Inferring Admixture Proportions from Molecular Data

Giorgio Bertorelle and Laurent Excoffier†*

*Department of Integrative Biology, University of California, Berkeley; and †Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva, Switzerland

We derive here two new estimators of admixture proportions based on a coalescent approach that explicitly takes into account molecular information as well as gene frequencies. These estimators can be applied to any type of molecular data (such as DNA sequences, restriction fragment length polymorphisms [RFLPs], or microsatellite data) for which the extent of molecular diversity is related to coalescent times. Monte Carlo simulation studies are used to analyze the behavior of our estimators. We show that one of them (m_Y) appears suitable for estimating admixture from molecular data because of its absence of bias and relatively low variance. We then compare it to two conventional estimators that are based on gene frequencies. m_Y proves to be less biased than conventional estimators over a wide range of situations and especially for microsatellite data. However, its variance is larger than that of conventional estimators when parental populations are not very differentiated. The variance of m_Y becomes smaller than that of conventional estimators only if parental populations have been kept separated for about N generations and if the mutation rate is high. Simulations also show that several loci should always be studied to achieve a drastic reduction of variance and that, for microsatellite data, the mean square error of m_Y rapidly becomes smaller than that of conventional estimators if enough loci are surveyed. We apply our new estimator to the case of admixed wolflike Canid populations tested for microsatellite data.

Introduction

Over the course of evolution, populations that have remained isolated from one another for a long period because of geographical, ecological, or cultural barriers are occasionally brought into contact. They then have the possibility of exchanging genes, leading to admixed or hybrid populations presenting some characteristic genetic features, such as gene frequencies that are intermediate between those of the parental populations (Cavalli-Sforza and Bodmer 1971), as well as linkage disequilibrium between independent loci (Nei and Li 1973). The problem of estimating the relative genetic contribution of the parental population to the admixed populations has been discussed repeatedly in the literature for more than 60 years (since Bernstein 1931). Even so, the term ''admixture'' can refer to quite distinct evolutionary processes. For some people, it specifically refers to the process described above, whereas for some others it refers to a special case of gene flow between populations.

Although most of the methodological developments concerning admixed populations have arisen in the human context, admixture is also very frequent in other species, involving not only different populations but also different interfertile groups or subspecies. The analysis of the composition of hybrid populations using observed genetic data has been the subject of many studies, and several estimators of the relative contributions of the source populations to the admixed population (the admixture coefficients) have been proposed (see Chakraborty 1986 for a review). All of these estimators are based on the comparison of allele or phenotype frequencies between the source population and the admixed populations. They rely on the simple fact that allele frequencies observed in the admixed population should be linear combinations of allele frequencies in the source populations, assuming that the effect of random drift occurring after the admixture event is negligible.

With the advent of molecular techniques, potentially more refined information has been made available to those interested in disentangling the respective contributions of founder populations to a given gene pool. However, even though they deal with molecular information, recent admixture studies (e.g., Hammer and Horai 1995; Horai et al. 1996) or mixed-population stocks studies (Xu, Kobak, and Smouse 1994) have used estimators of population contributions based only on allele frequencies. It appears important, therefore, to develop estimators of admixture that use molecular information explicitly because the number of evolutionary steps separating different alleles can also be meaningful. For instance, a hybrid population might be considered very close to one potential source population on the basis of its gene frequencies, but if the hybrid and a different source population share some very divergent alleles, the first conclusion would probably be altered.

We propose here two new estimators of admixture coefficients based on the mean coalescent time of genes drawn either within or between admixed and parental populations. Our estimators can be applied to any type of molecular data for which the amount of molecular diversity is related to those coalescent times, including DNA sequences, restriction fragment length polymorphisms (RFLPs), or microsatellite data. The behavior of our estimators is evaluated and compared with that of two conventional estimators (Roberts and Hiorns 1965; Chakraborty et al. 1992) by the use of Monte Carlo simulation studies. Finally, we apply our methodology to the case of hybrid wolf and coyote populations tested for 10 microsatellite loci (Roy et al. 1994).

Key words: wolflike Canids, admixture coefficient, coalescent, least-squares estimator, microsatellite, DNA sequences, Monte Carlo simulations.

Address for correspondence and reprints: Laurent Excoffier, Department of Anthropology, 12, rue G. Revilliod, 1227 Geneva, Switzerland. E-mail: Laurent.Excoffier@anthro.unige.ch.

Mol. Biol. Evol. 15(10):1298–1311. 1998

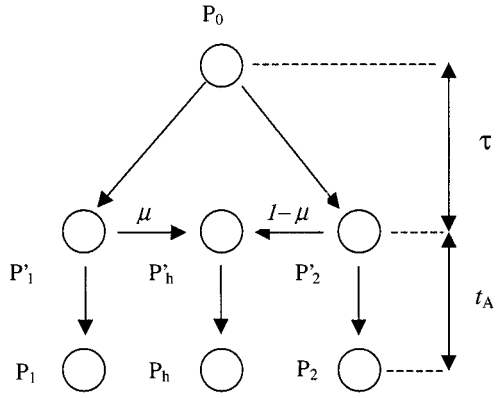

FIG. 1.—Demographic model of admixture used throughout this study. See text for description.

Materials and Methods

Derivation of the Estimators

Here, we introduce two estimators of the admixture coefficient based on the average coalescence times between pairs of genes sampled within and between populations. We consider the simple model of admixture shown in figure 1. An ancestral haploid population P_0 of size N splits into two parental populations P_1' and P_2', which evolve independently for τ generations. At that point, a hybrid population P_h' is instantaneously created by combining $N\mu$ genes taken at random from P_1' and $N(1-\mu)$ genes taken at random from P_2'. From then, and for t_A generations until the present time, the three populations are kept separate and do not exchange genes. During the whole process, all populations are assumed to have the same constant size, N. The parameter μ is thus the genetic contribution of population P_1' to the hybrid. This is the parameter we will try to estimate from samples of genes taken from three populations: P_1, P_2, and P_h.

The first estimator, m_X, is derived from the expected coalescence time between two genes sampled in the hybrid population P_h. Replacing the parameters with their estimations in equation (A4), derived in the *Appendix*, the estimator takes the form of

$$m_X = \frac{1}{2} \pm \frac{1}{2}\sqrt{1 - 2\frac{a}{b}}, \qquad (1)$$

where $a = \hat{\bar{t}}_h - \hat{\bar{t}}_0$, and $b = \hat{\tau}\exp(-\hat{t}_A/\hat{\bar{t}}_0)$. Here, $\hat{\bar{t}}_0$ is an estimator of the mean coalescence time between two genes drawn from the same parental population, simply obtained as $\hat{\bar{t}}_0 = (\hat{\bar{t}}_{11} + \hat{\bar{t}}_{22})/2$. An estimator of τ can be obtained as $\hat{\tau} = \hat{\bar{t}}_{12} - \hat{t}_A - \hat{\bar{t}}_0$. Finally, $\hat{t}_A$ can be roughly computed as the smallest coalescence time observed among all pairs of genes in which one gene of the pair is sampled from the hybrid and the other is sampled from one of the parental populations (Takahata and Nei 1985).

Even though quite simple, the use of equation (1) is not straightforward for several reasons. First, one can get negative terms under the square root if a/b is larger than 0.5, making expression (1) impossible to evaluate. This can happen in several instances: when the divergence between parental populations is recent (i.e., τ is small), when the admixture event is old (i.e., t_A is large), and when the contribution of each of the two parental populations is almost equal (when the true μ value is close to 0.5). Second, when the term under the square root is strictly positive, some external information, such as the genetic distance between the hybrid and parental populations, is needed to choose the correct solution for m_X. For example, if the hybrid population is genetically closer to population 1 than to population 2, one should take the larger of the two solutions as the correct solution.

A second estimator of μ, m_Y, can be inferred by considering the coalescent times of genes drawn from the admixed populations and from the parental populations P_1 and P_2 (see *Appendix*). The estimator m_Y can be derived from equation (A8) by replacing the parameters with their estimators, as follows:

$$m_Y = \frac{c\hat{\bar{t}}_{h1} - d\hat{\bar{t}}_{h2} + d^2 + \hat{\bar{t}}_{12}(\hat{\bar{t}}_{h2} - \hat{\bar{t}}_{h1} + e)}{c^2 + d^2 + 2e\hat{\bar{t}}_{12}}, \qquad (2)$$

where $c = \hat{t}_A + \hat{\bar{t}}_{11}$, $d = \hat{t}_A + \hat{\bar{t}}_{22}$, and $e = \hat{\bar{t}}_{12} - (c + d)$.

Admixture Proportions from Molecular Data

When dealing with molecular data, the coalescence times between two genes are not directly available, and mean coalescence times, $\bar{t}$'s, must be estimated from the genetic variability. Here, we consider two possible estimates based on two mutation models: the infinite-site model for DNA sequences and the single-step stepwise model for microsatellite loci.

For DNA sequences (or RFLPs), assuming that each new mutation occurs at a previously monomorphic site (the infinite-site model), mean coalescence times can be estimated from the mean number of pairwise differences, π, as $\hat{\bar{t}} = \pi/2u$, where u is the global mutation rate. It follows that all the coalescence times used to compute m_X and m_Y can simply be replaced by their corresponding π's because the mutation rate cancels out, as it appears in both numerators and denominators in equations (1) and (2). For instance, the estimator m_Y now takes the form

$$m_Y = \frac{c\hat{\pi}_{h1} - d\hat{\pi}_{h2} + d^2 + \hat{\pi}_{12}(\hat{\pi}_{h2} - \hat{\pi}_{h1} + e)}{c^2 + d^2 + 2e\hat{\pi}_{12}}. \qquad (3)$$

where $c = \hat{t}_A' + \hat{\pi}_{11}$, $d = \hat{t}_A' + \hat{\pi}_{22}$, $e = \hat{\pi}_{12} - (c + d)$, and $\hat{\pi}_{11}$, $\hat{\pi}_{22}$, $\hat{\pi}_{12}$, $\hat{\pi}_{h1}$, $\hat{\pi}_{h2}$, are the mean number of pairwise differences within P_1, within P_2, between P_1 and P_2, between the admixed population and P_1, and between the admixed population and P_2, respectively. Here, $\hat{t}_A'$ (which is now the age of the admixture event expressed in units of $1/(2u)$ generations) can be estimated, in practice, by the minimum number of pairwise differences observed between a gene drawn from the admixed population and a gene drawn from a parental population. It will thus have a value of zero if at least one sequence is found both in the admixed population and in the parental population. This situation can occur often if the admixture event is recent, implying that this

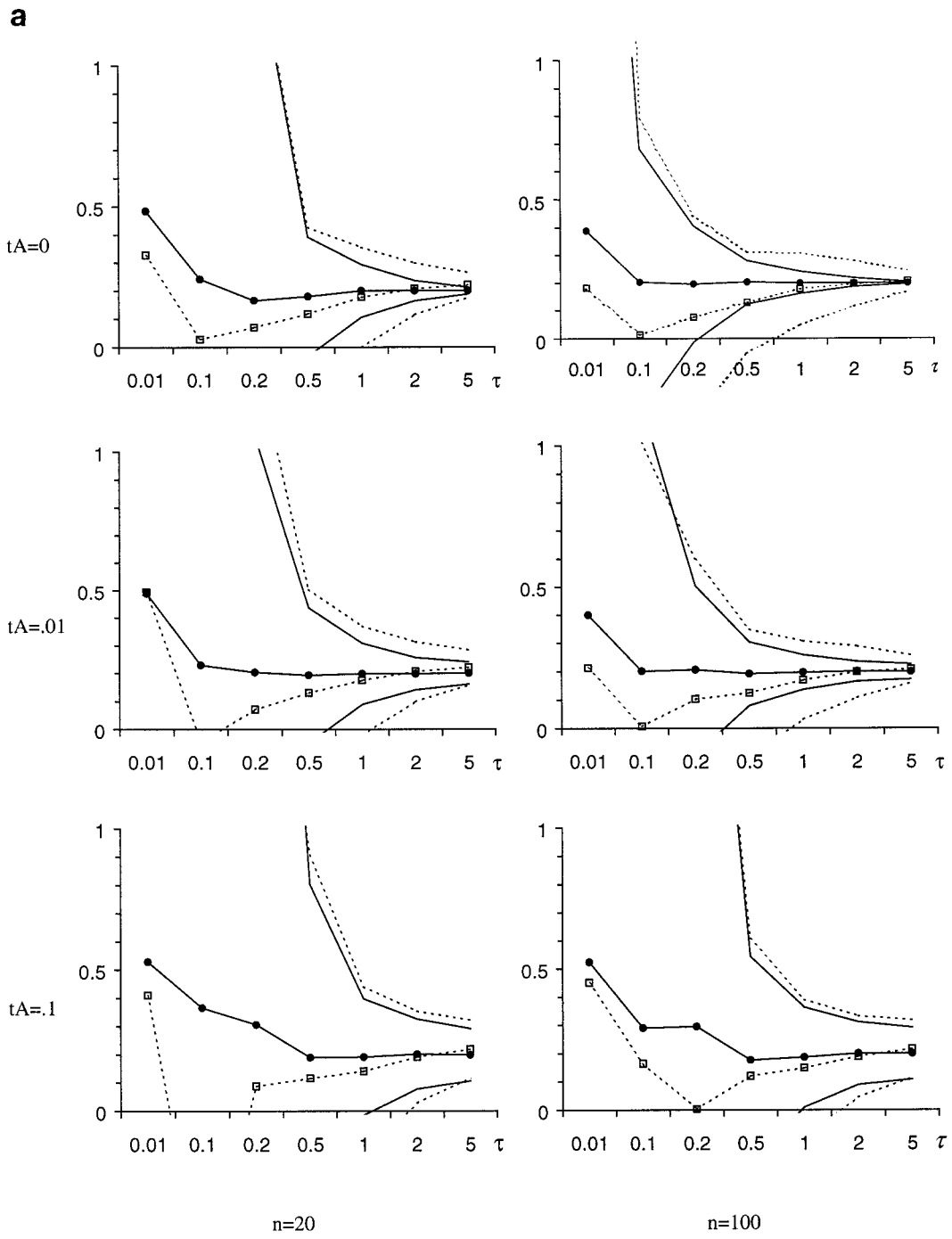

FIG. 2.—Results of the simulation study when the admixture coefficient (μ) is estimated directly from the coalescence times. Each graph reports the average of the coefficients computed from 1,000 iterations (central lines with markers), and the average ±1 SD within the range [0; 1] (external lines), as a function of the divergence time between parental populations (τ). Solid line: m_Y. Dashed line: m_X. t_A = age of the admixture event; n = sample size. (*a*) $\mu = 0.2$. (*b*) $\mu = 0.5$.

estimator of $\hat{t}'_A$ is certainly biased downward. Of course, external information (e.g., historical data) can also be used to estimate $\hat{t}'_A$ if the mutation rate is known.

When the molecular information comes from microsatellite loci, and the single-step stepwise model of mutation is assumed, the mean coalescent times can be estimated from the mutation rate and average squared difference in allele size, S, as $\hat{\bar{t}} = S/2u$ (Slatkin 1995). The single-step stepwise mutation model, under which each mutation can increase or decrease the allele size by a single repeat, has been widely used as an approximation of the process underlying the genetic diversity at microsatellite loci (e.g., Goldstein et al. 1995; Slatkin 1995; Zhivotovsky and Feldman 1995). The expression for m_Y based on microsatellite variability therefore becomes:

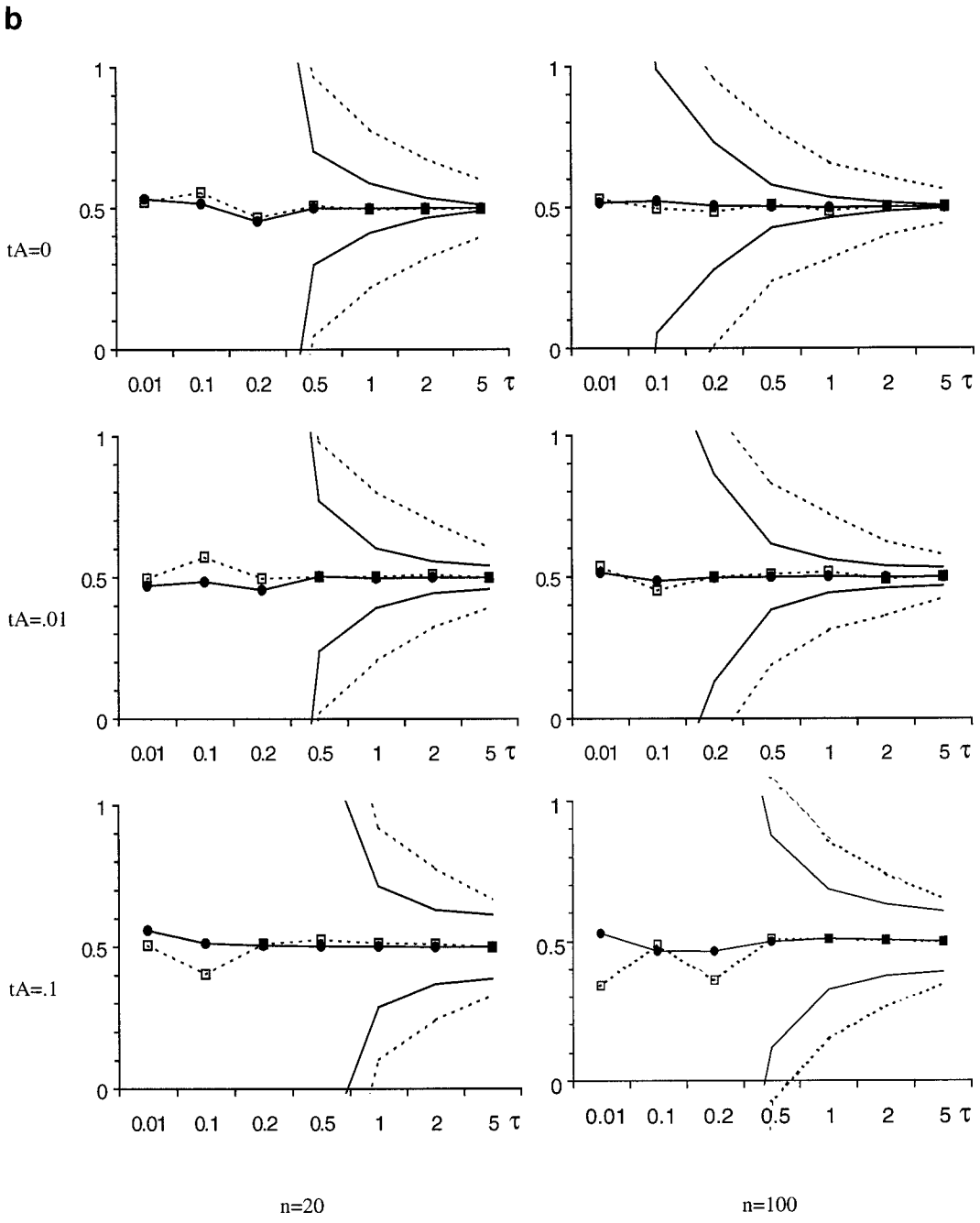

FIG. 2 *(Continued)*

$$m_Y = \frac{c\hat{\bar{S}}_{h1} - d\hat{\bar{S}}_{h2} + d^2 + \hat{\bar{S}}_{12}(\hat{\bar{S}}_{h2} - \hat{\bar{S}}_{h1} + e)}{c^2 + d^2 + 2e\hat{\bar{S}}_{12}}, \quad (4)$$

where $c = \hat{t}'_A + \hat{\bar{S}}_{11}$, $d = \hat{t}'_A + \hat{\bar{S}}_{22}$, $e = \hat{\bar{S}}_{12} - (c + d)$, and $\hat{\bar{S}}_{11}$, $\hat{\bar{S}}_{22}$, $\hat{\bar{S}}_{12}$, $\hat{\bar{S}}_{h1}$, $\hat{\bar{S}}_{h2}$, are the average squared difference in allele size within P_1, within P_2, between P_1 and P_2, between the admixed population and P_1, and between the admixed population and P_2, respectively. In this case, $\hat{t}_A'$ can be replaced by the minimum number of squared difference in allele size observed between a gene drawn from the admixed population and a gene from a parental population.

Multilocus Data

If data at more than one locus are available, a multilocus estimator of the admixture components can be constructed. As is advocated for computing means of ratios (see, e.g., Rice 1995, p. 206) and as has been done for other estimators of ratios, such as multilocus estimates of F_{ST} (e.g., Reynolds, Weir, and Cockerham 1983; Michalakis and Excoffier 1996), we propose to estimate the coalescence times separately for each locus and to use their average values in equations (1) and (2) to compute multilocus weighted averages $\bar{m}_X$ and $\bar{m}_Y$, instead of computing an average of m_X and m_Y values

over all loci. This approach is possible only when all loci have approximately the same mutation rate. Otherwise, the admixture coefficients should be computed separately for classes of loci with similar mutation rates, and a final estimate can be obtained from the average over classes, weighted by the number of loci involved in each class.

Monte Carlo Simulations

The statistical behavior of m_X and m_Y has been analyzed by simulation. Following a coalescent approach (Hudson 1990) and assuming the population model in figure 1, the genealogies of three samples of n genes (one sample from each of the populations P_1, P_2, and P_h) were reconstructed until the most recent common ancestor of all $3n$ genes. Poisson-distributed mutations were then introduced in the tree, assuming either the infinite-site model (ISM), for the simulation of DNA sequences, or the stepwise-mutation model (SMM), for the simulation of microsatellite data. Finally, the estimators m_X and m_Y were computed from equations (1) and (2), either directly with the mean coalescence times from the simulated genealogies or after replacement (as explained above) with the mean number of pairwise differences for the DNA sequences or with the mean square difference between number or repeats for microsatellite data.

We generated 1,000 random trees for each set of parameters to get an empirical distribution of our estimators, from which the means and the standard deviations of m_X and m_Y were computed. In some cases, it was necessary to generate more trees to end up with 1,000 values of m_X, because, as discussed above, there are cases where m_X cannot be computed. The multilocus estimators were computed for different numbers of loci (L). In this case, $1{,}000 \times L$ independent genealogies were generated for each set of parameters, thus assuming free recombination between loci.

Extensive simulations were performed to study the relative behavior of m_X and m_Y under different combinations of n, μ, τ and t_A when the mean coalescence times obtained from these simulated data were used directly to get the different estimators of the admixture coefficients. The molecular estimator with the most desirable statistical properties (which turned out to be m_Y) was then analyzed with the mean coalescence times estimated from the number of nucleotide differences or from the squared microsatellite allele size differences. Different values of the mutation parameter $\theta = 2Nu$ were considered, and the influence of the number of sampled loci was also analyzed. In these cases, we also compared the behavior of the molecular estimator m_Y with the behavior of two admixture coefficient estimators based on the frequencies of alleles present in populations P_1, P_2, and P_h. The first frequency-based estimator has been proposed by Roberts and Hiorns (1965) and will be designated m_R. It is a least-squares estimator. The second conventional estimator has been proposed by Chakraborty et al. (1992) as a closed-form expression of the maximum-likelihood estimator derived by Long (1991). It will be designated m_C.

Results

Results of the Simulation Study

Estimations based on Coalescence Times

The expected values of m_X and m_Y estimated from a single locus, together with their standard deviations, are reported in figure 2*a* for $\mu = 0.2$ and for different combinations of τ, t_A (both in N generations) and sample size n. In general, the standard deviations are very large unless the parental populations have differentiated for about N generations, and they are always larger for m_X than for m_Y. The effect of sample size on the estimators seems less important than the effect of τ and seems mostly appreciable when $t_A = 0$. On the other hand, if admixture is not recent ($t_A > 0$), the variance of the estimators remains large even when the parental populations have been separated for a long time.

In figure 2*a*, there is a bias towards 0.5 for m_Y for very small values of τ, but this bias is almost negligible for $\tau > 0.1$. The cause of this bias can be easily understood: for very small values of τ, the genetic constitution of the parental populations is almost identical, and thus, irrespective of their true contribution to the hybrid, the estimated admixture coefficient will tend to values close to 0.5. The absence of such a bias when $\mu = 0.5$ in figure 2*b* seems to confirm this view.

The behavior of the estimator m_X is not as good as that of m_Y, as m_X appears biased for most of the range of τ, t_A, and n, when $\mu = 0.2$ (fig. 2*a*), and has a larger associated variance. The systematic bias observed for small parental divergence times seems related to the noninclusion of the cases leading to negative arguments under the square root of equation (1). The bias of m_X becomes small only for very large values of τ ($\tau > 2N$ generations), making it practically unsuitable for the analysis of admixture between recently diverged populations. The estimator m_Y, however, performs consistently better than m_X, almost always having smaller variance and a low level of bias, even when parental populations have diverged for only $0.1N$ generations.

When more loci are used to estimate the admixture proportions, the standard deviations of both estimators decrease rapidly (fig. 3, left). The estimator m_Y still performs better than m_X for increasing values of L, with less bias and a smaller associated variance. We can see, for instance, that 10 loci seem sufficient to reduce the standard deviation of the estimator m_Y by one order of magnitude. In that case, the standard deviation evaluated for $\tau = 0.2$ becomes approximately as small as that of a single locus when $\tau = 1$ (see fig. 2*a*). Again, the estimator m_Y is nearly unbiased when more than five loci are considered, whereas m_X seems biased irrespective of the number of loci considered, passing from an underestimation of μ toward a slight overestimation with increasing values of L.

On the right side of figure 3, we report the expected values and standard deviations of the estimators obtained after an a posteriori selection of loci. We wanted to study the effect of selecting loci, whether those clearly indicating that parental populations are differentiated or those for which the hybrid sample presents a genetic

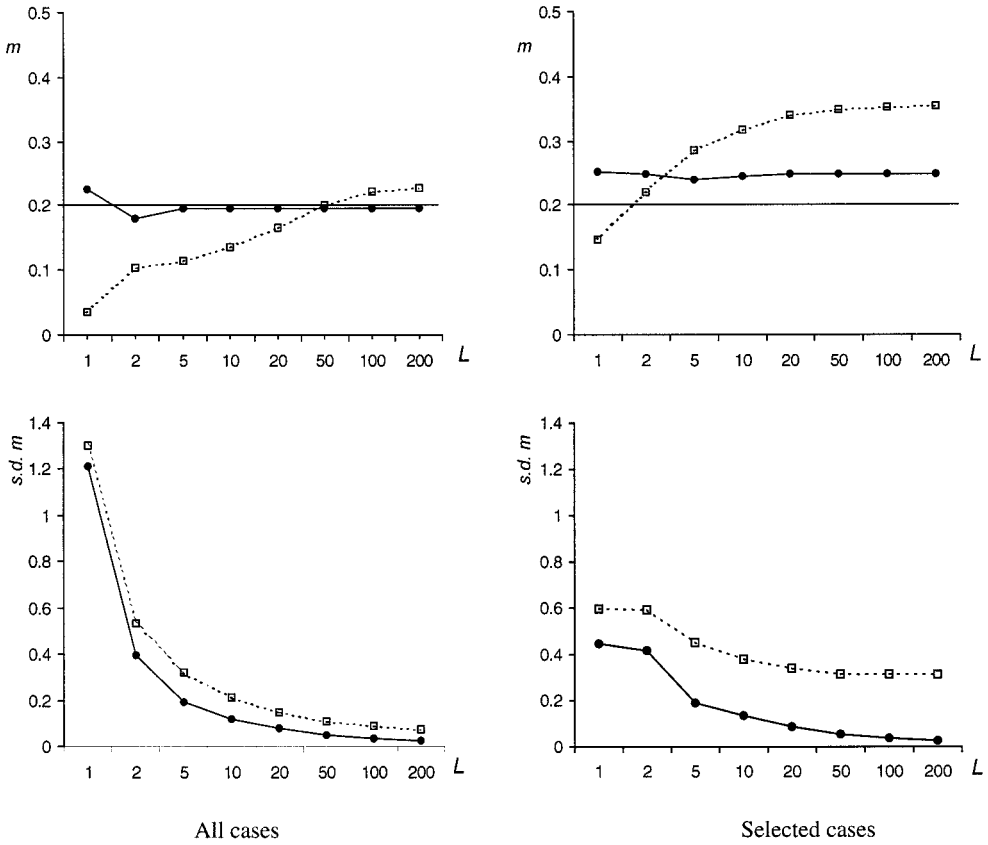

FIG. 3.—Results of the simulations for $\mu = 0.2$, $\tau = 0.2$, $t_A = 0$, and $n = 20$, as a function of different number of loci (L). On the left, there is no selection of loci for the analyses. On the right, loci are incorporated into the analysis only if they meet some criterion defined in the text. Solid line: m_Y. Dashed line: m_X. m = average over 1,000 iterations; $s.d.m$ = standard deviation of 1,000 iterations. The admixture coefficient is estimated directly from the coalescence times.

constitution intermediate between those of the parental populations, on the potential usefulness of those loci in admixture studies. Therefore, we computed multilocus estimators by including only the loci for which the average coalescence time between two genes sampled in different parental populations was equal to or larger than any other average coalescence time (i.e., $\bar{t}_{12} \geq \bar{t}_{11}$, $\bar{t}_{22}$, $\bar{t}_h$, $\bar{t}_{h1}$, $\bar{t}_{h2}$). Our results, reported in figure 3, right, show that this kind of selection does not provide better estimators but rather introduces a systematic bias for both m_X and m_Y. Note, however, that when only one locus is considered for selection, this selection has a positive effect by greatly reducing the variance of the estimator. When more than two loci are considered for selection, the bias remains but the variance increases as compared with the case of an absence of selection of the markers. The effect of the locus selection on m_X is even worse, as the amount of bias seems here to increase with L.

Estimations Based on the Number of Nucleotide Differences

The comparison between m_Y, computed from the average number of nucleotide differences shown in equation (3), and two frequency-based estimators, m_R and m_C, is reported in figure 4 for $\mu = 0.2$, $t_A = 0$, and $n = 50$, with $\tau = 0.2$ or $\tau = 2$. If parental populations have recently diverged ($\tau = 0.2$), frequency-based estimators clearly have a smaller mean square error (MSE; MSE = Variance + bias^2) than m_Y. However, the maximum-likelihood estimator m_C clearly increasingly overestimates μ with larger values of θ. This is probably because of the increase in the number of different alleles and because of the fact that the alleles absent in the hybrid populations are not considered by this estimator. When parental populations are more differentiated ($\tau = 2$), the molecular estimator m_Y has a smaller MSE than the others when $\theta > 5$. It is virtually unbiased for all possible mutation rates, which is clearly not the case for conventional estimators, and particularly for m_C, which is increasingly biased with θ.

In general, the stochastic errors introduced by the mutation process do increase the variance of m_Y, as compared with the case where we estimated m_Y directly from the coalescence times, but this increase becomes almost negligible for large values of θ. For example, when $\tau = 0.2$ and $\theta = 25$, or when $\tau = 2$ and $\theta > 5$, the standard deviations of m_Y are almost identical to those estimated, assuming that the coalescence times were directly available.

Estimations Based on the Number of Repeat Differences

In figure 5, we show the results of estimating admixture coefficients from microsatellite-like single locus

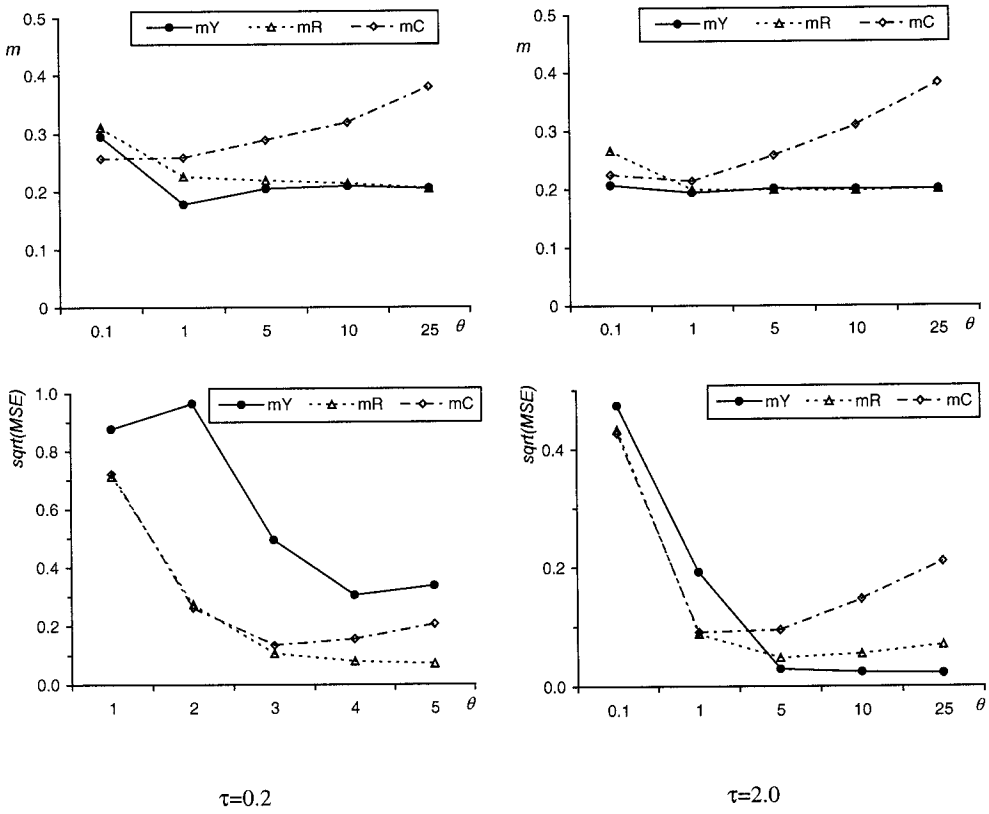

FIG. 4.—Results of the simulations for $\mu = 0.2$, $t_A = 0$, $n = 50$, $\tau = 0.2$ (left) or $\tau = 2$ (right), and for different values of the mutation parameter $\theta = 2Nu$ (N is the population size, and u is the global mutation rate). Here, the admixture coefficient is estimated from the average pairwise differences between 400-bp sequences. m = average over 1,000 iterations; *MSE* = mean square error of 1,000 iterations; (solid circle intersected by solid line) = m_Y; (open triangle intersected by dashed line) = m_R; (open diamond intersected by dashed and dotted line) = m_C.

data, using our molecular estimator m_Y, and two conventional estimators based only on allele frequencies. While having less bias than conventional estimators, the variance of m_Y is much larger than that of conventional estimators, contributing to the large MSE of m_Y over the whole range of parameters simulated here. Thus, in contrast with DNA data, and even though they appear to be biased, the conventional estimators here seem superior to ours when applied to a single locus, even in cases of large differentiation time between parental populations and high mutation rates at microsatellite loci. Note that the bias of m_R and m_C increases with the mutation rate, reflecting further departure from the assumption of allelic identity-by-descent on which these estimators are based. In contrast, the amount of bias of m_Y seems independent of the mutation rate and seems much less than that for the conventional estimators. These results suggest that for loci with very high mutation rates, our estimator may begin to perform quite well because of the high bias of conventional estimators. Finally, it should be noted that when the admixture coefficient μ is set to 0.8, m_R and m_C underestimate μ (results not shown). In other words, frequency-based estimators seem biased towards the central value of 0.5.

When several loci are used simultaneously to estimate the admixture coefficient from microsatellite-like data (fig. 6), the molecular estimator m_Y becomes essentially unbiased and its variance approaches that of conventional estimators, whereas m_R and m_C remain biased upward. It can be seen that the MSE of m_Y eventually becomes smaller than that of either m_R or m_C with increasing number of loci surveyed. For short divergence times between parental populations ($\tau < 0.2$), 50 loci need to be surveyed for our estimator to perform better than conventional ones, whereas only 5 to 10 loci are required when $\tau > 2$. Conventional estimators appear very sensitive to changes in sample size, as they show large bias with small samples. This is especially true for the maximum-likelihood estimator m_C, whose bias decreases remarkably when the sample size increases from 20 to 50.

Application to North American Wolflike Canids

We have applied the estimator m_Y to a study of North American wolflike Canids analyzed for microsatellite loci by Roy et al. (1994). In this paper, 10 loci were typed in seven populations of gray wolf, six populations of coyote, and one captive population of red wolf. Hybridization between gray wolf and coyote seemed to have occurred in two gray wolf populations and in two coyote populations, and past hybridization between these two species has been proposed to account for the origin of the red wolf (Roy et al. 1994).

Here we have pooled the samples of nonhybridized gray wolf populations and the samples of nonhybridized coyote populations to act as the parental population sam-

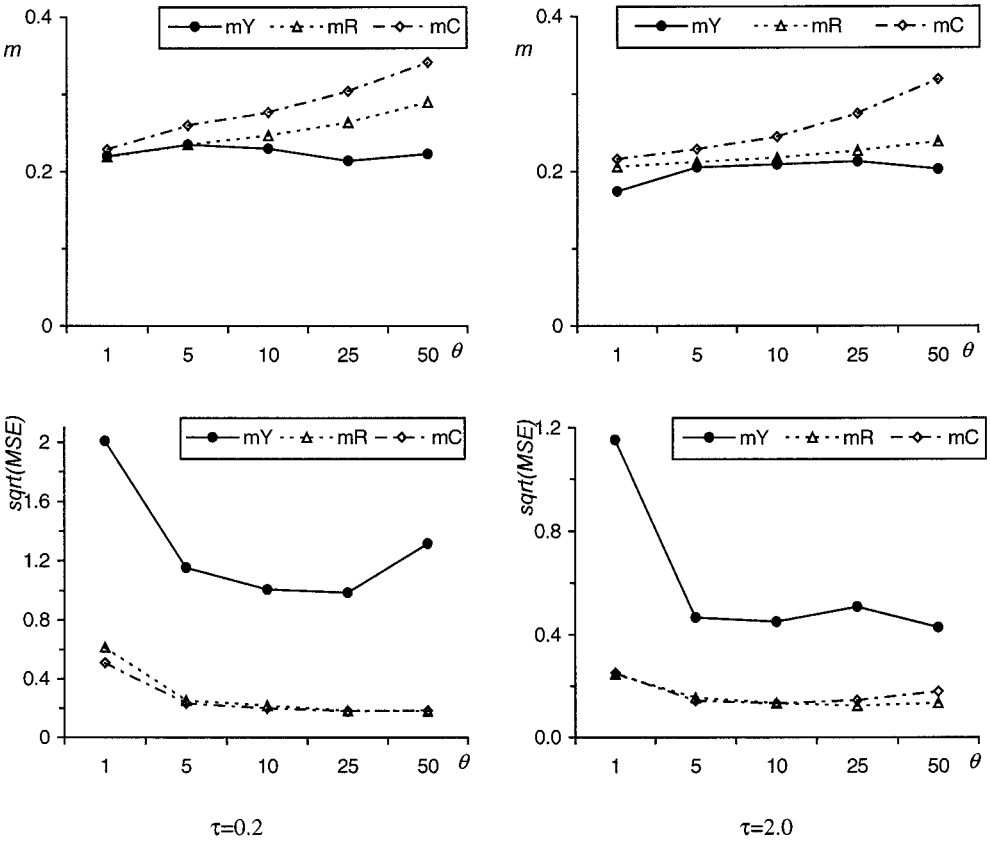

FIG. 5.—Results of the simulations for $\mu = 0.2$, $t_A = 0$, $n = 50$, $\tau = 0.2$ (left) or $\tau = 2$ (right), and for different values of the mutation parameter $\theta = 2Nu$ (N is the population size, and u is the global mutation rate). Here, the admixture coefficient is estimated from the average squared difference in allele size at a microsatellite-like locus following a single-step stepwise mutation model. m = average over 1,000 iterations; MSE = mean square error of 1,000 iterations; (solid circle intersected by solid line) = m_Y; (open triangle intersected by dashed line) = m_R; (open diamond intersected by dashed and dotted line) = m_C.

ples (P_1 and P_2 in the model of fig. 1). The genetic contribution of the gray wolf was then estimated in three hybrid samples: the hybridized gray wolf sample, the hybridized coyote sample, and the red wolf sample. The hybridized gray wolf and coyote samples were obtained by pooling the data of the corresponding original samples.

The sampling error of m_Y was estimated by bootstrap technique (Efron 1982): for each parental and hybrid sample and independently for each locus, 1,000 random samples of size identical to the original were generated by drawing, with replacement, the chromosomes from the original samples. The estimators m_Y, m_R, and m_C were then computed in each random sample and used to get their bootstrap average and standard deviation.

The results presented in table 1 show that the estimated contribution of the gray wolf to the hybridized gray wolf populations and to the hybridized coyote populations are approximately 50% and 10%–15%, respectively. The average bootstrap values of the three estimators are similar, with standard deviations between 0.05 and 0.14.

Interestingly, the inferred genetic contribution of the gray wolf to the red wolf gene pool was estimated to lie between 17% and 33% by frequency-based estimators, but we obtained a negative contribution (−33%) using the molecular estimator m_Y. This result could be the simple consequence of the large variance associated with the estimator, based on the average squared difference in allele size. However, our simulation results suggest that if the divergence time between parental populations is longer than $0.2N$ generations (which is probably the case; see, e.g., Lehman et al. 1991), then a negative admixture coefficient equal to or smaller than −0.33 is very unlikely ($P < 0.05$) to result from data on 10 microsatellite loci. In other words, the present microsatellite data do not seem compatible with the hypothesis that red wolves originated through hybridization between the gray wolf and the coyote during the past 300 years, as previously advocated (Roy et al. 1994). Alternatively, the admixture and/or the mutation models on which our methodology is based might not reflect the true processes that affected the genetic variability of wolflike Canids.

In order to see if negative admixture coefficients could be obtained for m_Y under alternative evolutionary models, we have simulated the evolution of gray wolf, red wolf, and coyote populations under three competing historical demographic scenarios, as described in figure 7. The results of the simulations of these three models are reported in table 2. Negative admixture coefficients

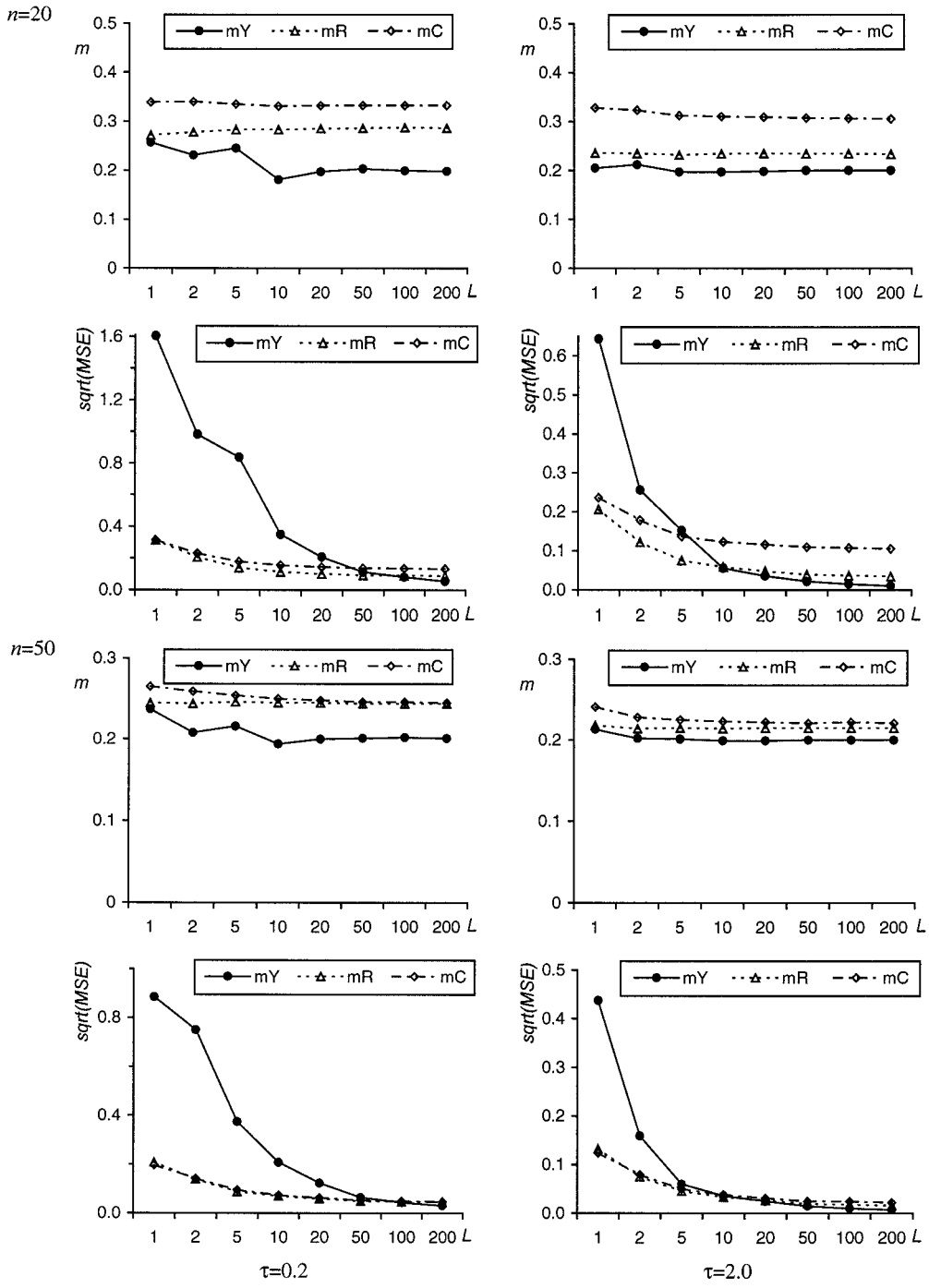

FIG. 6.—Results of the simulations for $\mu = 0.2$, for $t_A = 0$, $n = 20$ or $n = 50$, and $\tau = 0.2$ (left) or $\tau = 2$ (right), as a function of the number of microsatellite loci surveyed. For each locus, the mutation rate θ was set to 10. m = average over 1,000 iterations; *MSE* = mean square error of 1,000 iterations; (solid circle intersected by solid line) = m_Y; (open triangle intersected by dashed line) = m_R; (open diamond intersected by dashed and dotted line) = m_C.

are never observed in a pure admixture model (model a, fig. 7). They are seen at low frequencies for our molecular estimator m_Y in a model where an independent red wolf population would have incorporated gray wolf and coyote genes in the recent past (model c, fig. 7). Interestingly, negative estimates become quite common in a scenario where the red wolf and the coyote populations are sister species that diverged only recently (model b, fig. 7). In that case, 31.7% of the simulations led to negative m_Y estimates, whereas negative m_C and m_R estimates were only observed in less than 3% of the cases. The results for the red wolf, shown in table 1 (regarded as a hybrid between the gray wolf and the coyote), are thus quite compatible with model b. This does not necessarily mean that model b is true for wolf-like Canids nor that model b is the most likely one.

Table 1
Contribution of the Grey Wolf Population to Different Hybrid Populations

Case	Estimator	Estimated admixture coefficient	Bootstrap average[a]	Bootstrap standard deviation
Gray wolf hybrid	m_Y	0.468	0.478	0.120
	m_R	0.481	0.486	0.051
	m_C	0.641	0.507	0.130
Coyote hybrid	m_Y	0.157	0.147	0.137
	m_R	0.091	0.107	0.048
	m_C	0.112	0.137	0.089
Red wolf hybrid	m_Y	−0.343	−0.333	0.172
	m_R	0.170	0.187	0.050
	m_C	0.331	0.258	0.190

[a] Gray wolf contribution to the admixed hybrid population, computed as an average over 1,000 random bootstrap samples.

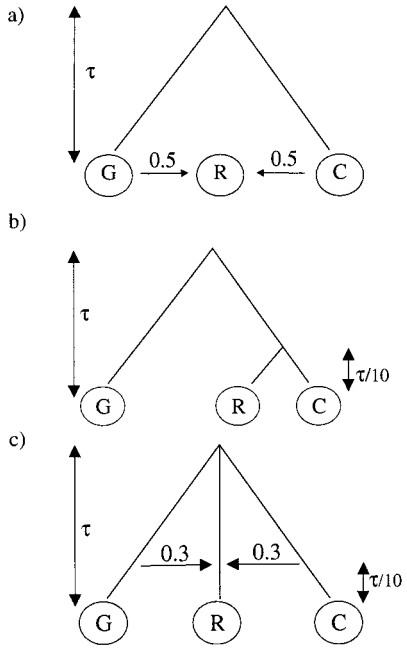

FIG. 7.—Simple alternative historic demographic scenarios considered for the origin of the red wolf. (*a*) The red wolf species as a hybrid of the gray wolf and the coyote. Both parental populations have contributed equally to the red wolf population. (*b*) The red wolf as a sister species to the coyote. The separation of the gray wolf lineage is 10 times older than the separation of the red wolf lineage. (*c*) The gray wolf, the red wolf, and the coyote diverged simultaneously from a common stock τ generations ago. $\tau/10$ generations ago, the red wolf population received 30% of its current genes from both gray wolf and coyote populations.

However, it shows that a plausible departure from a pure admixture model can generate values for the three estimators close to our observation. There is thus no need to assume a flaw in our estimator m_Y. On the contrary, it appears interesting that this estimator is more sensitive to departures from the pure admixture model than are estimators based on allele frequencies. This suggests that negative estimates of m_Y obtained from several independent loci are indicative of evolutionary processes other than a simple admixture event.

Discussion

In this paper, we have shown how molecular data can be simply used to efficiently estimate admixture proportions. Two new estimators have been introduced that explicitly consider not only the frequencies of different alleles, but also their level of divergence. While the molecular estimators presented here have been derived for haploid populations of size N, they can be used as well in diploid populations of the same size, with N being replaced by the number of gene copies, $2N$. The admixture model we have adopted here (fig. 1) may seem impractical, as the admixture event is usually not instantaneous and can last for much more than one generation. Although we did not study this case of long-lasting admixture, such a departure from our model should not drastically alter our conclusions if the number of generations taken to form the hybrid population is small compared with the divergence time between the parental populations. While these molecular estimators have been derived in the context of admixture analysis, we note that our estimator m_Y could easily be extended to a different demographic model involving several parental populations contributing to the hybrid, thus making it suitable also for mixed stock analysis based on molecular markers (Ferris and Berg 1987; Xu, Kobak, and Smouse 1994; Brown et al. 1996).

The first estimator we considered, here called m_X, is based mainly on the mean coalescence time within a hybrid population. The second estimator, m_Y, depends instead on mean coalescence times, both within the parental populations and between the parental and the hybrid populations. Despite its simplicity, m_Y seems to perform better than m_X under most conditions. The variance of m_X is always larger than that of m_Y, probably because the former estimator does not consider directly the highly informative distances between hybrid and parental populations. Moreover, m_X cannot be computed for all data sets, which is not only a serious limitation in itself but also a possible source of bias. The use of m_Y instead of m_X therefore seems highly recommendable when molecular information is available; this is why we have focused on m_Y when comparing our molecular estimator to conventional ones.

One of the most important factors affecting the efficiency of m_Y in recovering the true admixture coefficient μ is the level of divergence between the parental populations. If the parental populations have been isolated for only a few generations (say, $\tau < 0.1$), our simulations show that m_Y is biased toward 0.5 and has a huge associated variance (fig. 2). In such cases, these unfortunate properties cannot be alleviated by simply increasing the sample size (fig. 2). Note, however, that the bias described here should also be expected for any other estimator. This is because an absence of differentiation between the parental populations will prevent a correct inference of their respective contribution to the hybrid, whatever the method used to estimate this contribution. While quite obvious, this bias does not seem to have been described before for conventional estimators based on allele frequencies. In addition, because the estimator m_Y depends on the comparison of average co-

Table 2
Estimated Proportions of Grey Wolf Genes in the Red Wolf Populations Obtained Under Three Possible Historical Demographic Scenarios Accounting for the Origin of Wolflike Canids

	m_Y			m_R			m_C		
Demographic Model[a]	Average[b]	SD	Negative Estimates (%)	Average[b]	SD	Negative Estimates (%)	Average[b]	SD	Negative Estimates (%)
a)	0.500	0.033	0.0	0.502	0.029	0.0	0.501	0.038	0.0
b)	0.052	0.129	31.7	0.184	0.101	2.8	0.437	0.136	0.1
c)	0.490	0.222	1.6	0.501	0.110	0.0	0.500	0.122	0.0

Note.—In each case, the admixture coefficients were estimated from 10 independent microsatellite loci. The simulation conditions were the following: sample size = 50 haploid individuals per population; divergence time $\tau = 2.0$; and mutation parameter $\theta = 10$ per locus (single-step mutation model assumed).

[a] The demographic models are those reported in figure 7.

[b] Obtained from 1,000 simulations.

alescence within and between populations, which are quantities known to have large variances (Tajima 1983), the information contained in average coalescence times is limited by the stochasticity of the genealogical process. Therefore, the applicability of m_Y to single-locus data sets seems restricted to two particular situations. First, when parental populations have been isolated for more than N generations (or $2N$ for diploids), the variance of m_Y becomes reasonably low, making it well adapted for estimating introgression coefficients between subspecies. Second, in the case of a relatively short divergence time between parental populations and when the mutation rate is extremely high at a given locus following the stepwise mutation model, the conventional estimators become increasingly biased (fig. 5), whereas m_Y remains unbiased despite a large variance, making it more likely to be suitable than conventional estimators. High mutation rates and quite long divergence times between parental populations are the conditions under which our molecular estimator should perform best as compared with conventional estimators, whatever the mutation model (infinite-site, fig. 4; or stepwise, fig. 5). Under these conditions, the mutational process probably becomes more important and informative than genetic drift (which, supposedly, is the main process affecting the classical estimators). In other words, if the parental populations are different in mutational terms, m_Y should perform better, whereas classical estimators should perform better when the parental populations had the time to differentiate only through genetic drift.

As observed before (e.g., Thompson 1973; Wijsman 1984; Chakraborty 1986; Cavalli-Sforza, Menozzi, and Piazza 1994), the estimation of admixture proportions when the hybrid population is not recent ($t_A > 0$) can be very problematic, and this effect also applies to m_Y. Our simulations confirm that stochastic factors occurring after the admixture event, such as genetic drift and mutations, lead to biased estimations of admixture coefficients, much more so if parental populations have diverged recently (fig. 2). Roughly speaking, the use of m_Y should be avoided if the admixture is older than $0.01N$ generations. It should be noted, however, that the estimator m_Y, unlike most conventional estimators, includes the age of the admixture event in its computational expression. This implies that the effect of postadmixture stochastic factors can be explicitly accounted for if one gets an external estimation of the age of the admixture event.

When more independent loci are considered, both the bias (observed for small τ) and the variance of m_Y rapidly decrease (fig. 3). Multilocus data therefore seem suitable for estimating admixture proportions with m_Y, and they really are necessary when the parental populations are not strongly differentiated. For example, if the parental populations have remained isolated for only $0.2N$ generations and the admixture event is recent, 10 loci are enough to reduce the standard deviation of the estimator to the level obtained in the single-locus case for parental populations separated for N generations. In general, however, it seems more difficult to reduce variance of the estimator by studying more loci when the admixture event is not recent. For instance, when $t_A = 0.1$ and $\tau = 5$, five loci are enough to obtain a variance of m_Y similar to the single-locus case for $t_A = 0.01$, but more than 50 loci are needed to approach the variance obtained when $t_A = 0$ for the single-locus case (data not shown). Interestingly, our results strongly suggest that if one does not want to introduce a bias, all loci should be considered when inferring admixture coefficients, without a priori consideration of their usefulness in such studies (fig. 3). However, we have studied here cases where differences in the amount of observed polymorphism were due to the stochasticity of the coalescent process and not due to differences in mutation rates among loci. Although this latter case has not been explicitly studied here, it might be safer to estimate admixture coefficients separately for classes of loci with similar mutation rates and then to compute an average estimate weighted by the number of loci involved in each class. The combination of information gathered from nuclear autosomal DNA, cytoplasmic DNA, and sex chromosomes may prove difficult, as individuals from different sexes may not have contributed equally to the admixed population. The comparison of estimates obtained from maternally and paternally transmitted genes may, however, provide interesting insights on the admixture process.

Generally, m_Y is clearly less biased than frequency-based estimators for most of the cases considered here. In particular, when highly variable markers are used (such as mitochondrial DNA or microsatellites), the two

Table 3
Effect of Unequal Population Sizes on Estimators of Admixture Coefficients

	m_Y		m_R		m_C	
	Average	sqrt(MSE)	Average	sqrt(MSE)	Average	sqrt(MSE)
$N_2 = 15{,}000$						
$t_A = 0.00$	0.201	0.035	0.217	0.049	0.249	0.085
$t_A = 0.01$	0.203	0.121	0.232	0.194	0.331	0.288
$N_2 = 50{,}000$:						
$t_A = 0.0$	0.194	0.087	0.268	0.091	0.243	0.082
$t_A = 0.1$	0.198	0.201	0.271	0.235	0.237	0.334

NOTE.—The average values of the estimators were obtained after 1,000 simulations. The true admixture coefficient was set to $\mu = 0.2$. In all cases, the population size of the admixed population was fixed to $N_h = 10{,}000$ haploid individuals, like that of the ancestral population ($N = 10{,}000$). The size of the first parental population was set to $N_1 = 5{,}000$. The divergence time between parental populations was set to $\tau = 2.0N$ generations. The mutation parameter θ for a DNA sequence of 400 bp is here equal to 5, and the sample sizes are all equal to 50.

classical estimators often appear biased toward 0.5, whereas m_Y is always unbiased. This result might be explained by the fact that if the number of alleles is high, some of them will be sampled in the hybrid population but not in the parental populations. They therefore will not be informative for estimators based on frequencies only, whereas they will still be informative for the molecular estimator m_Y, which takes into consideration their molecular distance from other alleles.

The admixture model we considered here assumed that the sizes of all populations were identical, which may not hold true in practical cases. We therefore carried out some additional simulations to analyze the influence of unequal population sizes on the admixture estimators. The results, summarized in table 3, show that the MSE of all estimators tends to increase (even if not dramatically) with the difference between the sizes of the parental populations. As for the equal population size simulations, however, the molecular estimator m_Y had smaller bias than the frequency-based estimators; this was also the case when the parental population sizes differed by an order of magnitude.

Finally, it is important to remember that any method of estimation of admixture proportions relies on the correct identification of the hybrid and parental populations. A potential hybrid population can sometimes be detected by its genetic heterogeneity or by its intermediacy between the putative parental populations, but the parental populations are almost always defined a priori and are assumed to be the true parental populations. Although we have not specifically addressed the problem of deciding whether a genetically intermediate population really is an admixed population, the results of an admixture analysis obtained on populations that have experienced evolutionary processes other than a pure admixture model (table 2) suggest that the estimator m_Y is quite sensitive to departure from the admixture model. This property, which certainly deserves further analysis, might prove useful in preventing biologically meaningless computations of the admixture proportions.

Acknowledgments

Thanks to Stefan Schneider for stimulating comments on the use of least-squares estimators. We are also grateful to Monty Slatkin, Naruya Saitou, and two anonymous reviewers for their helpful comments on an earlier version of the manuscript. G.B. was partly supported by Swiss NSF grant No. 32-47053.96 to L.E. A computer program to compute the admixture estimators and bootstrap standard deviations is available from G.B. upon request.

APPENDIX

Following the population differentiation scheme shown in figure 1, the coalescence time of two genes drawn from the admixed population P_h can be estimated by assuming a continuous time approximation (see, e.g., Hudson 1990) and by considering three distinct periods in the coalescent process, as

$$\bar{t}_h = \int_{t=0}^{t_A} \frac{1}{N} e^{-t/N} t\, dt + (\mu^2 + (1-\mu)^2) \int_{t=t_A}^{t_A+\tau} \frac{1}{N} e^{-t/N} t\, dt + (\mu^2 + (1-\mu)^2) \int_{t=t_A+\tau}^{\infty} \frac{1}{N} e^{-t/N} t\, dt + (1 - \mu^2 - (1-\mu)^2) \int_{t=t_A+\tau}^{\infty} \frac{1}{N} e^{-(t-\tau)/N} t\, dt, \quad \text{(A1)}$$

where μ is the contribution of P_1 to the hybrid population. The first term of the second member of equation (A1) considers the coalescent events that will occur from the present time until the admixture event. The second term considers the coalescent events during the period when the parental populations were kept separated. During that time, the coalescent events can occur only between genes that comigrated in the same parental population. Finally, the third and fourth terms consider the coalescences occurring in the ancestral population. These events have different probabilities depending on whether the two genes comigrated in the same population or not. Equation (A1) reduces to

$$\bar{t}_h = N + 2\mu(1-\mu)\tau e^{-t_A/N}, \quad \text{(A2)}$$

showing that the mean coalescence time in an admixed population is increased by the factor $2\,\mu(1-\mu)\,\tau\,e^{-tA/N}$, as compared with $\bar{t}_0 = N$, the mean coalescence time in an isolated and stationary population of size N (Kingman 1982). Note that by a similar way of reasoning, we can obtain the second moment of the coalescent time, and, therefore, its variance, as

$$V_h(t) = N^2 + 2\mu(1 - \mu)\tau e^{-t_A/N} \cdot (2t_A + \tau - 2\mu(1 - \mu)\tau e^{-t_A/N}). \quad \text{(A3)}$$

Solving equation (A2) for μ leads to

$$\mu_X = \frac{1}{2} \pm \frac{\sqrt{\tau e^{-t_A/N}(\tau e^{-t_A/N} + 2N - 2\bar{t}_h)}}{2\tau e^{-t_A/N}}. \quad \text{(A4)}$$

To derive another estimator of μ, we consider now the coalescent times of genes drawn from the admixed populations and from the parental populations P_1 and P_2. A gene sampled from the admixed population may originally come from P_1 or P_2, with probabilities of μ and $1-\mu$, respectively. In the first case, its mean coalescence time with a gene from P_1 is simply the mean coalescence time between two genes drawn from P_1 ($\bar{t}_{11}$), plus the time elapsed since the admixture (t_A). In the second case, its mean coalescence time with a gene from P_1 is just the mean coalescence time between a gene drawn from P_1 and a gene drawn from P_2 ($\bar{t}_{12}$), a quantity equal to $\bar{t}_{12} = t_A + \tau + \bar{t}_0$, the sum of time since admixture, divergence time, and the mean coalescence time between two genes drawn from the ancestral population. Therefore, the mean coalescence time $\bar{t}_{h1}$ between a gene sampled in the hybrid population P_h and a gene sampled in population P_1 is simply

$$\bar{t}_{h1} = \mu(\bar{t}_{11} + t_A) + (1 - \mu)\bar{t}_{12}. \quad \text{(A5)}$$

Similarly, $\bar{t}_{h2}$ is simply

$$\bar{t}_{h2} = \mu\bar{t}_{12} + (1 - \mu)(\bar{t}_{22} + t_A). \quad \text{(A6)}$$

Much in the same way as has been done for allele frequencies (Roberts and Hiorns 1965; Chakraborty 1986), a least-squares estimator of μ resulting from the combination of (A5) and (A6) can be found by minimizing the MSE

$$(\bar{t}_{h1} - \mu(\bar{t}_{11} + t_A) - (1 - \mu)\bar{t}_{12})^2 + (\bar{t}_{h2} - (1 - \mu)(\bar{t}_{22} + t_A) - \mu\bar{t}_{12})^2, \quad \text{(A7)}$$

and solving for μ. This leads to

$$\mu_Y = \frac{c\bar{t}_{h1} - d\bar{t}_{h2} + d^2 + \bar{t}_{12}(\bar{t}_{h2} - \bar{t}_{h1} + e)}{c^2 + d^2 + 2e\bar{t}_{12}}. \quad \text{(A8)}$$

where $c = t_A + \bar{t}_{11}$, $d = t_A + \bar{t}_{22}$, and $e = \bar{t}_{12} - (c + d)$. We note here that by taking into account mean coalescence times between the hybrid and additional populations and adding terms to equation (A7), this least-squares estimator can be readily extended to the case where more than two populations have contributed to the hybrid population.

LITERATURE CITED

Bernstein, F. 1931. Die geographische Verteilung der Blutgruppen und ihre anthropologische Bedeutung. Pp. 227–243 *in* Comitato Italiano per o studio dei problemi della populazione. Instituto Poligrafico dello Stato, Roma.

Brown, B. L., J. M. Epifanio, P. E. Smouse, and C. J. Kobak. 1996. Temporal stability of mtDNA haplotype frequencies in American shad stocks: to pool or not to pool across years? Can. J. Fish. Aquat. Sci. **53**:2274–2286.

Cavalli-Sforza, L. L., and W. F. Bodmer. 1971. The genetics of human populations. W. H. Freeman and Co., San Francisco.

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza. 1994. The history and geography of human genes. Princeton University Press, Princeton, N.J.

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

Chakraborty, R., M. I. Kamboh, M. Nwankwo, and R. E. Ferrell. 1992. Caucasian genes in American Blacks: new data. Am. J. Hum. Genet. **50**:145–155.

Efron, B. 1982. The jackknife, the bootstrap and other resampling plans. Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia.

Ferris, S. D., and W. J. Berg. 1987. The utility of mitochondrial DNA in fish genetics and fishery management. Pp. 277–299 *in* N. Ryman and F. Utter, eds. Population genetics and fisheries management. University of Washington Press, Seattle, Wash.

Goldstein, D. B., A. R. Linares, L. L. Cavalli-Sforza, and M. W. Feldmann. 1995. An evaluation of genetic distances for use with microsatellite loci. Genetics **139**: 463–471.

Hammer, M. F., and S. Horai. 1995. Y chromosomal DNA variation and the peopling of Japan. Am. J. Hum. Genet. **56**:951–962.

Horai, S., K. Murayama, K. Hayasaka, S. Matsubayashi, Y. Hattori, G. Fucharoen, S. Harihara, K. Sook Park, K. Omoto, and I.-H. Pan. 1996. mtDNA polymorphism in East-Asian populations, with special reference to the peopling of Japan. Am. J. Hum. Genet. **59**:579–590.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. Pp. 1–44 *in* D. J. Futuyma and J. D. Antonovics, eds. Oxford surveys in evolutionary biology. Oxford University Press, New York.

Kingman, J. F. C. 1982. On the genealogy of large populations. J. Appl. Proba. **19A**:27–43.

Lehman, N., A. Eisenhawer, K. Hansen, L. D. Mech, R. O. Peterson, P. J. P. Gogan, and R. K. Wayne. 1991. Introgression of coyote mitochondrial DNA into sympatric North American gray wolf populations. Evolution **45**: 104–119.

Long, J. C. 1991. The genetic structure of admixed populations. Genetics **127**:417–428.

Michalakis, Y., and L. Excoffier. 1996. A generic estimation of population subdivision using distances between alleles with special reference to microsatellite loci. Genetics **142**:1061–1064.

Nei, M., and W. H. Li. 1973. Linkage disequilibrium in subdivided populations. Genetics **75**:213–219.

Reynolds, J., B. S. Weir, and C. C. Cockerham. 1983. Estimation for the coancestry coefficient: basis for a short-term genetic distance. Genetics **105**:767–779.

Rice, J. A. 1995. Mathematical statistics and data analysis. Duxburry Press, Belmont, Calif.

Roberts, D., and R. Hiorns. 1965. Methods of analysis of the genetic composition of a hybrid population. Hum. Biol. **37**:38–43.

Roy, M. S., E. Geffen, D. Smith, E. A. Ostrander, and R. K. Wayne. 1994. Patterns of differentiation and hybridization in North American wolflike canids, revealed by analysis of microsatellite loci. Mol. Biol. Evol. **11**: 553–570.

SLATKIN, M. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics **139**: 457–462.

TAJIMA, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics **105**:437–460.

TAKAHATA, N., and M. NEI. 1985. Gene genealogy and variance of interpopulational nucleotide differences. Genetics **110**:325–344.

THOMPSON, E. A. 1973. The Icelandic admixture problem. Ann. Hum. Genet. **37**:69–80.

WIJSMAN, E. M. 1984. Techniques for estimating genetic admixture and application to the problem of the origin of the Icelanders and the Ashkenazi Jews. Hum. Genet. **67**:441–448.

XU, S., C. J. KOBAK, and P. E. SMOUSE. 1994. Constrained least squares estimation of mixed population stock composition from mtDNA haplotype frequency data. Can. J. Fish. Aquat. Sci. **51**:417–425.

ZHIVOTOVSKY, L. A., and M. W. FELDMAN. 1995. Microsatellite variability and genetic distances. Proc. Natl. Acad. Sci. USA **92**:11549–11552.

NARUYA SAITOU, reviewing editor

Accepted July 13, 1998