Mountain 1997) preferred to use coefficients that are the inverse of the number of alleles, corresponding to a more U-shaped distribution of allelic frequencies. This is why we redid the computations with such values of the Dirichlet coefficients. This proved to be much more satisfactory, especially for large values of N_e. However, some residual bias is still visible for small values of N_e. This change of the Dirichlet coefficients is pretty ad hoc and it is clear that a more convincing way of dealing with this step of the method is needed.

Comparison of methods

As already mentioned, all six methods perform equally well under optimal conditions (highly differentiated parental populations and recent hybridization event). However, they can display very different performances when departing from these optimal conditions. We will review shortly the positive and negative characteristics of each method.

Gene identities. In our tests, this method had good performances at central admixture proportions ($\lambda_H = 0.5$): it provided unbiased and precise estimates in most situations, especially when effective population sizes are small and the hybridization is old, although in the latter case, the confidence interval success for estimating the current admixture (λ_P) is below expectation. When the admixture proportions are more marginal ($\lambda_H = 0.1$), this method appeared to be in the middle range for bias and precision, but the λ CI successes decreased dramatically when the hybridization age increased beyond 100 generations. Note that this method was mainly applied on racial admixture between Black and White North Americans, an admixture which occurred not more than 15 generations ago. Also, parental populations are well differentiated and the admixture is thought to be rather high (between 25 and 50%) (Chakraborty 1986). Unless the admixture is high and the hybridization is recent (within the 50 last generations), this method provides less trustable estimations and therefore should not be applied (Korey 1978).

Least square regression. In most cases, this method performs satisfactorily. This might be one of the causes of its long-time success (Roberts & Hiorns 1962, 1965; Elston 1971; Long & Smouse 1983; Long 1991). However, it is not always

applicable, especially with marginal admixture ($\lambda_H = 0.1$), small effective sizes and old hybridization.

Private alleles. The main drawback of this method is its poor applicability, which decreases sharply when the admixture gets older. This is especially true when the effective sizes (N_e) are low. When N_e is too low, not only can this method be applied in only a very few number of cases, but it also gives biased estimations and its confidence interval is so large that it is of no practical use. One cause of this may be that when the effective sizes are too low, genetic drift becomes so high that private alleles in the hybrid population disappear quickly.

Maximum likelihood. This method is very applicable. This is simply because, in our algorithm, the maximum likelihood is numerically sought within the interval [0; 1]. The only cases where computations fail are when parental populations have identical allele frequencies at all loci. This occurs mainly (if not only) when loci are fixed for the same allele, a situation favoured by small effective population sizes and low divergence times between parental populations. Its bias and standard deviation are always among the best, which certainly makes this method one of the most often advised (Krieger *et al*. 1965; Roberts & Hiorns 1965; Elston 1971; Korey 1978). The best point of this method is that the bias stays at a low level even for ancient hybridization, point already noted by Korey (1978). Note however, that for the last criterion (λ_H CI success) this method ranks among the worst when admixture is old and differentiation is low.

Coalescence times. This method is among those that give the lowest bias except when differentiation is low. Like Bertorelle & Excoffier (1998), we note that the bias is low even for ancient hybridizations. It also provides confidence intervals that are generally closest to the expected score of 90% for λ_H CI success. Yet, the interval length is not particularly high. However, again like Bertorelle & Excoffier (1998), we note a slightly higher variance of the estimations given by this method and this is certainly its main although minor drawback. All in all, this method appears to be a good choice in most situations.

Monte Carlo Markov chain method. This method takes much longer than the five preceding methods, even with the improvements achieved here. However, there is no definite advantage of choosing the MC method. The only clearly superior feature is its 100% applicability. The fact that the Markov chain can 'stick' from time to time is not a drawback per se because there are known simple remedies, for example, increasing the number of iterations of the importance sampling scheme (which was arbitrarily set to 50 in our study) or recomputing the likelihood at each proposed value of the admixture proportion. It is probable that the performance of this method would have been increased slightly if the sticking had been avoided. Having said that, one has to recognize that the performance is generally among the best. The method performs better for marginal admixture proportions, especially with a low N_e and a low differentiation of parental populations (upper row of Fig. 2A), i.e. a situation where other methods behave rather badly.

The methods assayed in this are necessarily partial. An article published at the time we were finishing our study (Wang 2003) provides another maximum likelihood-based method that seems superior to some of the methods assayed here. Like Chikhi *et al*.'s (2001) MC method, it is based on the computation of the likelihood of the samples, but instead of using a sampling scheme, it uses the transition matrix method to compute the probability of observed allele frequencies. Several tricks are used to lower the computational load, so that computations are fast enough. Also, the genealogical scheme is considered up to the ancestral population of the two parental populations. The method can be adapted to various situations of admixture (more than two parental populations, admixture due to constant migration).

Wang's (2003) method is a pure drift method like the other methods studied here, except the CT method. When considering highly variable markers such as microsatellite loci, mutations may have some impact on estimations. Of all the methods considered (except CT), the MC is the only one that can readily incorporate mutations, by changing the importance sampling scheme. because this method needs further improvements to get rid of the residual bias evidenced here, why not looking for a new MCMC method before making further comparisons?

Acknowledgements

This study benefited from a grant by the French BRG (Bureau des Ressources Génétiques). We are grateful to Arnaud Estoup and Réjane Streiff for critically reading a former version of the manuscript. We also thank the Associate Editor Laurent Excoffier for suggesting that we include the MC method in our comparative analysis and Mark Beaumont who gave us the original source code of the LEA software and suggested improving the hybridization step in the importance sampling algorithm.

References

Abernethy K (1994) The establishment of a hybrid zone between red and sika deer (genus *Cervus*). *Molecular Ecology*, **3**, 551–562.

Bernstein F (1931) Die geographische Verteilung der Blutgruppen und ihre anthropologische Bedeutung. *Comitato Italiano Per Lo Studio Dei Problemi Della Populazione*, pp. 227–243. Instituto Poligraphico Dello Stato, Rome.

Bertorelle G, Excoffier L (1998) Inferring admixture proportion from molecular data. *Molecular Biology and Evolution*, **15**, 1298–1311.

Byard PJ, Schanfield MS, Crawford MH (1985) Admixture and heterozygosity in West Alaskan populations. *Journal of Biosocial Science*, **15**, 207–216.

Cavalli-Sforza LL, Bodmer WF (1971) *The Genetics of Human Populations*. Freeman, San Fransisco.

Chakraborty R (1975) Estimation of race admixture — a new method. *American Journal of Physical Anthropology*, **42**, 507–511.

Chakraborty R (1986) Gene admixture in human populations: models and predictions. *Yearbook of Physical Anthropology*, **29**, 1–43.

Chakraborty R, Ilyas Kamboh M, Ferrell E (1991) 'Unique' alleles in admixed populations: a strategy for determining 'hereditary' population differences of disease frequencies. *Ethnicity and Disease*, **1**, 245–256.

Chakraborty R, Weiss KM (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proceedings of National Academy of Sciences of the USA*, **85**, 9119–9123.

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

Cornuet J-M, Albisetti J, Mallet N, Fresnaye J (1982) Etude biométrique de populations d'abeilles landaises. *Apidologie*, **19**, 355–366.

Cornuet J-M, Daoudi A, Chevalet C (1986) Genetic pollution and number of mating in a black honey bee (*Apis mellifera mellifera*) population. *Theoretical and Applied Genetics*, **73**, 223–227.

De Wayne Shoemaker D, Ross KG, Arnold ML (1996) Genetic structure and evolution of a fire ant hybrid zone. *Evolution*, **50**, 1958–1976.

Destro-Bisol G, Maviglia R, Caglia A *et al.* (1999) Estimating European admixture in African Americans by using microsatellites and a microsatellite haplotype (CD4/Alu). *Human Genetics*, **104**, 149–157.

Ellstrand NC, Prentice HC, Hancock JF (1999) Gene flow and introgression from domesticated plants into their wild relatives. *Annual Review of Ecology and Systematics*, **30**, 539–563.

Elston RC (1971) The estimation of admixture in racial hybrids. *Annals of Human Genetics*, **35**, 9–17.

Estoup A, Garnery L, Solignac M, Cornuet J-M (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 AIJW, Sullivan C, Cornuet JM, Moritz C (2001) Inferring population history from microsatellite and enzyme data in serially introduced cane toads, *Bufo marinus*. *Genetics*, **159**, 1671–1687.

Franck P, Garnery L, Celebrano G, Solignac M, Cornuet J-M (2000) Hybrid origins of honeybees from Italy (*Apis mellifera ligustica*) and Sicily (*A. m. sicula*). *Molecular Ecology*, **9**, 907–921.

Franck P, Garnery L, Solignac M, Cornuet J-M (1998) The origin of west European subspecies of honeybees (*Apis mellifera*): new insights from microsatellite and mitochondrial data. *Evolution*, **52**, 1119–1134.

Futuyama D (1998) *Evolutionary Biology*. Sinauer Associates, Sunderland, MA.

Garnery L, Cornuet J-M, Solignac M (1992) Evolutionary history of the honey bee *Apis mellifera* inferred from mitochondrial DNA analysis. *Molecular Ecology*, **1**, 145–154.

Garnery L, Franck P, Baudry E *et al.* (1998) Genetic biodiversity of the west European honey bee (*Apis mellifera meliffera* and *A. m. iberica*) — II. Microsatellite loci. *Genetic, Selection and Evolution*, **30**, 49–74.

Giuffra E, Guyomard R, Forneris G (1996) Phylogenetic relationships and introgression pattern between incipient parapatric species of Italian brown trout (*Salmo trutta* L. complex). *Molecular Ecology*, **5**, 207–220.

Glass B (1955) On the unlikelihood of significant admixture of genes from the North American Indians in the present composition of the Negroes of the United States. *American Journal of Human Genetics*, **7**, 368–385.

Glass B, Li CC (1953) The dynamics of racial intermixture — an analysis based on the American Negro. *American Journal of Human Genetics*, **5**, 1–19.

Goodman SJ, Barton NH, Swanson G, Abernethy K, Pemberton JM (1999) Introgression through rare hybridization: a genetic study of a hybrid zone between red and sika deer (genus *Cervus*) in Argyll, Scotland. *Genetics*, **152**, 355–371.

Goostrey A, Carss DN, Noble LR, Piertney SB (1998) Population introgression and differentiation in the great cormorant *Phalacrocorax carbo* in Europe. *Molecular Ecology*, **7**, 329–338.

Gotteli D, Sillero-Zubiri C, Applebaum GD *et al.* (1994) Molecular genetics of the most endangered canid: the Ethiopian wolf *Canis simensis*. *Molecular Ecology*, **3**, 301–312.

Griffith RC, Tavaré S (1994) Simulating probability distributions in the coalescent. *Theoretical Population Biology*, **46**, 131–159.

Haig SM, Gratto-Trevor CL, Mullins TD, Colwell MA (1997) Population identification of western hemisphere shorebirds throughout the annual cycle. *Molecular Ecology*, **6**, 413–428.

Hanis CL, Chakraborty R, Ferrell RE, Schull WJ (1986) Individual admixture estimates: disease associations and risk of diabetes and gallbladder disease among Mexican Americans of Starr County, Texas. *American Journal of Physical Anthropology*, **70**, 433–441.

Korey KA (1978) A critical appraisal of methods for measuring admixture. *Human Biology*, **50**, 343–360.

Krieger H, Morton NE, Mi MP *et al.* (1965) Racial admixture in north-eastern Brazil. *Annals of Human Genetics*, **29**, 113–125.

Long JC (1991) The genetic structure of admixed populations. *Genetics*, **127**, 417–428.

Long JC, Smouse PE (1983) Intertribal gene flow between the Ye'cuana and Yanomama: genetic analysis of an admixed village. *American Journal of Physical Anthropology*, **61**, 411–422.

Madansky A (1959) The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association*, **54**, 173–205.

McKeigue PM (1997) Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. *American Journal of Human Genetics*, **60**, 188–196.

McLean CJ, Workman PL (1973) Genetic studies on hybrid populations. II. Estimation of distribution of ancestry. *Annals of Human Genetics*, **36**, 459–465.

Neel JV (1973) 'Private' genetic variants and the frequency of mutation among South American Indians. *Proceedings of the National Academy Sciences of the USA*, **70**, 83–87.

Nielsen EE, Hansen MM, Loeschcke V (1997) Analysis of microsatellite DNA from old scale samples of Atlantic salmon *Salmo solar*: a comparison of genetic composition over 60 years. *Molecular Ecology*, **6**, 487–492.

Paetkau D, Calvert W, Stirling I, Strobeck C (1995) Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology*, **4**, 347–354.

Paetkau D, Strobeck C (1994) Microsatellite analysis of genetic variation in black bear populations. *Molecular Ecology*, **3**, 489–495.

Parra EJ, Marcini A, Akey J *et al*. (1998) Estimating African American admixture proportions by use of population-specific alleles. *American Journal of Human Genetics*, **63**, 1839–1851.

Poteaux C, Bonhomme F, Berrebi P (1998) Difference between nuclear and mitochondrial introgression of brown trout populations from a restocked main river and its unrestocked tributary. *Biology Journal of the Linnean Society*, **63**, 379–392.

Rannala B, Mountain JL (1997) Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences of the USA*, **94**, 9197–9201.

Reed TE (1969) Caucasian genes in American Negroes. *Science*, **165**, 762–768.

Reed JZ, Tollit DJ, Thompson PM, Amos W (1997) Molecular scatology: the use of molecular genetic analysis to assign species, sex and individual identity to seal faeces. *Molecular Ecology*, **6**, 225–234.

Reich DE, Wayne RK, Goldstein DB (1999) Genetic evidence for a recent origin by hybridization of red wolves. *Molecular Ecology*, **8**, 139–144.

Roberts DF (1955) The dynamics of racial admixture in American Negro: some anthropological considerations. *American Journal of Human Genetics*, **7**, 361–367.

Roberts DF, Hiorns RW (1962) The dynamics of racial admixture. *American Journal of Human Genetics*, **14**, 261–277.

Roberts DF, Hiorns RW (1965) Methods of analysis of the genetic composition of a hybrid population. *Human Biology*, **37**, 38–43.

Simonsen KL, Churchill GA, Aquadro CF (1995) Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*, **141**, 413–429.

Szathmary EJE, Reed TE (1978) Calculation of the maximum amount of gene admixture in a hybrid population. *American Journal of Physical Anthropology*, **48**, 29–34.

Szymura JM, Barton NH (1986) Genetic analysis of a hybrid zone between the fire-bellied toads, *Bombina bombina* and *B. variegata*, near Cracow in Southern Poland. *Evolution*, **40**, 1141–1159.

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

Wayne RK, Jenks SM (1991) Mitochondrial DNA analysis implying extensive hybridization of the endangered red wolf, *Canis rufus*. *Nature*, **351**, 565–568.

Williams RC, Steinberg AG, Knowler WC, Pettitt DJ (1986) Gm3,5,13,14 and stated-admixture: independent estimates of admixture in American Indians. *American Journal of Human Genetics*, **39**, 409–413.

The present study forms the main part of Marc Choisy's MSc under the supervision of Jean-Marie Cornuet whose interests are on evolutionary genetics of honeybees. Pierre Franck completed a PhD thesis on evolutionary sociobiology and phylogeography of honeybees under the same supervisor. He is currently working on integrated pest management in apple trees.
