Microsatellite null alleles and estimation of population differentiation

Marie Pierre Chapuis, Arnaud Estoup

RESEARCH ARTICLES

Microsatellite Null Alleles and Estimation of Population Differentiation

Marie-Pierre Chapuis†‡ and Arnaud Estoup**

*Centre de Biologie et de Gestion des Populations, Institut National pour la Recherche Agronomique, Campus International de Baillarguet, Montferrier/Lez, France; †Génétique et Evolution des Maladies Infectieuses, UMR 274 CNRS-IRD, Montpellier, France; and ‡Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Campus International de Baillarguet, Montpellier, France

Microsatellite null alleles are commonly encountered in population genetics studies, yet little is known about their impact on the estimation of population differentiation. Computer simulations based on the coalescent were used to investigate the evolutionary dynamics of null alleles, their impact on F_{ST} and genetic distances, and the efficiency of estimators of null allele frequency. Further, we explored how the existing method for correcting genotype data for null alleles performed in estimating F_{ST} and genetic distances, and we compared this method with a new method proposed here (for F_{ST} only). Null alleles were likely to be encountered in populations with a large effective size, with an unusually high mutation rate in the flanking regions, and that have diverged from the population from which the cloned allele state was drawn and the primers designed. When populations were significantly differentiated, F_{ST} and genetic distances were overestimated in the presence of null alleles. Frequency of null alleles was estimated precisely with the algorithm presented in Dempster et al. (1977). The conventional method for correcting genotype data for null alleles did not provide an accurate estimate of F_{ST} and genetic distances. However, the use of the genetic distance of Cavalli-Sforza and Edwards (1967) corrected by the conventional method gave better estimates than those obtained without correction. F_{ST} estimation from corrected genotype frequencies performed well when restricted to visible allele sizes. Both the proposed method and the traditional correction method have been implemented in a program that is available free of charge at http://www.montpellier.inra.fr/URLB/. We used 2 published microsatellite data sets based on original and redesigned pairs of primers to empirically confirm our simulation results.

Introduction

Microsatellites are popular and versatile molecular markers for addressing questions in population genetics and evolution (Estoup and Angers 1998). Observed microsatellite alleles are DNA fragments of different sizes detected by initial amplification using polymerase chain reaction (PCR) and visualization via electrophoresis. Size polymorphism reflects variation in the number of repeats of a simple DNA sequence (2–6 bases long). However, sequencing studies indicate that changes in flanking region sequences also occur at a nonnegligible rate (e.g., Angers and Bernatchez 1997; Grimaldi and Crouau-Roy 1997). Such variation in the nucleotide sequences of flanking regions may prevent the primer annealing to template DNA during amplification of the microsatellite locus by PCR, resulting in a null allele. The molecular origin of null alleles (substitution and indel mutations) resulting from polymorphism in the annealing region has been assessed directly by sequencing the annealing sites of microsatellite locus primers for both null and visible alleles (Callen et al. 1993). Other possible causes of microsatellite null alleles include the preferential amplification of short alleles (due to inconsistent DNA template quality or quantity) or slippage during PCR amplification (Gagneux et al. 1997; Shinde et al. 2003). These technical problems associated with amplification will not be considered here.

The presence of microsatellite null alleles has been reported frequently in PCR primer characterization and in population genetics studies (Dakin and Avise 2004). Although microsatellite null alleles have been found in a wide range of taxa, some taxa have a particularly high frequency of null alleles; examples include insects (Lepidoptera, reviewed in Meglecz et al. 2004; Diptera, Lehmann et al. 1997; and Orthoptera, Chapuis et al. 2005) and mollusks (Li et al. 2003; Astanei et al. 2005). Interestingly, these are species with large effective population sizes. The association between the presence of null alleles and highly variable flanking regions has been demonstrated repeatedly in molecular studies, and several studies have suggested that the sequences flanking microsatellites may be less stable than those in other genomic regions (Angers and Bernatchez 1997; Grimaldi and Crouau-Roy 1997; Meglecz et al. 2004). On the other hand, no correlation has been found between null allele frequency and microsatellite unit-repeat length or motif complexity (Li et al. 2003), 2 factors related to the mutation rate of the microsatellite repeat region (Jin et al. 1996; Chakraborty et al. 1997). The null allele frequency in a congeneric species has been shown to rapidly increase with increasing phylogenetic distance from a focal species (e.g., in the oyster *Crassostrea*; Li et al. 2003). Despite the known prevalence of null alleles, the evolutionary dynamics and patterns of variation of these alleles in populations has never been examined analytically or by computer simulation.

Ninety percent of articles reporting microsatellite loci with null alleles include these loci in their analyses without correction for potential bias (reviewed in Dakin and Avise 2004). Yet null alleles may affect the estimation of population differentiation, for instance, by reducing the genetic diversity within populations (e.g., Paetkau and Strobeck 1995). Markedly, F_{ST} and genetic distances values generally increase with decreasing within-population genetic diversity (Slatkin 1995; Paetkau et al. 1997). The extent to

Key words: coalescent, microsatellite, null alleles, population differentiation, *F* statistics, genetic distances.

E-mail: chapuimp@ensam.inra.fr.

Mol. Biol. Evol. 24(3):621–631. 2007
doi:10.1093/molbev/msl191
Advance Access publication December 5, 2006

which null alleles may overestimate population differentiation has never been investigated.

Null alleles can be detected in population studies by carefully testing for Hardy–Weinberg (HW) proportions, provided that observed heterozygote deficiencies have no other origin (e.g., Wahlund effect). Various null allele frequency estimators ($\hat{r}$) making use of this property have been developed (Dempster et al. 1977; Chakraborty et al. 1992; Brookfield 1996). Some authors have attempted to correct for null alleles in population genetic studies by statistical adjustment of the visible allele and genotype frequencies, based on $\hat{r}$ and assuming a single new null allele size common to all genotyped populations (Roques et al. 1999). However, experimental studies using various amplifications (null and nonnull) to determine the null allele sizes have suggested that null alleles often correspond to alleles with different sizes and that alleles with the same size may correspond to both null and visible states (Callen et al. 1993; Paetkau and Strobeck 1995; Lehmann et al. 1996). The efficiency of the null allele frequency estimators and the existing correcting method has not been assessed.

We used computer simulations based on the coalescent (Hudson 1990) to investigate the prevalence and distribution of null allele sizes at microsatellite loci. We then assessed the impact of such null alleles on 2 statistics traditionally used to estimate population differentiation, F_{ST} and genetic distance. We evaluated the available methods for estimating null allele frequency and population differentiation from data sets with null alleles and propose a new method for estimating F_{ST} in the presence of null alleles. We illustrate our simulation results by verifying empirically the presence and impact of null alleles in 2 published microsatellite data sets based on original and redesigned pairs of primers (Paetkau and Strobeck 1995; Lehmann et al. 1996).

Materials and Methods

Simulation Method

We used a 3-step simulation approach described schematically in figure 1.

Step 1. Genotypic data were simulated from an algorithm based on the coalescent (Leblois et al. 2003; Paetkau et al. 2004). Two population models were assumed: a migration model and a split population model. In the migration model, 2 populations of equal effective size N_e exchange migrants at a rate m. In the population split model, an ancestral population of N_e individuals splits into 2 populations, each with the same effective size N_e; these 2 populations then do not exchange any genes for t generations. After the coalescent tree was constructed, we simulated mutational events on this tree, both within the repeat region of the microsatellite locus (hereafter referred to as R; mutation rate μ_R) and in the bases flanking the microsatellite locus for which a mutation is likely to prevent primer binding (hereafter referred to as B; mutation rate μ_B). We chose B as the 10 bp binding to the 3′ end of each 20 bp–long primer, so that only half of the mutations at the binding sites precluded PCR amplification. R and B were assumed to be completely linked. This assumption is reasonable because of the short physical distance between these regions (less than 300 bp). The number of mutations in R and B was simulated along each branch of the tree, according to a Poisson distribution with parameters $L\mu_R$ and $L\mu_B$, respectively, where L is the length of the branch in generations. Mutation rates μ_R were assumed to be equal for all loci. The same assumption was made for the mutation rates μ_B. Mutations in R followed a symmetric generalized stepwise mutation model (GSM) without allele size constraints (Zhivotovsky et al. 1997; Estoup et al. 2002). Changes in the number of repeat units followed a geometric distribution with a variance of 0.36 (Estoup et al. 2001). Mutations in B followed an infinite allele model (Kimura and Crow 1964). Once genotypic data had been simulated for both R and B, we randomly selected a gene copy used for the design of the microsatellite primers from a single focal population. This imitates the work of molecular biologists, who design PCR primers based on the sequence of a single gene copy in a given population. The allele state of the B region of the selected gene copy (hereafter referred to as B-cloned allele state) corresponded to the state of B for which PCR amplification was successful. All other B allele states were assumed to preclude PCR amplification. Therefore, any R gene copy not associated with the B-cloned allele state bore a null allele.

Step 2. From a single set of genotypic data, 3 data sets, composed of 60 genes (or 30 diploid individuals) for each population, were generated simultaneously. In the first, all B allele states were assumed to allow PCR amplification, so no null alleles were present (VA data set for visible alleles data set). Using the second data set, the R alleles not associated with the B-cloned allele state were assumed to be null (NA data set). The simulated NA genotype data set was corrected for null alleles following the approach of all empirical population genetics studies to date (CNA data set; Roques et al. 1999). Null and visible allele frequencies were first estimated with the algorithm described in Dempster et al. (1977) and the Supplementary Material online, which performed best of all the null allele frequency estimators tested (see Results). Homozygous genotype frequencies were then adjusted. We partitioned apparent homozygous counts n_{ii} into true n^*_{ii} and false n^*_{i0} homozygous counts. The true homozygote frequency is $p^*_{ii}=[n^*_{ii}/(n^*_{ii}+n^*_{i0})](n_{ii}/n)$ with n the number of individuals. Based on the relationships between true genotype counts and frequencies, we obtained the following estimate for homozygote frequency: $\hat{p}_{ii}=[\hat{p}_i/(\hat{p}_i+2\hat{r}_D)]\ (n_{ii}/n)$, with $\hat{r}_D$ the estimate of null allele frequency. Finally, all null alleles were given a single arbitrary allele size, not present in the original data set.

Step 3. The available method for estimating population differentiation in the presence of null alleles uses CNA genotype data sets and is referred to as INA (i.e., including null alleles). The F_{ST} estimate at a given locus is the appropriate combination of allele-based estimates for several alleles (Weir 1996). We hence propose a new correction for estimating F_{ST} in the presence of null alleles, in which F_{ST} is estimated from CNA data sets, but the calculation is restricted to visible allele sizes (referred to as ENA for excluding null alleles). Note that, in this case, the sums of the frequencies of alleles and genotypes

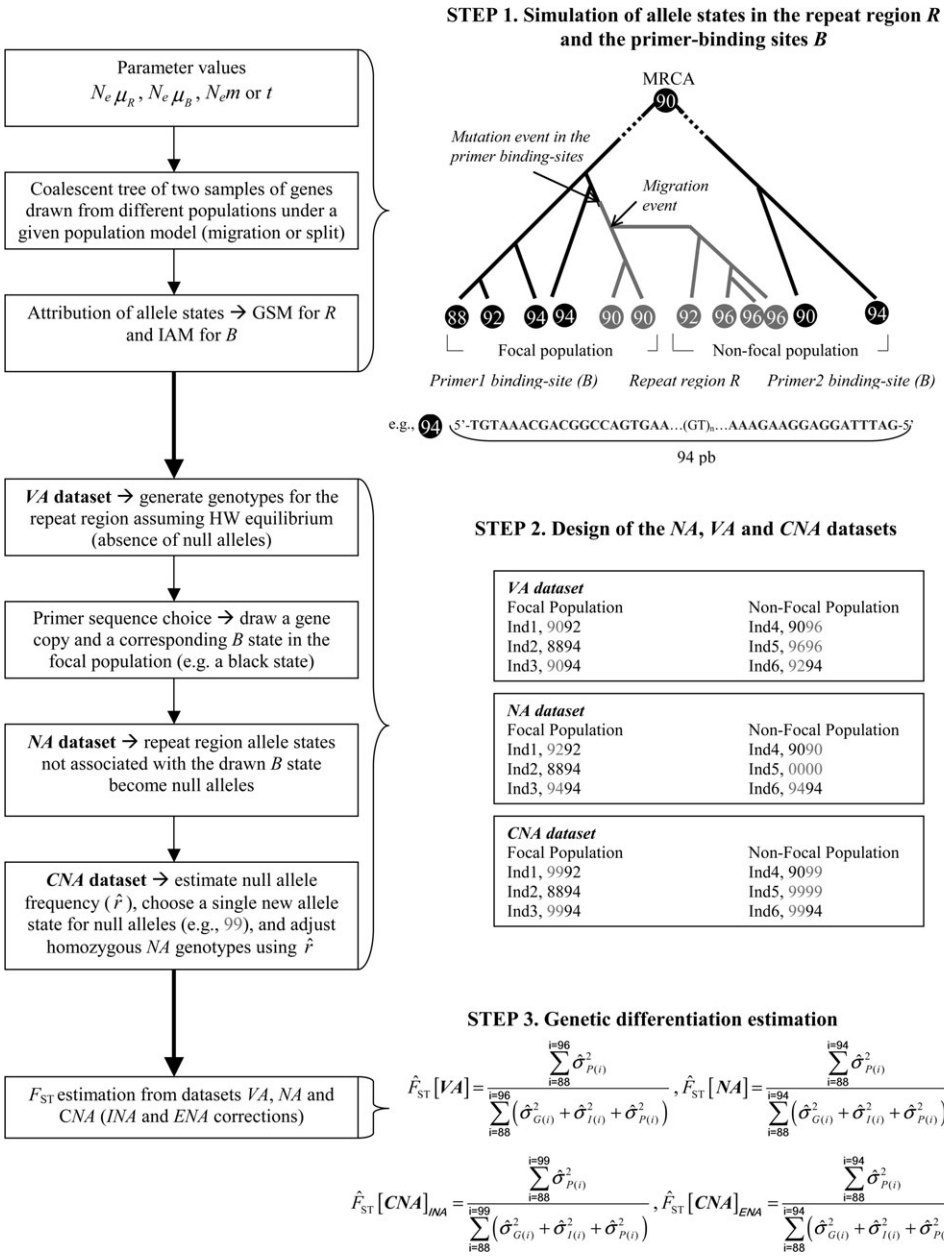

FIG. 1.—Synopsis of the simulation method. A single iteration is presented. In the coalescent tree, the allele state in the binding sites of the 2 microsatellite primers of the "gray" gene copies leads to null alleles. Estimation of genetic differentiation is illustrated by estimation of F_{ST}. $\hat{\sigma}^2_P$, $\hat{\sigma}^2_I$, and $\hat{\sigma}^2_G$ are the estimated components of variance for populations, individuals within populations, and genes within individuals, respectively. GSM, generalized stepwise mutational model (Zhivetorsky et al. 1997; Estoup et al. 2002); IAM, infinite allele model (Kimura and Crow 1964); *R*, repeat region; and *B*, primer-binding sites.

are not adjusted to 1. This approach cannot be used in the calculation of genetic distances, however, because genetic distances are expressed in terms of the proportions of similar alleles between and within populations, and so the lowest level of integration for such measures is the locus (i.e., the entire set of visible and null alleles).

Tests on Simulated Data Sets

We generated 10,000 simulated data sets for 35 different couples of values of the mutational parameter $N_e\mu_B$ (10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, and 1) and the populational parameter N_em (0.01, 0.1, 1, and 10) or t (1,000, 10,000, and 100,000) according to the population model considered. It is worth stressing here that the product $N_e\mu_B$, not μ_B alone, determines the level of variation in binding sites and hence the prevalence of null alleles in population gene samples. Preliminary simulations showed that the prevalence and allele size distribution of null alleles remained similar for a large range of $N_e\mu_R$ values (results not shown). We therefore fixed the product $N_e\mu_R$ at 1 for all simulations. This resulted in heterozygosity values spanning a large part of the range of heterozygosity generally observed at microsatellite markers (0.5–0.8; Takezaki and Nei 1996).

We first tested observations stemming from molecular studies that null alleles at a microsatellite locus are likely to be encountered in populations with a large effective size and/or an unusually high mutation rate in the flanking regions (i.e., large $N_e\mu_B$ values) and in populations that have diverged from the population from which the cloned allele state was drawn and the primers designed. To do so, we determined the range of values and/or combinations of the parameters $N_e\mu_B$ and N_em or t (according to the population model considered) favoring the presence of null alleles in population gene samples by simulating single-locus NA data sets. The simulated loci were categorized, separately for the focal and nonfocal population, into 3 classes of null allele frequency: negligible ($r < 0.05$), moderate ($0.05 \leq r < 0.20$), or large ($r \geq 0.20$). We then tested whether all null alleles in both populations correspond to a single shared allele. Distributions of null allele sizes, within and between populations, were characterized for data sets harboring null alleles. This allowed us to estimate the within-population percentages of allele sizes associated with null gene copies for the focal and the nonfocal populations and the percentage of allele sizes associated with null gene copies that are shared by both populations.

In the remaining tests, we simulated data sets of 10 and 100 loci. Researchers typically counter the large variances of differentiation estimators by examining between 5 and 20 loci. Ten loci thus mimic a typical empirical data set. However, larger numbers of loci (e.g., several hundreds) are required for reliable estimates of between-population parameters, such as migration rates (Whitlock and McCauley 1999) or times of population splitting events (Zhivotovsky and Feldman 1995). We assessed the effect of null alleles on population differentiation estimation by evaluating the Weir's (1996) unbiased estimator of F_{ST}, the genetic distance of Cavalli-Sforza and Edwards (1967) (D_C), and Nei's (1978) standard genetic distance (D_S). We compared the differentiation estimators for VA and NA data sets that correspond to the same set of parameters. We then estimated null allele frequencies averaged over the 2 populations, using the 3 methods of Dempster et al. (1977; Dempster method; estimate $\hat{r}_D$), Chakraborty et al. (1992; Chakraborty method; estimate $\hat{r}_C$), and Brookfield (1996; Brookfield method; estimate $\hat{r}_B$). Details about the null allele frequency estimates are provided as Supplementary Material online. We evaluated the methods according to 1) their applicability, expressed as the percentage of times an estimate was successfully produced and 2) a comparison of the means of estimated and simulated frequencies of null alleles averaged over the 2 populations.

Finally, we assessed the performance of available (INA) and new (ENA; for F_{ST} only) methods for estimating population differentiation with data sets that included null alleles. The efficiency of correction for estimates of F_{ST} was evaluated with respect to Weir's (1996) F_{ST} values calculated with VA data sets or Li's (1976) equilibrium value:

$$F_{ST} = \frac{1}{1 + 2N_e\left(2\mu_R + 2\frac{n_d}{n_d - 1}m\right)},$$

with the number of demes $n_d = 2$. As the 2 comparisons gave similar results (details not shown), only the comparison with Weir's (1996) F_{ST} values calculated with VA data sets is shown. As the relationship $D_S = 2\mu_R t$ (Nei 1972) does not hold under a GSM (Takezaki and Nei 1996), it was not considered in our comparisons. The performances of INA and ENA were evaluated by 1) comparing the distributions of each estimator of F_{ST}, D_S, and D_C calculated from CNA data sets, according to INA and ENA for F_{ST} and INA for D_S and D_C, with those calculated from VA data sets and 2) calculating a success index for the corrections. This index corresponds to the percentage of times the differentiation estimate obtained with the VA data set was closer to the differentiation estimate obtained with the CNA data set, by INA or ENA, than to the differentiation estimate obtained with the NA data set. For instance, for F_{ST}, we calculated the percentage of times $|\hat{F}_{ST[CNA]} - \hat{F}_{ST[VA]}| < |\hat{F}_{ST[NA]} - \hat{F}_{ST[VA]}|$.

Application to Empirical Molecular Data

In some studies, the inference that null alleles are present leads to the design of new primers for PCR amplification of DNA from all individuals originally identified as homozygous or null (reviewed in Dakin and Avise 2004). Although the 2 data sets obtained in this way are the empirical equivalents of our simulated NA and VA data sets, redesigning new primers does not guarantee that all null alleles are recovered (Ishibashi et al. 1996). To illustrate our simulation results with empirical molecular data, we reanalyzed 2 such published microsatellite data sets: a single locus from 3 Kenyan populations of the mosquito *A. gambiae* (Lehmann et al. 1996) and a single locus from 3 brown bear (*U. americanus*) populations sampled in Canadian National Parks (Paetkau and Strobeck 1995). These data sets represent different taxa, microsatellite loci, null allele frequencies, gene diversities, and levels of population differentiation. We first checked the recovery of HW equilibrium for each population using the genotype data sets obtained with new primers (Fisher's exact tests, as implemented in Genepop; Raymond and Rousset 1995). For each data set, we then calculated an empirical null allele frequency as the frequency of gene copies amplified only with the new primers. We compared this empirical estimation with estimates of null allele frequency calculated from the original data set, applying the 3 previously described methods. We compared global F_{ST} and mean genetic distance statistics calculated from the original data set, the new data set, and the original data set corrected for the presence of null alleles.

Results

Null Allele Prevalence and Distribution

We first tested the prediction that genetic diversity in binding sites B (determined by $N_e\mu_B$ values) substantially affects null allele prevalence in the focal population (fig. 2, dotted line). For values of $N_e\mu_B$ below 0.001, the prevalence of null alleles was low for most loci ($r < 0.05$). For values of $N_e\mu_B$ greater than 0.1, the incidence of null alleles was high, with most loci having a high frequency of null alleles ($r \geq 0.20$ for 71% of loci). For intermediate values of $N_e\mu_B$, a substantial proportion of loci

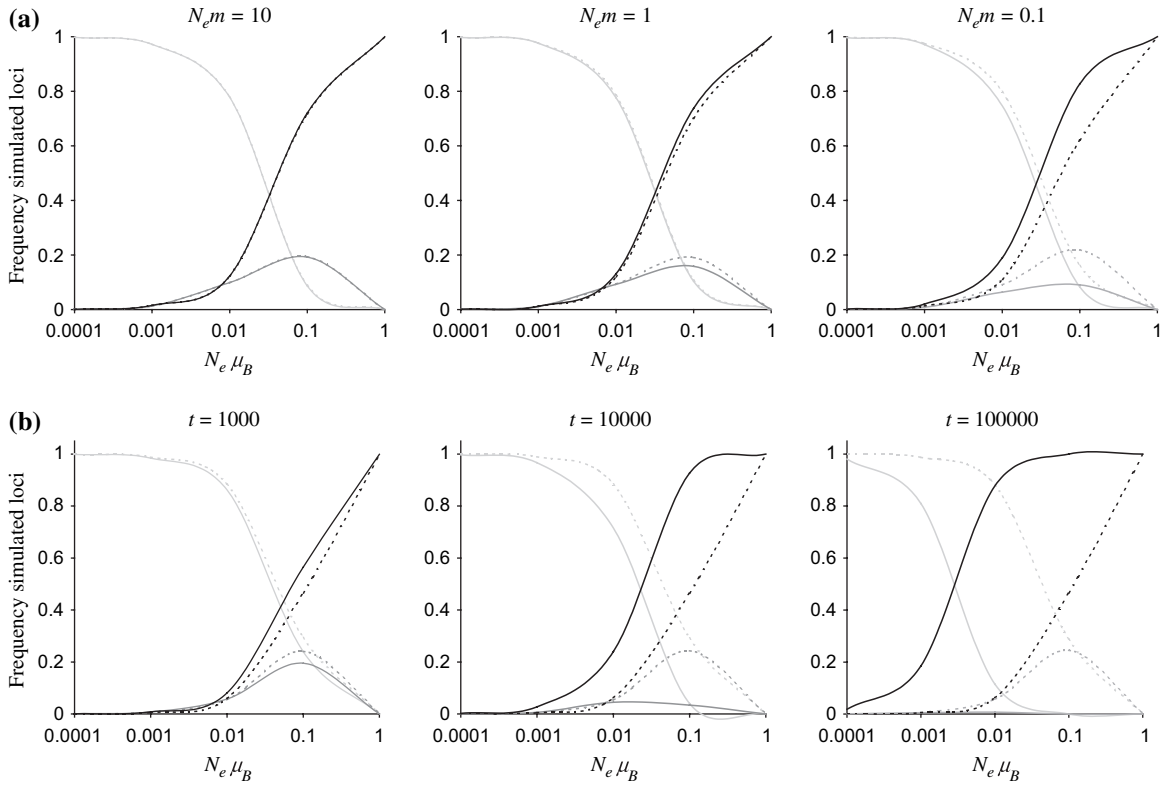

FIG. 2.—Prevalence of null alleles. Frequencies of simulated loci with a null allele frequency $r < 0.05$ (light gray), $0.05 \leq r < 0.20$ (dark gray), and $r \geq 0.20$ (black) as a function of the parameter $N_e\mu_B$ (x axis). Dotted lines represent the focal population and solid lines represent the nonfocal population. Different levels of gene flow and splitting time are tested for a migration model (*a*) and a population split model(*b*).

had a high frequency of null alleles ($r \geq 0.20$), and a moderate proportion of loci had an intermediate null allele frequency ($0.05 \leq r \leq 0.20$ for less than 19% of loci).

We then investigated how genetic differentiation from the focal population might favor null allele prevalence in the nonfocal population. Gene flow had a low to moderate impact on null allele prevalence (fig. 2*a*). The focal and nonfocal populations behaved similarly under high gene flow conditions ($N_em = 10$). However, for low values of gene flow ($N_em = 0.1$), the nonfocal population was more strongly affected by null alleles. In the population split model, in which there was assumed to be no gene flow (fig. 2*b*), both populations had very similar distributions of loci harboring null alleles at various frequencies for short to moderate splitting times ($t < 1,000$ generations). For longer times, the nonfocal population was much more strongly affected by null alleles, even for low $N_e\mu_B$ values.

Finally, we investigated whether all null alleles in all populations correspond to a single shared allele size. Figure 3 shows the distribution of null alleles according to allele sizes, within and between populations. For both population models, a large number of allele sizes harbored null gene copies whatever the value of $N_e\mu_B$ (fig. 3*a*). In the migration model, the focal and nonfocal populations behaved similarly for moderate to high levels of gene flow, with more than 34% of allele sizes harboring null gene copies. For low values of gene flow (i.e., $N_em = 0.1$), the nonfocal population displayed a slightly higher number of allele sizes with null gene copies (results not shown). In the population split model, the nonfocal population displayed a much larger number of allele sizes with null gene copies ($\geq 60\%$ for $t = 10,000$) than the focal population. This result held for a large range of splitting times (results not shown). In the migration model, less than half of the allele sizes harboring null gene copies were shared between the 2 populations for almost all combinations of parameter values tested (fig. 3*b*). The proportion of shared null allele sizes decreased with lower gene flow and $N_e\mu_B$. In the population split model, for all splitting times tested, populations shared very few allele sizes harboring null gene copies (less than 20% in most cases).

Effect of Null Alleles on the Estimation of Population Differentiation

We tested the prediction that the presence of null alleles causes bias in differentiation estimators (fig. 4, black and gray lines). The presence of null alleles led to overestimation of both F_{ST} and genetic distance. In the migration model, bias in F_{ST} was moderate for intermediate null allele frequencies or high levels of gene flow. Larger bias was observed for high null allele frequencies and low levels of gene flow, with the F_{ST} distributions based on VA and NA data sets becoming almost nonoverlapping. In the population split model, the effect on genetic distances remained moderate, even for large null allele frequencies and large splitting times. D_C was found to be slightly less affected by null alleles than D_S. D_S could not be calculated

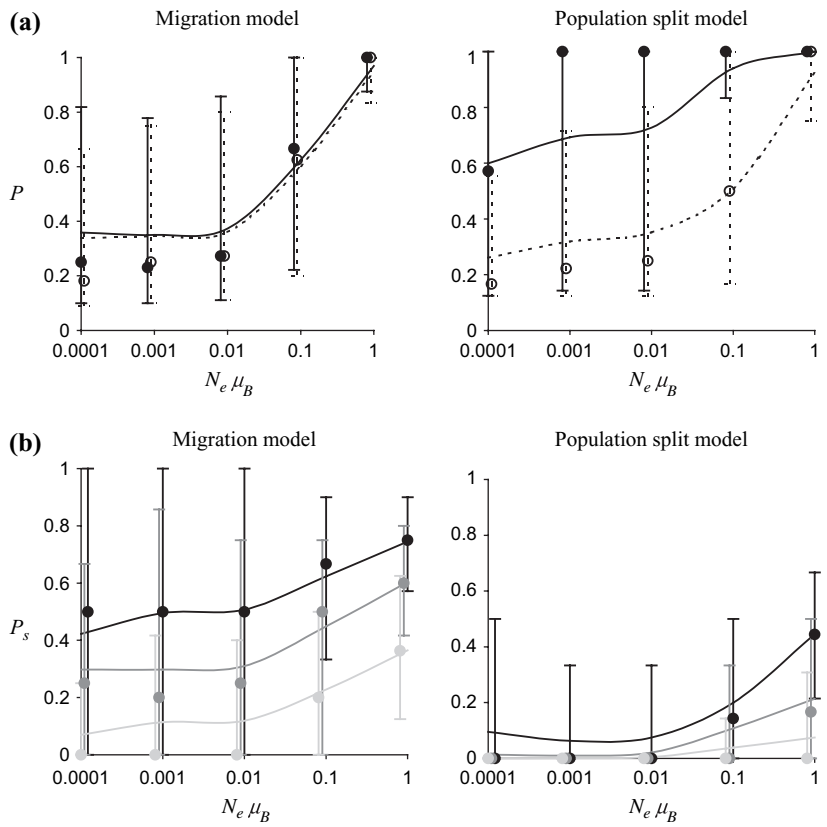

FIG. 3.—Allele sizes harboring null gene copies. Distribution of null allele sizes within (*a*) and between (*b*) populations presented along the *y* axis as a function of the parameter $N_e\mu_B$ (*x* axis). *P*, proportion within population of allele sizes harboring null gene copies; P_s: proportion of allele sizes harboring null gene copies that are shared by both populations. Mean estimates (line), 50 (points), and 10 and 90 (bars) percent quantile values are represented. (*a*) Both the focal (dotted line) and the nonfocal (solid line) populations are presented. The parameter N_em is fixed at 1 for the migration model. The parameter *t* is fixed at 10,000 for the population split model. (*b*) For the migration model, the tested values of gene flow are $N_em = 0.1$ (light gray), $N_em = 1$ (dark gray), and $N_em = 10$ (black). For the population split model the tested values of splitting time are $t = 100{,}000$ (light gray), $t = 10{,}000$ (dark gray), and $t = 1{,}000$ (black).

for a diverse range of $N_e\mu_B$ and *t* values (results not shown). These failures to calculate D_S corresponded to paired populations that did not share at least one allele state. This situation is likely for large splitting times, even in the absence of null alleles. However, in the presence of null alleles, the probability of sharing no allele increases (results not shown). Increasing the number of loci reduced the variance of estimation for both F_{ST} and genetic distances, but did not change the null allele bias.

Estimation of Null Allele Frequency

The performance of the methods for estimating null allele frequency under the migration model and for genotype data sets of 10 loci are presented in figure 5. The results obtained for the population split model and for genotype data sets of 100 loci were similar and are therefore not shown. The Chakraborty method generated negative estimates of null allele frequency (fig. 5*a*) when the simulated null allele frequency was close to 0 for at least 1 of the 2 populations and when the number of visible genotypes for 1 population was too small for correct estimation of the observed heterozygozity. The Chakraborty method was also not applicable for monomorphic populations. The Chakraborty method gave a small positive bias and a large variance, especially for large values of null allele frequency (fig. 5*b*). This may simply result from sample size being reduced in this case because estimation with the Chakraborty method is carried out for individuals with at least one visible band. Other methods had an applicability of 1 for all sets of parameter values tested. The Brookfield method displayed a slight positive bias and its variance was low. The Dempster method provided unbiased and low variance estimates of null allele frequencies. Results were similar for a wide range of N_em and *t* values and number of loci (results not shown). We therefore conclude that the Dempster method was the best method of the 3 for estimating null allele frequencies.

Correction Methods for Estimating Population Differentiation

Figure 4 shows the differentiation estimates obtained from CNA data sets including (INA) or excluding (ENA) the null allele size for different categories of null allele frequency and numbers of loci. The INA correction continued to generate biased values of F_{ST}. This procedure partially educed the bias induced by null alleles in the presence of high levels of gene flow, generating values of F_{ST} estimates smaller than those obtained with uncorrected data sets.

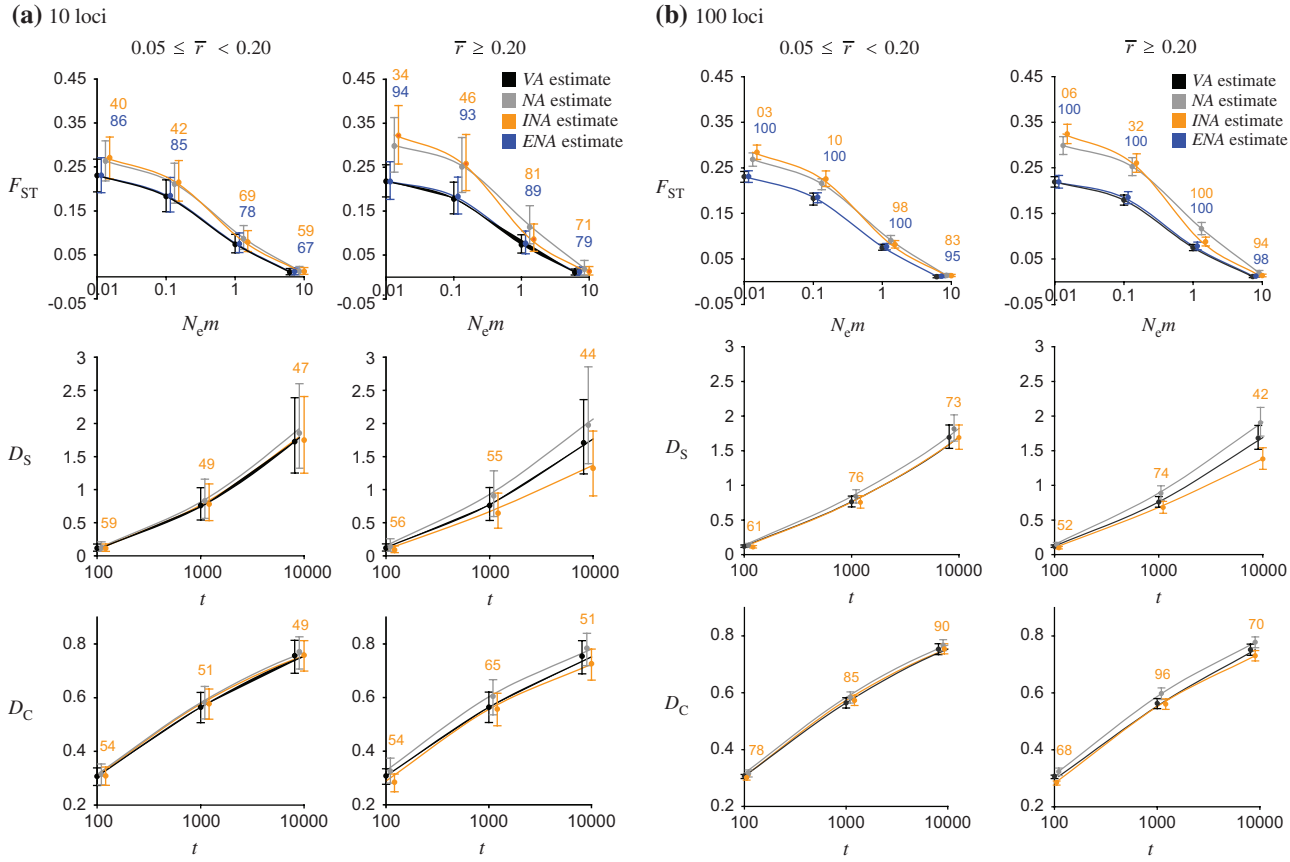

FIG. 4.—Effects of null alleles on estimation of population differentiation and performance of correction methods for a genotyping effort of 10 loci (*a*) and 100 loci (*b*). F_{ST} and genetic distance estimates (*y* axis) are presented as a function of gene flow (N_em) for F_{ST} or splitting time (*t*) for genetic distances (*x* axis). The differentiation estimates are based on VA data sets (black line), NA data sets (gray line), and CNA data sets including (INA, orange line) or excluding (ENA, blue line) the null allele size. Null allele frequency is estimated using the Dempster method. Mean estimates (line), 50 (points), and 10 and 90 (bars) percent quantile values are represented. Numbers refer to success indices corresponding to the percentages of differentiation estimates based on the VA data sets that are closer to the differentiation estimates based on the CNA data sets than to the differentiation estimates based on the NA data sets. The CNA data set estimate was generated following the INA (orange) or ENA (blue) correction method. All estimates were calculated for 2 classes of mean null allele frequency $\bar{r}$: $0.05 \leq \bar{r} \leq 0.20$ and $\bar{r} \geq 0.20$. D_S: Nei's (1978) standard distance; D_C: the distance of Cavalli-Sforza and Edwards (1967).

However, this procedure increased the bias induced by null alleles in the presence of low levels of gene flow, with F_{ST} estimates reaching values larger than those obtained from uncorrected data sets. In contrast, the newly proposed ENA method almost entirely resolved the bias induced by null alleles, regardless of null allele frequency, the level of gene flow, and the number of loci. Variance estimates for the ENA method were only slightly larger than those with VA data sets. These results were confirmed by success index values, which were larger than 67% for 10 loci and 95% for 100 loci (fig. 4).

INA decreased the bias in genetic distance estimation, almost eliminating it for moderate null allele frequencies. However, INA gave a negative bias for high null allele frequencies. These findings applied to both D_S and D_C, but the bias was substantially less pronounced for D_C than for D_S. For 10 loci, INA only marginally improved genetic distance estimation, as confirmed by success index values (fig. 4*a*). Increasing the number of loci to 100 increased the success index values for INA, in spite of similar biases (e.g., between 68% and 96% for D_C; fig. 4*b*). This probably results from a much smaller variance of distance estimation for large number of loci. Thus, there appears to be a gain in using D_C corrected by conventional methods, at least for data sets with a large number of loci.

Application to Empirical Molecular Data Sets

H_O and H_E values and tests of HW disequilibrium showed that null alleles were largely eliminated by the design of new primers for both *Anopheles gambiae* and *Ursus americanus* (table 1). However, the heterozygote deficit remained significant for *A. gambiae*. As some null genotypes were still observed and Lehmann et al. (1997) excluded the Wahlund effect as an explanation of HW deviations in the genotype data set obtained with the original primer set, the smaller, but still significant, HW deviation in the data set obtained with the new primers may reflect the presence of nonrecovered null alleles. Population estimates of null allele frequency $\hat{r}$ were generally close to the empirical values, estimated as the frequency of gene copies amplified only with the new primers. However, all $\hat{r}$ values were larger than the empirical *r* values for *A. gambiae* populations, probably due to the incomplete recovery of null

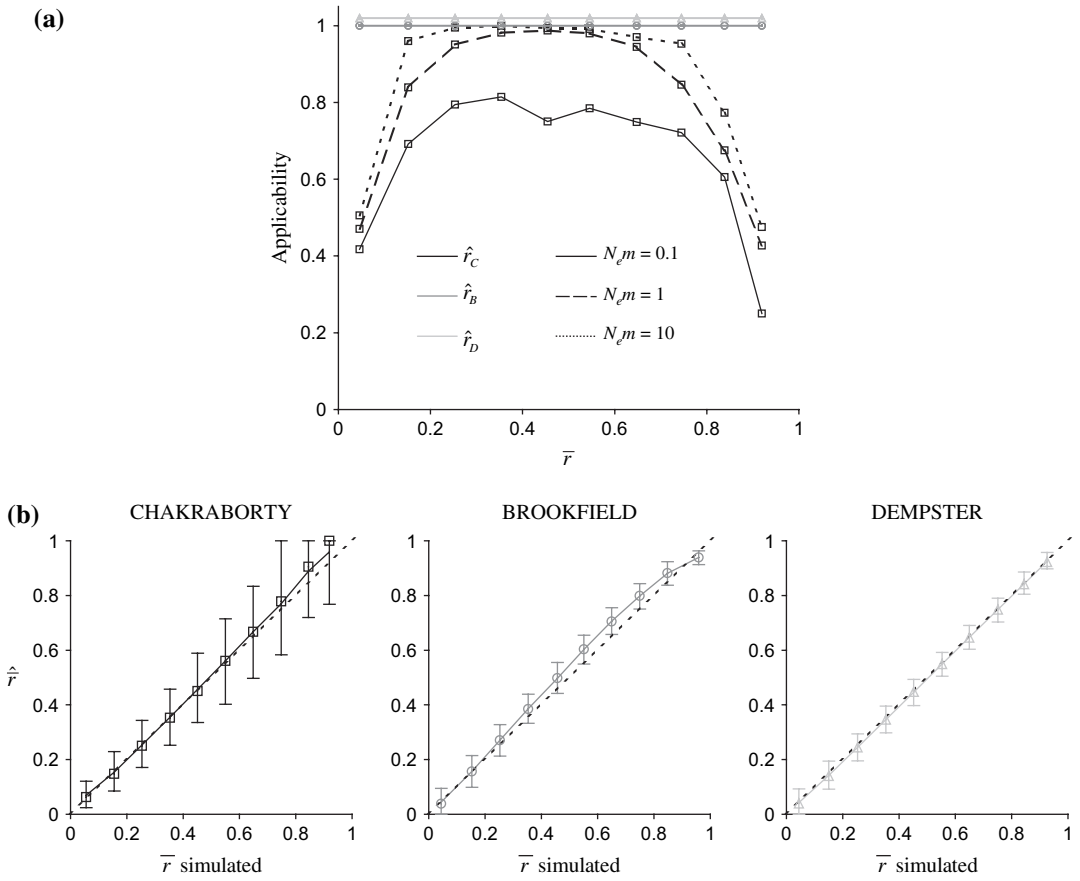

FIG. 5.—Performance of methods for estimating null allele frequency. Applicability (*a*) and mean and quantile values (*b*) of null allele frequency estimates (*y* axis) were plotted as a function of the simulated mean null allele frequency $\bar{r}$ (*x* axis) grouped into classes of 0.1 units. The methods evaluated are those of Chakraborty ($\hat{r}_C$:□, black), Brookfield ($\hat{r}_B$: ○, dark gray), and Dempster ($\hat{r}_D$: △, light gray), as described in Supplementary Material online. Calculations were performed under the migration model and for genotype data sets of 10 loci. (*a*) The applicability is the percentage of times an estimate is successfully produced. Different values of gene flow were tested: $N_em = 0.1$ (solid line), $N_em = 1$ (broken line), and $N_em = 10$ (dotted line). For $\hat{r}_B$ and $\hat{r}_D$, the corresponding different lines are merged. (*b*) Mean estimates (lines), 50 (points), and 10 and 90 (bars) percent quantile values are presented. N_em was fixed at 1.

alleles in this species. Moreover, $\hat{r}_C$ values also appeared to be overestimated in *U. americanus*, especially for the population from Fundy, probably due to its small sample size.

The conclusions drawn from the population differentiation tests were the same for all 3 data sets (original primer data set, new primer data set, and corrected original primer data set): no significant differentiation in *A. gambiae*

Table 1
Null Alleles in Empirical Molecular Data. Data Set Details and Estimation of Null Allele Frequency

		Original Primer Set				New Primer Set				Null Allele Frequencies			
Sample Site	*n*	n_0	H_O	H_E	HW Test	n_0	H_O	H_E	HW Test	*r*	$\hat{r}_C$	$\hat{r}_B$	$\hat{r}_D$
Anopheles gambiae[a]													
Village 3	39	7	0.415	0.860	*	4	0.705	0.870	*	0.174	0.344	0.427	0.367
Village 7	54	3	0.352	0.854	*	0	0.737	0.867	*	0.184	0.412	0.344	0.316
Village 15	70	7	0.378	0.848	*	3	0.731	0.860	*	0.224	0.380	0.373	0.331
Ursus americanus[b]													
Fundy	11	3	0.000	0.600	*	0	0.636	0.502	n.s.	0.591	1.000	0.643	0.589
La Mauricie	31	2	0.379	0.878	*	0	0.774	0.856	n.s.	0.242	0.389	0.351	0.322
Terra Nova	26	1	0.080	0.520	*	0	0.385	0.529	n.s.	0.192	0.729	0.351	0.336

NOTE.—Sample size in diploid individuals (*n*), number of null genotypes (n_0), observed (H_O), and expected (H_E) heterozygosities for original and new primer sets. Null allele frequencies in original data sets were calculated as described by Chakraborty, $\hat{r}_C$; Brookfield, $\hat{r}_B$; and Dempster, $\hat{r}_D$ (see Supplementary Material Online). *r* is the ''real'' estimate of null allele frequency calculated as the frequency of genes amplified only with new primers. HW test: HW exact rest as implemented in GENEPOP (Raymond and Rousset 1995), *: significant departure at $\alpha = 0.05$, and n.s.: not significant.

[a] Data sets originally published in Lehmann et al. (1996).

[b] Data sets were originally published in Paetkau and Strobeck (1995).

Table 2
Full Alleles Empirical Molecular Data Estimation of Genetic Differentiation

Differentiation Estimator	Population Set	Original Primer Data Set	New Primer Data Set	Original Primer Data Set Corrected	
				INA	ENA
Global F_{ST}	*A. gambiae*	−0.011	−0.005	−0.005	−0.005
	U. americanus	0.177	0.150	0.078	0.092
Mean D_S	*A. gambiae*	−0.036	−0.026	−0.019	n.a.
	U. americanus	0.727	0.354	0.234	n.a.
Mean D_C	*A. gambiae*	0.122	0.135	0.110	n.a.
	U. americanus	0.566	0.498	0.445	n.a.

NOTE.—Original data sets were corrected using $\hat{r}_D$ estimation. D_S: Nei's (1978) standard distance; D_C: the distance of Cavalli-Sforza and Edwards (1967); INA: calculation of the differentiation measures (F_{ST} and genetic distance) from the data set corrected for null alleles when the null allele size is included; ENA: calculation of the F_{ST} from the data set corrected for null alleles when the null allele size is excluded; and n.a.: not applicable.

populations and significant differentiation in *U. americanus* populations (results not shown). In agreement with our simulation results, F_{ST} and genetic distances were considerably larger in the original data set harboring null alleles than in the data set obtained with the new primers, at least when genetic differentiation was significant (i.e., in *U. americanus*; table 2). The corrected data set gave lower D_S and D_C values than the new primer data set, consistent with simulation results. However, the F_{ST} value obtained for *U. americanus* with the new primer data set was more similar to that calculated from the original data set than to that calculated from the corrected data set. This may be due to the large variance observed in our simulations for single-locus F_{ST} estimation, regardless of the data set considered (results not shown).

Discussion

Null Allele Prevalence

Our simulations showed that null alleles were likely to be encountered in populations with high levels of diversity in flanking sequences, particularly for $N_e\mu_B \geq 0.001$. Assuming a frequency of point mutations at a specific basepair of 10^{-9} (Li et al. 1985), the mutation rate in key regions of the binding sites for microsatellite primers (i.e., the 10 bp binding to the 3′ end of each 20 bp-long primer), μ_B, is expected to be about 2×10^{-8}. Hence, null alleles are likely to be found only in populations with large effective sizes (i.e., $N_e \geq 50{,}000$ and even larger population sizes if some mutations in the 10 bp binding sites do not preclude PCR amplification in spite of the primer mismatch to the DNA template). The prevalence of null alleles varies considerably between studies, but microsatellite null alleles have been found in a wide range of taxa, including species for which N_e is not necessarily large (Dakin and Avise 2004). High mutation rates in the flanking sequences of microsatellite loci would be required to reconcile such empirical results with our simulations. In agreement with this, several molecular studies suggest that microsatellite flanking regions may be more unstable than is generally thought (Angers and Bernatchez 1997; Grimaldi and Crouau-Roy 1997; Meglecz et al. 2004). A simpler nonexclusive explanation for the frequent presence of null alleles in most real data sets is the high level of differentiation that may exist between the focal population and the genotyped populations. In agreement with molecular studies (Li et al. 2003), our simulations showed that the nonfocal population was more strongly affected by null alleles than the focal population, even for low $N_e\mu_B$ values.

Effect of Null Alleles on the Estimation of Population Differentiation

Simulated and empirical data sets showed that the presence of null alleles led to the overestimation of both F_{ST} and genetic distance in cases of significant population differentiation. F_{ST} estimates were unbiased in the absence of population structure, but were considerably affected in the presence of low levels of gene flow (i.e., strongly differentiated populations). The presence of null alleles may be particularly problematic in studies comparing different sets of populations with different frequencies of null alleles and/or patterns of gene flow, especially when one or several population sets are characterized by low levels of gene flow. The distance (D_C) Cavalli-Sforza and Edwards (1967) performed better than Nei's (1978) standard distance (D_S): D_C was less affected by null alleles and the bias remained similar for a large range of splitting times. This feature is important because genetic distances based on microsatellites are usually calculated for the construction of dendrograms of related taxa. If all pairwise D_S distances are similarly biased, then the tree topology should be roughly unchanged.

Correction Methods for Estimating Population Differentiation

Although the frequency of null alleles can be estimated precisely by the Dempster method, the conventional correction based upon this estimate of null allele frequency did not perform well. Bias in F_{ST} is larger after correction for null alleles in the presence of low levels of gene flow. Genetic distances calculated from corrected data sets were underestimated when null allele frequencies were high. However, the absolute bias on the distance of Cavalli-Sforza and Edwards (1967) was lower than that for uncorrected data sets. Our simulations demonstrated that null alleles often corresponded to multiple allele sizes, some of which were similar to those of visible alleles. This is due to the mutational model of the repeat region of the microsatellite, in which the loss or gain of a variable number of repeat units generates alleles identical in state but not in descent (i.e., allele size homoplasy; Estoup et al. 2002). This issue was

more pronounced in higher levels of population differentiation, where population differences in allele sizes of null gene copies were larger. The conventional assumption of a single null allele size common to all studied populations, rather than the actual allele sizes, amounts to considering these alleles as slowly evolving and so decreases the apparent overall mutation rate of the locus. As F_{ST} increases with decreasing $N_e\mu_R$ (Slatkin 1995), we would expect F_{ST} values calculated with the INA procedure to be overestimated with respect to F_{ST} values calculated from VA data sets (particularly in low gene flow conditions). Conversely, as genetic distance decreases with decreasing $\mu_R t$ (Nei 1972), the genetic distances values calculated with the INA procedure should be lower than those calculated from VA data sets. The assumption of arbitrarily choosing a single allele size common to all null alleles can be relaxed, at least when estimating F_{ST}, by restricting F_{ST} calculation from corrected data sets to visible allele sizes. F_{ST} calculation with the ENA procedure was unbiased and resulted in a variance only slightly larger than that for data sets without null alleles.

Supplementary Material

Methods for estimating null allele frequency are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We would like to thank T. Lehmann and D. Paetkau for providing us with the data sets on which their publications were based. We thank S. Baird, D. Bourguet, C. Brouat, J. M. Cornuet, K. Kim, Y. Michalakis, G. Roderick, T. Sappington, and 2 anonymous reviewers for constructive comments on an earlier version of the manuscript. This work was partly funded by the scientific Santé des Panteset Environement department of Institut National de a Recherche Agronomique. M.P.C. was supported by a grant from the Centre National de a Recherche Scientifique.

Literature Cited

Angers B, Bernatchez L. 1997. Complex evolution of a salmonid microsatellite locus and its consequences in inferring allelic divergence from size information. Mol Biol Evol. 14:230–238.

Astanei I, Gosling E, Wilson J, Powell E. 2005. Genetic variability and phylogeography of the invasive zebra mussel, *Dreissena polymorpha* (Pallas). Mol Ecol. 14:1655–1666.

Brookfield JFY. 1996. A simple new method for estimating null allele frequency from heterozygote deficiency. Mol Ecol. 5:453–455.

Callen DF, Thompson AD, Shen Y, Phillips HA, Richards RI, Mulley JC. 1993. Incidence and origin of 'null' alleles in the (AC)*n* microsatellite markers. Am J Hum Genet. 52:922–927.

Cavalli-Sforza LL, Edwards AWF. 1967. Phylogenetic analysis: models and estimation procedures. Am J Hum Genet. 19:233–257.

Chakraborty R, De Andrade M, Daiger SP, Budowle B. 1992. Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. Ann Hum Genet. 56:45–57.

Chakraborty R, Kimmel M, Stivers DN, Davison LJ, Deka R. 1997. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc Natl Acad Sci USA. 94:1041–1046.

Chapuis M-P, Loiseau A, Michalakis Y, Lecoq M, Estoup A. 2005. Characterization and PCR multiplexing of polymorphic microsatellite loci for the locust *Locusta migratoria*. Mol Ecol Notes. 5:554–557.

Dakin EE, Avise JC. 2004. Microsatellite null alleles in parentage analysis. Heredity. 93:504–509.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B. 39:1–38.

Estoup A, Angers B. 1998. Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations. In: Carvalho G, editor. Advances in molecular ecology. Amsterdam: IOS Press. p. 55–86. (NATO ASI series).

Estoup A, Jarne P, Cornuet JM. 2002. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Mol Ecol. 11:1591–1604.

Estoup A, Wilson IJ, 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.

Gagneux P, Boesch C, Woodruff DS. 1997. Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair. Mol Ecol. 6:861–868.

Grimaldi MC, Crouau-Roy B. 1997. Microsatellite allelic homoplasy due to variable flanking sequences. J Mol Evol. 44:336–340.

Hudson RR. 1990. Gene genealogies and the coalescent process. In: Futuyama D, Antonovics J, editors. Oxford surveys in evolutionary biology. Oxford: Oxford University Press. p. 1–44.

Ishibashi Y, Saitoh T, Abe S, Yoshida MC. 1996. Null microsatellite alleles due to nucleotide sequence variation in the grey-sided vole *Cleithrionomy rufocanus*. Mol Ecol. 5:589–590.

Jin L, Macaubas C, Hallmayer J, Kimura A, Mignot E. 1996. Mutation rate varies among alleles at a microsatellite locus: phylogenetic evidence. Proc Natl Acad Sci USA. 93:15285–15288.

Kimura M, Crow JF. 1964. The number of alleles that can be maintained in a finite population. Genetics. 49:725–738.

Leblois R, Estoup A, Rousset F. 2003. Influence of mutational and sampling factors on the estimation of demographic parameters in a "Continuous" population under isolation by distance. Mol Biol Evol. 20:491–502.

Lehmann T, Besanky NJ, Hawley WA, Fahey TG, Kamau L, Collins FH. 1997. Microgeographic structure of *Anopheles gambiae* in western Kenya based on mtDNA and microsatellite loci. Mol Ecol. 6:243–253.

Lehmann T, Hawley WA, Collins FH. 1996. An evaluation of evolutionary constraints on microsatellite loci using null alleles. Genetics. 144:1155–1163.

Li G, Hubert S, Bucklin K, Ribes V, Hedgecock D. 2003. Characterization of 79 microsatellite DNA markers in the Pacific oyster *Crassostrea gigas*. Mol Ecol Notes. 3:228–232.

Li W-H, Luo C-C, Wu C-I. 1985. Evolution of DNA sequences. In: Macintryre RJ, editor. Molecular evolutionary genetics. New York: Plenum Press. p. 1–94.

Meglecz E, Petenian F, Danchin E, Coeur d'Acier A, Rasplus JY, Faure E. 2004. High similarity between flanking regions of different microsatellites detected within each of two species of lepidoptera: *Parnassius apollo* and *Euphydryas aurinia*. Mol Ecol. 13:1693–1700.

Nei M. 1972. Genetic distance between populations. Am Nat. 106:283–291.

Nei M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics. 89:583–590.

Paetkau A, Slade R, Burden M, Estoup A. 2004. Genetic assignment methods for the direct, real-time estimation of migration

rate: a simulation-based exploration of accuracy and power. Mol Ecol. 13:55–65.
Paetkau D, Strobeck C. 1995. The molecular basis and evolutionary history of a microsatellite null allele in bears. Mol Ecol. 4:519–520.
Paetkau D, Waits LP, Clarkson PL, Craighead L, Strobeck C. 1997. An empirical evaluation of genetic distance statistics using microsatellite data from bear (Ursidae) populations. Genetics. 147:1943–1957.
Raymond M, Rousset F. 1995. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. Heredity. 86:248–249.
Roques S, Duchesne P, Bernatchez L. 1999. Potential of microsatellites for individual assignment: the North Atlantic redfish (genus *Sebastes*) species complex as a case study. Mol Ecol. 8:1703–1717.
Shinde D, Lai YL, Sun FZ, Arnheim N. 2003. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)(n) and (A/T)(n) microsatellites. Nucleic Acids Res. 31:974–980.
Slatkin M. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics. 139:457–462.
Takezaki N, Nei M. 1996. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. Genetics. 144:389–399.
Weir BS. 1996. Genetic data analysis II. Sunderland (MA): Sinauer Associates.
Whitlock MC, McCauley DE. 1999. Indirect measures of gene flow and migration: F_{ST} not equal $1/(4Nm+1)$. Heredity. 82:117–125.
Zhivotovsky LA, Feldman MW. 1995. Microsatellite variability and genetic distances. Proc Natl Acad Sci USA. 92:11549–11552.
Zhivotovsky LA, Feldman MW, Grishechkin SA. 1997. Biased mutations and microsatellite variation. Mol Biol Evol. 14:926–933.

Lauren McIntyre, Associate Editor

Accepted November 29, 2006