Multiple Levels of Single-Strand Slippage at Cetacean Tri- and Tetranucleotide Repeat Microsatellite Loci

Per J. Palsbøll,*,† Martine Bérubé† and Hanne Jørgensen†

**Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92697-2525, †Department of Population Biology, University of Copenhagen, DK-2100 Copenhagen, Denmark*

Manuscript received July 2, 1998
Accepted for publication September 21, 1998

ABSTRACT

Between three and six tri- and tetranucleotide repeat microsatellite loci were analyzed in 3720 samples collected from four different species of baleen whales. Ten of the 18 species/locus combinations had imperfect allele arrays, *i.e.*, some alleles differed in length by other than simple integer multiples of the basic repeat length. The estimate of the average number of alleles and heterozygosity was higher at loci with imperfect allele arrays relative to those with perfect allele arrays. Nucleotide sequences of 23 different alleles at one tetranucleotide repeat microsatellite locus in fin whales, *Balaenoptera physalus*, and humpback whales, *Megaptera novaeangliae*, revealed sequence changes including perfect repeats only, multiple repeats, and partial repeats. The relative rate of the latter two categories of mutation was estimated at 0.024 of the mutation rate involving perfect repeats only. It is hypothesized that single-strand slippage of partial repeats may provide a mechanism for counteracting the continuous expansion of microsatellite loci, which is the logical consequence of recent reports demonstrating directional mutations. Partial-repeat mutations introduce imperfections in the repeat array, which subsequently could reduce the rate of single-strand slippage. Limited computer simulations confirmed this predicted effect of partial-repeat mutations.

ANALYSES of microsatellite loci are now commonplace in evolutionary and genetic studies of natural populations. Microsatellite loci are nucleotide sequences of one to five nucleotides arranged in tandem (Tautz 1989; Weber and May 1989), with mutation rates as high as 10^{-5}–10^{-2} (Weber and Wong 1993; Talbot *et al.* 1995; Amos and Rubinsztein 1996a; Primmer *et al.* 1996). The allelic states at microsatellite loci are usually scored from their molecular weight, and the subsequent data analysis relies on a mutational mechanism of single-strand slippage during replication (Levinson and Gutman 1987a,b), mainly of single repeats (Schlötterer and Tautz 1992; Mahtani and Willard 1993; Weber and Wong 1993; Talbot *et al.* 1995; Amos and Rubinsztein 1996a; Primmer *et al.* 1996; but see also Grimaldi and Crouau-Roy 1997; Orti *et al.* 1997). This stepwise mode of mutation combined with the high mutation rates violates the assumptions of the commonly used infinite allele/site models. This, in turn, has necessitated development of novel measures of genetic divergence specifically for the analysis of microsatellite data (*e.g.*, Goldstein *et al.* 1995a,b; Shriver *et al.* 1995; Slatkin 1995; Kimmel and Chakraborty 1996). Although most investigations of these novel statistics presented so far have been based on a simple symmetrical stepwise mutation model, the proposed statistics also accommodate more complicated distributions of changes in repeat numbers, including asymmetrical and multirepeat mutations (Kimmel and Chakraborty 1996; Kimmel *et al.* 1996).

Several reports have presented analyses of microsatellite data, which demonstrated deviations from null expectations of the simple symmetrical, stepwise mutation model. Likely explanations for the observed deviations are constraints on the number of repeats (Garza *et al.* 1995), presence of multirepeat mutations (Di Rienzo *et al.* 1994; Amos and Rubinsztein 1996a; Primmer *et al.* 1996), and/or directional mutation toward more repeats (Ellegren *et al.* 1995; Rubinsztein *et al.* 1995; Amos and Rubinsztein 1996a; Primmer *et al.* 1996).

A serious obstacle to additional insight into the mode of evolution at microsatellite loci is the fact that the only phylogenetic signal contained in the repeat array itself is the number of repeats. Hence, investigations of the mode of evolution at microsatellite loci have mainly relied on indirect analyses of deviations from the null-expectations, either by estimating the probability of the observed data under specific evolutionary models (*e.g.*, Shriver *et al.* 1993; Di Rienzo *et al.* 1994; Nielsen 1997), by including sequence data from other linked loci (*e.g.*, Jin *et al.* 1996; Orti *et al.* 1997), or by direct identification of germ-line mutations (*e.g.*, Amos and Rubinsztein 1996b; Primmer *et al.* 1996). An alternative approach, which has been pursued by several authors, is analyses of loci with interrupted or compound microsatellite repeat arrays (*e.g.*, Estoup *et al.* 1995; Garza *et al.* 1995; Garza and Freimer 1996; Messier *et*

Corresponding author: Per J. Palsbøll, School of Biological Sciences, University of Wales, Deiniol Rd., Bangor, Gwyneed LL57 2UW, Wales. E-mail: p.palsboll@bangor.ac.uk

al. 1996; Angers and Bernatchez 1997). These studies showed that the nucleotide sequence of the microsatellite array at such loci often provides additional evolutionary data not obtainable from the molecular weight alone and that, indeed, the difference in molecular weight may not be a reliable indicator of evolutionary distance. However, because one of the major advantages of microsatellite analyses over traditional sequence analyses in population assays is that alleles are scored by the molecular weight, such an additional sequence analysis will require a substantial increase in effort.

Here we present the results from a study of cetacean tri- and tetranucleotide repeat microsatellite loci where some alleles differ in length by other than simple multiples of the basic repeat length. These microsatellite loci differ from the interrupted and compound microsatellite loci presented previously by the fact that alleles at a locus can be divided into groups that represent different evolutionary lineages from the molecular weight alone. Sequencing of alleles at one locus in two species revealed a complex pattern of single-strand slippage on several levels that involved not only single repeats, but also multiple and partial repeats. We estimated the rate of mutations that included such imperfect or partial repeats at ~2.4% of the rate of mutations involving only perfect repeats (most of which are presumably single-step mutations).

MATERIALS AND METHODS AND RESULTS

Sample collection: A total of 3720 tissue samples were analyzed, the majority of which were obtained from free-ranging whales as skin biopsies (Palsbøll *et al.* 1991) or sloughed skin (Clapham *et al.* 1993), and a few that were obtained during whaling operations (coastal subsistence hunting and premoratorium commercial whaling operations). Samples from fin whales, *Balaenoptera physalus*, were collected in the North Atlantic, the Mediterranean Sea, and the Sea of Cortez in the North Pacific Ocean (Bérubé *et al.* 1998). Minke whale, *B. acutorostrata*, blue whale, *B. musculus*, and humpback whale, *Megaptera novaeangliae*, samples (Palsbøll *et al.* 1997a) were collected only in the North Atlantic (Table 1). Samples were preserved either by freezing at −20° to −80°, conservation in saturated NaCl with 20% DMSO (Amos and Hoelzel 1991), or both.

Genotyping of microsatellite loci: Total-cell DNA was extracted after standard procedures of cell lysis by addition of 1% SDS, overnight digestion with proteinase K, multiple extractions with phenol/chloroform, and finally ethanol precipitation (Maniatis *et al.* 1982). Three to six tri- and tetranucleotide repeat microsatellite loci were analyzed in each sample as described (Palsbøll *et al.* 1997b; Tables 1 and 3). In addition, the first 289–302 nucleotides of the mitochondrial control region were sequenced and the sex was determined for each sample following the procedures outlined in Palsbøll *et al.* (1995) and Bérubé and Palsbøll (1996a,b). Any two samples with identical genotypes at all analyzed microsatellite loci, mitochondrial control region sequences, and sex were inferred as duplicate samples from the same individual whale (in both the fin whale and blue whale one dinucleotide repeat microsatellite locus was analyzed as well). Using these criteria, the 3720 samples were collected from a total of 2975 individual whales (Table 1).

Number of alleles and allele-length distributions: Two kinds of intraspecific allele-length distributions were observed in the analyzed samples: "perfect allele arrays," in which the length of all alleles differed by simple integer multiples of the basic repeat length, and "imperfect allele arrays," where some alleles differed in length by other than simple integer multiples of the basic repeat length (see Table 2 for an example). The alleles at each imperfect allele array could be further subdivided into "subarrays," each containing alleles that differed in length only by simple multiples of the basic repeat length (Table 2).

Of the 18 species/loci combinations analyzed, 10 had imperfect and 8 perfect allele arrays (Table 3). Within and among species, we observed a higher number of alleles at loci with imperfect allele arrays relative to loci with perfect allele arrays. We observed an average of 8.2 (range: 6–11) and 14.6 (range: 8–27) alleles at loci with perfect and imperfect allele arrays, respectively, and between two and four subarrays at loci with imperfect allele arrays. Not surprisingly (given the difference in the number of alleles), we estimated a higher degree of heterozygosity (H) at loci with imperfect allele arrays as well (Table 7).

To test if the observed number of alleles and heterozygosity

TABLE 1

Number and origin of samples

Species	Samples	Individuals[a]	Sampling localities
B. acutorostrata[b]	161	160	North Atlantic: Barents Sea, Gulf of Maine, Gulf of St. Lawrence, Iceland, West and East Greenland
B. musculus[b]	92	89	Western North Atlantic: Gulf of St. Lawrence, West Greenland
B. physalus[c]	407	358	North Atlantic: Gulf of Maine, Gulf of St. Lawrence, off West Greenland, Iceland, eastern Spain Mediterranean Sea: Ligurian Sea North Pacific: Sea of Cortez
M. novaeangliae[d]	3,060	2,368	North Atlantic: Eastern Canada, Barents Sea, Gulf of Maine off Iceland, West Greenland

[a] Samples with identical mitochondrial control region sequence, sex, and genotype across all analyzed loci were considered duplicate samples from the same individual.
[b] P. Palsbøll (unpublished data).
[c] Bérubé *et al.* (1998)
[d] Palsbøll *et al.* (1997a)

at perfect loci indeed was significantly lower than that of loci with imperfect allele arrays, we ranked the observed number of alleles (Table 3) or estimated heterozygosity (Table 7) within each species. The test statistic (S_{OBS}) was calculated as the sum, across all species, of the ranks assigned to the loci with perfect allele arrays.

The probability of S_{OBS} was estimated from 10,000 permutations. For each permutation and each species, the observed ranks were randomly reassigned to the analyzed loci, and the sum of the ranks (S_{SIM}) assigned to loci with perfect allele arrays was calculated. The probability of S_{OBS} was estimated as the proportion of simulations where S_{SIM} was equal or smaller than S_{OBS}. The tests did not include the data from *B. musculus*, as only loci with perfect allele arrays were observed in this species.

The probability of the observed ranking regarding the number of alleles (Table 3) and estimated heterozygosity (Table 7) at loci with perfect and imperfect allele arrays was estimated at 0.0086 (S_{OBS} = 10.5) and 0.073 (S_{OBS} = 10.5), respectively. This result implied that a significantly higher number of alleles was observed at loci with imperfect allele arrays. The degree of heterozygosity was similarly higher, but not significantly so, at loci with imperfect allele arrays.

Sequence analysis of locus GATA028 alleles: To gain further insight into the kind of changes at the sequence level that generated the imperfect allele arrays, we sequenced individual alleles of different lengths at locus GATA028 in fin and humpback whale samples.

For the fin whale, one copy of each allele length detected among the 358 individual whales analyzed was sequenced (a total of 19 alleles). The alleles were preferably isolated and sequenced in homozygous individuals. Alleles not detected in a homozygous state were amplified and sequenced in the heterozygous individual, where we observed the largest difference in allele lengths. In practice, this meant that the sequenced alleles were sampled from several different and quite divergent populations, such as the Sea of Cortez, the Mediterranean Sea, and the Gulf of St. Lawrence. In the humpback whale, only two alleles of each subarray were sequenced.

Individual alleles were sequenced directly from asymmetrically amplified PCR products after an initial symmetrical amplification (Gyllensten and Erlich 1988). For alleles found only in heterozygous individuals, the symmetrical amplification products were separated before the subsequent asymmetrical amplification by electrophoresis through 4% NuSieve low-melting agarose, and the relevant band was excised and dissolved in distilled water. Symmetrical and asymmetrical amplifications were performed under conditions similar to those used during the population analyses, except that both oligonucleotide primers (the same as used for the population analyses) were added in 1 μm concentrations. For the asymmetrical amplifications, the concentration of the limiting oligonucleotide primer was reduced to 0.01 μm. Symmetrical and asymmetrical amplifications were performed in 10- and 50-μl volumes, respectively.

The limiting oligonucleotide primer used for the asymmetrical amplification was used as a sequencing primer following the manufacturer's instructions (Sequenase Version 2.0; United States Biochemical, Cleveland). The sequence reaction products were separated by electrophoresis, as described for the population analyses, and visualized by overnight autoradiography.

Nucleotide composition of locus GATA028 alleles: The alleles at locus GATA028 sequenced in the fin whale could be divided into four categories, each corresponding to the four subarrays identified in the population analyses. The sequenced alleles of the subarray denoted 1 (Table 4) all con-

TABLE 2

Allele lengths and frequencies at locus GATA028 in the Gulf of St. Lawrence fin whale sample

	Subarray[a]			
Allele length[b]	0	1	2	3
143	14			
151	2			
156		13		
160		23		
164		11		
167	1			
168		5		
171	8			
172		4		
173			6	
175	22			
179	36			
183	26			
186				2
187	21			

[a] See Table 4. The subarray designations 0–3 indicate the difference in allele length relative to the alleles in subarray 0 (which contain the shortest allele) and, thus, do not reflect any actual sequence changes.
[b] Lengths in base pairs.

TABLE 3

Number of alleles and subarrays per locus

			n_a for each subarray[c]			
Species and locus	$2n$[a]	n_a[b]	0	1	2	3
B. acutorostrata						
GATA028[d]	320	14	2	1	11	
GATA098	320	7				
GATA417[d]	320	12	4			8
B. musculus						
GATA028	178	8				
GATA098	178	8				
ACCC392	178	6				
GATA417	178	11				
B. physalus						
TAA023[d]	716	8	7	1		
GATA028[d]	716	19	10	6	2	1
GATA053[d]	716	13	5	7		1
GATA098	716	8				
GGAA520[d]	716	17	11		6	
M. novaeangliae						
GATA028[d]	4736	10	3			7
TAA031[d]	4736	14	7		7	
GATA053	4736	9				
GATA098	4736	8				
GATA417[d]	4736	17	7		2	8
GGAA520[d]	4736	27	17	4	1	5

[a] Number of genotyped chromosomes.
[b] Number of different alleles.
[c] The subarray designations 0–3 indicate the difference in allele length relative to the alleles in subarray 0.
[d] Loci with imperfect allele arrays.

TABLE 4

Nucleotide sequences of the microsatellite array for alleles detected at locus GATA028 in fin and humpback whales

Species	Subarray[a]	Nucleotide sequence of microsatellite array
B. physalus	0	$(gata)_{5–18}$ (gta gata gata gata)$_3$
	1	$(gata)_{10–16}$ (gta gata gata gata)$_2$
	2	$(gata)_8$ (ta) $(gata)_{4–6}$ (gta gata gata gata)$_3$
	3	$(gata)_2$ (gat) $(gata)_{13}$ (gta gata gata gata)$_3$
M. novaeangliae	0	$(gata)_{5–7}$ (gta gata gata gata)
	3	$(gata)_{5–15}$ (gta gata gata)$_2$

[a] The subarray designations 0–3 indicate the difference in allele length relative to the alleles in subarray 0 (which contain the shortest allele) and, thus, *do not* reflect any actual sequence changes.

tained a duplicated, 15-nucleotide repeat at the 3′ end of the microsatellite array, each composed of one imperfect (GTA) followed by three perfect (GATA) repeats. All the remaining three subarrays (denoted 0, 2, and 3; Table 4) also contained the 15-nucleotide repeat, but in these alleles, it was repeated three times. Of these last three subarrays, two contained imperfect repeats (a TA or a GAT repeat, subarrays 2 and 3, respectively; Table 4) within what was a perfect array of GATA repeats in the third subarray (subarray 0; Table 4).

The sequences of GATA028 alleles in the humpback whale could also be subdivided into two categories, each corresponding to the two observed subarrays. Alleles belonging to the subarray denoted 0 (Table 4) contained a 15-nucleotide repeat sequence at the 3′ end that was identical to the one found in the fin whales, although not repeated. Alleles of the subarray denoted 3 (Table 4) in the humpback whale did not contain the 15-nucleotide repeat sequence, but rather they contained a duplicated 11-nucleotide repeat sequence consisting of one imperfect (GTA) repeat followed by two perfect (GATA) repeats (Table 4).

The nucleotide sequences of alleles at locus GATA028 revealed that the main mutational mechanism within each subarray at loci with imperfect allele arrays probably was (as anticipated for microsatellite loci) single-strand slippage of perfect GATA repeats. However, the mutations responsible for the transitions between subarrays were imperfect mutations, *i.e.*, not simple loss or gain of single, perfect repeats. Two kinds of imperfect mutations were observed: gain or loss (presumably by single-strand slippage) of multiple repeats, of which one was an imperfect repeat (*e.g.*, the 15- or 11-nucleotide repeat sequences in the fin and humpback whale, respectively), or single-strand slippage involving partial repeats. Alternatively, the latter kind of imperfect mutations could also result from a deletion of one or two nucleotides not generated by single-strand slippage.

The new allele generated from such an imperfect mutation may differ in length from the parental allele by other than a simple integer multiple of the basic repeat length, as observed in the present study. Hence, the new allele, as well as its descendant alleles generated by single-strand slippage of perfect repeats, will form a lineage (or subarray) that is readily distinguishable from other alleles by molecular weight alone. The occurrence of imperfect mutations thus explained why we observed an elevated number of alleles at loci with imperfect allele arrays. In the absence of imperfect mutations, many mutations will yield allele lengths that already are present in the population and, thus, do not add to the overall number of discernible alleles.

Relative rate of imperfect to perfect mutations: The fact that we observed imperfect allele arrays at 10 of 18 loci indicated that imperfect mutations were relatively frequent. To obtain an estimate of the frequency of imperfect mutations from the combined data sets of all four species, we estimated the frequency of imperfect mutations as the relative rate (R) of imperfect to perfect mutations. For simplicity, we assumed that all perfect mutations were stepwise mutations. R was defined as

$$R = \frac{\mu_{[I]}}{\mu_{[S]}}$$

and estimated as

$$\hat{R} = \frac{\Sigma_i \hat{\theta}_{[I]_i}}{\Sigma_i \hat{\theta}_{[S]_i}}, \quad (1)$$

where $\hat{\theta}_{[I]_i}$ and $\hat{\theta}_{[S]_i}$ are the estimates of the composite parameters $\theta_{[I]}$ and $\theta_{[S]}$, respectively, at the ith locus. The parameters $\theta_{[I]}$ and $\theta_{[S]}$ equal $4N_e\mu_{[I]}$ and $4N_e\mu_{[S]}$, where N_e denotes the effective population size, and $\mu_{[I]}$ and $\mu_{[S]}$ denote the mutation rate of imperfect and single-step mutations, respectively. The term $\mu_{[S]}$ is equal to the mutation rate under a symmetrical single-step model. Under other and less simple mutation models (*e.g.*, asymmetrical and multistep mutations), the term $\mu_{[S]}$ is equal to the product of the mutation rate and the variance of the symmetrized distribution of changes in allele size (see Kimmel and Chakraborty 1996; Kimmel *et al.* 1996). Hence, in principle, the estimations below are valid for other and more complicated stepwise mutation models than the simple symmetrical single-step mutation model.

Estimation of R: Depending on the rate and nature of the imperfect mutations, the parameter $\theta_{[I]}$ can be estimated under either a single-step mutation model and/or an infinite allele model. The parameter $\theta_{[S]}$, however, is most appropriately estimated under a stepwise model.

Estimation of $\theta_{[S]}$ at a single locus: We estimated $\theta_{[S]}$ at each locus under the simplest possible stepwise mutation model, namely gain or loss of only a single repeat, each with an equal probability. Under such a strict single-step mutation model and assuming equilibrium conditions, $\theta_{[S]}$ can be estimated from the sample variance in repeat number per chromosome at the locus, *i.e.*,

$$\hat{\theta}_{[S]} = \frac{2}{(n-1)} \sum_i (j_i - \bar{j})^2, \quad (2)$$

where n is the number of chromosomes sampled, j_i is the number of repeats detected at the ith copy, and $\bar{j}$ is the mean number of repeats for all sampled chromosomes (Moran 1975; Valdes *et al.* 1993).

It is straightforward to estimate $\theta_{[S]}$ in this manner at loci with perfect allele arrays, as the variance can be estimated directly from relative difference in allele lengths divided by the repeat length. However, for loci with imperfect allele arrays, we cannot deduce the relative difference in the number of repeats between alleles from different subarrays unless the nucleotide composition of alleles at each subarray is known (which was the case for only two species/locus combinations in this study). An overall estimate of the parameter $\theta_{[S]}$ could be obtained by simply adding the contribution from each subarray, *i.e.*,

$$\hat{\theta}_{[S]} = \sum_j \hat{\theta}_{[S]j}, \quad (3)$$

where $\hat{\theta}_{[S]j}$ is the estimate of $\theta_{[S]}$ for the jth subarray obtained

as described by Equation 2. This approach is similar to that suggested by Hudson and Kaplan (1986), who found that the expected number of segregating sites in a nested subsample (based on allelic class) was approximately equal to the population frequency of the subsample times the expected number of segregating sites in the entire sample. Hence, if $E(\hat{\theta}_{[S]j}) = \theta_S x_j$, where x_j is the population frequency of the jth subarray, it follows that $E(\hat{\theta}_{[S]}) \approx \theta_{[S]}$.

Estimation of $\theta_{[I]}$ at a single locus: The parameter $\theta_{[I]}$ was estimated as $\hat{\theta}_{[I]}$ from the heterozygosity in the sample using the bias correction suggested by Chakraborty and Weiss (1991) as the solution to the equation

$$\hat{\theta}_{[I]}^3 + (7 - t)\hat{\theta}_{[I]}^2 + (8 - 5t)\hat{\theta}_{[I]} - 6t = 0, \qquad (4)$$

where $t = H/(1 - H)$ and $H = 1 - \Sigma_i x_i^2$, where x_i is the frequency of the ith allele.

Evaluating the estimation of R: To evaluate if indeed Equation 1 provided an unbiased estimate of R, coalescence simulations were performed as described by Hudson (1990). During these simulations, two kinds of mutations were allowed: single-step gain or loss of repeats, each with a mutation rate of $\theta_{[S]}/2$, as well as less frequent mutations, at a rate of $\theta_{[I]}$, each generating a new discernible allele, *i.e.*, corresponding to imperfect mutations. For each combination of $\theta_{[I]}$ and $\theta_{[S]}$, we conducted 1000 simulations, each with six loci and 200 chromosomes, where $\theta_{[I]}$, $\theta_{[S]}$, and R were estimated as $\hat{\theta}_{[I]}$, $\hat{\theta}_{[S]}$, and $\hat{R}$ in the manner described above (Equations 4, 3, and 1, respectively).

Our simulations revealed that R was consistently overestimated over a wide range of parameter values (see Table 5) when $\theta_{[S]}$ was estimated from subarrays (Equation 2). The degree of bias, however, was ~40% and did not appear to be affected by the value of $\theta_{[I]}$, $\theta_{[S]}$, or R. The bias was mainly caused by underestimation of $\theta_{[S]}$ when estimated from subarrays (Equation 3).

The severity of the bias introduced by the estimation of R in the above manner from the subarrays should be evaluated in terms of the overall variance in the estimation of R. As is evident from Table 5, the variance of R is quite considerable and exceeds by far the bias introduced by the estimation from subarrays for the values of $\hat{\theta}_{[I]}$, $\hat{\theta}_{[S]}$, and R observed in this study (Table 7).

Effects of population expansion on the estimation of R: While R is the ratio of the same parameter ($4N_e\mu$) for two different kinds of mutations, the estimates of $\theta_{[I]}$ and $\theta_{[S]}$ are, however, obtained from two different aspects of the data. Chakraborty and Kimmel have recently shown (Chakraborty *et al.* 1997; Kimmel *et al.* 1998) that these two aspects respond differently to temporal changes in N_e, and, thus, our estimate of R may not only reflect the ratio $\mu_{[I]}/\mu_{[S]}$ during and after changes in N_e.

Analyses of mitochondrial control region sequences in the samples included in this study using the program Fluctuate in the Lamarc computer package (Kuhner *et al.* 1998) indicated that several of the populations included in this study probably have growth rates that deviate significantly from zero (P. J. Palsbøll, unpublished data; data and results not shown). We investigated the possible effect of such changes in N_e to our estimates of R by coalescence simulations. The simulations were conducted under a model of exponential growth in the manner described by Slatkin and Hudson (1991), with two kinds of mutations corresponding to either an infinite allele or a stepwise mutation model and equivalent to $\theta_{[I]}$ and $\theta_{[S]}$, respectively. Simulations were performed with parameter values of $\theta_{[I]}$ and $\theta_{[S]}$ ranging from 0.001 to 100, and α (rN_e where r is the growth rate) ranging from 5 to 5000 (Table 6). A total of 1000 simulations were undertaken per combination of $\theta_{[I]}$, $\theta_{[S]}$, and α, each with six loci and 200 chromosomes. The estimate of R was obtained for each simulation using Equations 1–4.

Although the simulations revealed that $\hat{\theta}_{[S]}$ and $\hat{\theta}_{[I]}$ (Equations 3 and 4) underestimated $\theta_{[I]}$ and $\theta_{[S]}$ during population growth, the bias of the estimate of R itself was relatively modest (Table 6). The simulations that yielded mean values of $\hat{\theta}_{[S]}$, $\hat{\theta}_{[I]}$, and $\hat{R}$ observed during this study indicated that R (on average) was underestimated by ~20% (Table 6).

Observed estimates of R: Using Equations 1–4, we estimated R, the relative mutation rate of the imperfect to single-step mutations, at all loci with imperfect allele arrays.

As explained above, the estimations rely on population equilibrium conditions, *i.e.*, constant population size, no recombination, and that the sampled chromosomes are from a single, panmictic population with no migration. It is not possible with the current knowledge to assess if all these assumptions are met for all the species and populations included in this study. However, to minimize possible violations of the assumptions, we did confine our estimation of R to populations that are currently believed to constitute part of a single panmictic population (although migration most likely does occur). The analyzed populations were West Greenland minke whales ($n = 69$), western North Atlantic blue whales ($n = 89$), Gulf of St. Lawrence fin whales ($n = 97$), and West Indian humpback whales (1992 only, $n = 596$). Additional estimations were also obtained from the Sea of Cortez ($n = 51$) and Mediterranean Sea ($n = 58$) fin whale populations. As mentioned above, an analysis of the mitochondrial control region sequences using the program Fluctuate in the Lamarc computer package (Kuhner *et al.* 1998) based upon the population samples used in this study yielded estimates of growth rates in blue and humpback whale populations that did not deviate significantly from zero (data not shown). For the remaining populations, we estimated positive growth rates that deviated significantly from zero. The only exception were the Mediterranean Sea fin whales, where we estimated a negative growth rate that deviated significantly from zero.

Estimates of $\theta_{[S]}$ for individual loci ranged from 0 to 430, with most values in the range of 5–30 (Table 7). Extreme values outside this range (*e.g.*, *B. musculus*, locus GATA098, $\hat{\theta}_{[S]} = 430$; Table 7) did have allele frequency distributions that deviated significantly from the null expectations under a single-step mutation model (for further discussion see Nielsen and Palsbøll 1999).

The estimates of R obtained at loci with imperfect allele arrays ranged from 0 to 0.065 (Table 8), with an overall mean of 0.024. The highest estimates of R were observed in the minke whale (*B. acutorostrata*; Table 8) and fin whale (*B. physalus*; Table 8), where analysis of mitochondrial control region sequences indicated population growth (data not shown). Hence, it appears (as our simulations suggested) that the population expansions have not greatly influenced our estimate of R. Our results imply that (on average) ~2.4% of the mutations at these tri- and tetranucleotide repeat microsatellite loci were imperfect mutations, *i.e.*, mutations other than simple gain or loss of perfect repeats (Table 8).

Separate estimates from other fin whale populations in the Mediterranean Sea and the Sea of Cortez yielded similar estimates of R (Table 9).

Because the nucleotide sequence of each allele length detected at locus GATA028 was known in the fin and humpback whales, we were able to estimate $\theta_{[S]}$ directly from the variance in repeat number (Equation 2) after exclusion of the imperfect mutations responsible for the generation of subarrays (Table 4). During this estimation, we assumed that all copies of equal length had a nucleotide sequence similar to that of the sequenced allele (Table 4).

The values of R estimated in this manner at locus GATA028,

TABLE 5

Estimated values of $\hat{R}$ from simulations under a model of constant population size

Parameter values			Estimated values					
$\theta_{[I]}$	$\theta_{[S]}$	$\hat{\theta}_{[I]}$[a]	$V_{\hat{\theta}[I]}$[b]	$\hat{\theta}_{[S]}$[c]	$V_{\hat{\theta}[S]}$[b]	$\hat{R}$[d]	$V_{\hat{R}}$[b]	Bias $\hat{R}/R$
0.001	1.0	0.00092	0.00007	1.0	0.26	0.0013	0.00021	1.32
0.005	1.0	0.0061	0.00046	0.98	0.24	0.0074	0.00082	1.48
0.01	1.0	0.012	0.00088	0.96	0.26	0.014	0.0016	1.41
0.05	1.0	0.059	0.0047	0.99	0.29	0.074	0.011	1.48
0.1	1.0	0.11	0.0088	0.99	0.26	0.14	0.022	1.43
0.005	5.0	0.0073	0.00062	5.1	7.6	0.0017	0.000038	1.66
0.025	5.0	0.029	0.0023	4.9	4.7	0.0071	0.00016	1.42
0.05	5.0	0.057	0.0045	4.8	5.4	0.014	0.00032	1.37
0.25	5.0	0.27	0.023	4.3	3.1	0.071	0.0024	1.41
0.5	5.0	0.54	0.046	4.3	3.1	0.14	0.0053	1.41
0.01	10	0.011	0.00081	9.8	23	0.0013	0.000015	1.29
0.05	10	0.057	0.0048	9.8	23	0.0069	0.000087	1.38
0.1	10	0.11	0.0096	9.6	18	0.013	0.00016	1.35
0.5	10	0.54	0.046	8.5	11	0.070	0.0012	1.41
1	10	1.0	0.097	7.9	6.2	0.14	0.0033	1.41
0.025	25	0.026	0.0020	25	140	0.0012	0.000005	1.25
0.125	25	0.14	0.011	24	110	0.0068	0.000034	1.36
0.25	25	0.28	0.025	23	100	0.014	0.000098	1.41
1.25	25	1.27	0.12	20	38	0.069	0.00065	1.38
2.5	25	2.4	0.31	18	17	0.14	0.0017	1.40
0.05	50	0.056	0.0041	48	480	0.0014	0.000003	1.36
0.25	50	0.27	0.023	44	310	0.0068	0.000018	1.35
0.5	50	0.54	0.045	43	230	0.014	0.000049	1.40
2.5	50	2.4	0.31	37	74	0.068	0.00042	1.36
5	50	4.7	0.72	35	44	0.14	0.0012	1.40
0.1	100	0.12	0.0087	95	1800	0.0014	0.000002	1.44
0.5	100	0.54	0.051	86	990	0.0070	0.000014	1.41
1	100	1.0	0.093	79	670	0.014	0.000029	1.40
5	100	4.7	0.69	68	170	0.070	0.00026	1.40
10	100	9.2	1.8	63	83	0.15	0.00075	1.47
Mean								1.40

[a] Equation 4.
[b] The observed variance.
[c] Equations 2 and 3.
[d] Equation 1.

in three fin whale populations and one humpback whale population (Table 10), yielded estimates of $\theta_{[S]}$ that were approximately half of the estimates obtained by our indirect approach (Equation 3). In all four cases, the estimate of R was at least twice that of the estimate obtained by the indirect approach (Equation 3). Given the large variance in the estimation of $\theta_{[S]}$ itself from the number of repeats (Equation 5) and the fact that some of the populations share a recent common ancestry and, thus, do not constitute independent observations, no generalizations can be drawn from these relatively few observations.

The values in Table 8 suggested a positive correlation between $\theta_{[I]}$ and $\theta_{[S]}$. The existence of such a correlation was assessed by using the same approach that was used when testing whether the observed number of alleles and heterozygosity was higher at loci with imperfect allele arrays compared to loci with perfect allele arrays (see above). The loci within each species (Table 8) were ranked according to $\hat{\theta}_{[S]}$ and subsequently partitioned into loci with perfect or imperfect allele arrays. The probability of the observed sum of the ranks for the loci with perfect allele arrays ($S_{OBS} = 9.0$) was estimated from 10,000 Monte Carlo simulations to 0.025, which implies there was a positive correlation between $\theta_{[I]}$ and $\theta_{[S]}$.

DISCUSSION

Multiple levels of single-strand slippage at microsatellite arrays: The findings of this study suggest that single-strand slippage mutations at microsatellite loci involve not only single-step mutations, but also relatively high frequencies of multi- as well as partial-repeat mutations. The frequency of the two latter categories of mutations was estimated at a mean of 2.5% of the rate of single-step mutations. The estimate was obtained from several

TABLE 6

Estimates of $\hat{R}$ under a model of population expansion

Parameter values				Estimated values						
α	$\theta_{[I]}$	$\theta_{[S]}$	R	$\hat{\theta}_{[I]}{}^{a}$	$V_{\hat{\theta}_{[I]}}{}^{b}$	$\hat{\theta}_{[S]}{}^{c}$	$V_{\hat{\theta}_{[S]}}{}^{b}$	$\hat{R}^{d}$	$V_{\hat{R}}{}^{b}$	Bias $\hat{R}/R$
5	0.05	10	0.005	0.092	0.035	18	26	0.053	0.00012	1.05
10	0.05	10	0.005	0.064	0.021	12	11	0.0055	0.00017	1.09
50	0.05	10	0.005	0.020	0.0062	4.1	1.4	0.0050	0.00042	1.00
100	0.05	10	0.005	0.012	0.0037	2.5	0.56	0.0049	0.00057	0.98
500	0.05	10	0.005	0.0026	0.00026	0.68	0.11	0.046	0.00084	0.91
1000	0.05	10	0.005	0.0029	0.00091	0.37	0.047	0.0081	0.0045	1.62
5000	0.05	10	0.005	0.00030	0.000007	0.091	0.0084	0.0049	0.0034	0.98
5	1.00	10	0.1	2.1	1.0	21	32	0.10	0.0026	1.01
10	1.00	10	0.1	1.3	0.53	14	14	0.094	0.0027	0.94
50	1.00	10	0.1	0.39	0.11	4.6	1.8	0.086	0.0051	0.86
100	1.00	10	0.1	0.23	0.050	2.6	0.67	0.091	0.0086	0.91
500	1.00	10	0.1	0.061	0.011	0.70	0.10	0.10	0.033	1.01
1000	1.00	10	0.1	0.036	0.0092	0.38	0.047	0.12	0.10	1.18
5000	1.00	10	0.1	0.0089	0.0017	0.094	0.0089	0.17	1.1	1.67
5	0.25	50	0.005	0.48	0.19	95	640	0.0052	0.000023	1.04
10	0.25	50	0.005	0.32	0.11	64	250	0.0051	0.000028	1.02
50	0.25	50	0.005	0.097	0.026	21	18	0.0046	0.000057	0.91
100	0.25	50	0.005	0.062	0.016	13	6.3	0.0048	0.000088	0.97
500	0.25	50	0.005	0.017	0.0036	3.5	0.75	0.0050	0.00027	1.00
1000	0.25	50	0.005	0.0085	0.0016	1.9	0.31	0.0042	0.00034	0.85
5000	0.25	50	0.005	0.0017	0.00044	0.48	0.051	0.0041	0.0025	0.81
5	5.00	50	0.1	14	14	150	750	0.099	0.00072	0.99
10	5.00	50	0.1	8.9	7.1	110	370	0.086	0.00064	0.86
50	5.00	50	0.1	2.4	1.0	32	44	0.075	0.00086	0.75
100	5.00	50	0.1	1.2	0.41	18	15	0.070	0.0011	0.70
500	5.00	50	0.1	0.31	0.066	4.0	1.2	0.078	0.0035	0.78
1000	5.00	50	0.1	0.17	0.032	2.1	0.42	0.081	0.0065	0.81
5000	5.00	50	0.1	0.042	0.0069	0.50	0.066	0.096	0.039	0.96
5	0.50	100	0.005	0.99	0.42	200	2600	0.0052	0.000012	1.04
10	0.50	100	0.005	0.66	0.26	130	860	0.0050	0.000014	1.00
50	0.50	100	0.005	0.19	0.048	44	69	0.0043	0.000021	0.87
100	0.50	100	0.005	0.10	0.022	25	21	0.0041	0.000034	0.82
500	0.50	100	0.005	0.029	0.0046	6.9	1.7	0.0043	0.000091	0.86
1000	0.50	100	0.005	0.017	0.0023	3.9	0.64	0.0043	0.00014	0.86
5000	0.50	100	0.005	0.0033	0.00014	0.98	0.11	0.0039	0.00018	0.77
5	10.00	100	0.1	36	46	320	2400	0.11	0.00055	1.13
10	10.00	100	0.1	24	33	260	1700	0.096	0.00052	0.96
50	10.00	100	0.1	5.5	3.2	86	220	0.064	0.00034	0.64
100	10.00	100	0.1	2.8	1.2	45	67	0.062	0.00043	0.62
500	10.00	100	0.1	0.65	0.17	9.3	5.2	0.070	0.0015	0.70
1000	10.00	100	0.1	0.35	0.067	4.8	1.8	0.074	0.0025	0.74
5000	10.00	100	0.1	0.085	0.014	0.98	0.14	0.094	0.019	0.94

[a] Equation 4.
[b] The observed variance.
[c] Equations 2 and 3.
[d] Equation 1.

loci and across four different species. The multistep mutations detected in this study included an imperfect repeat, and, thus, were contingent on a previous imperfect mutation, *i.e.*, by partial-repeat slippage. Of the four imperfect mutations detected from the sequences at locus GATA028, two involved multiple repeats. Hence, our study yielded an approximate rate of multirepeat and partial-repeat mutations of roughly 1.25% each. As our study only detected multirepeat mutations that included an imperfect repeat, this rate is most likely an underestimate of the overall rate of multirepeat mutations. The occurrence of imperfect mutations was not

TABLE 7

Observed values of $\hat{\theta}_{[S]}$ and heterozygosity ($\hat{H}$) for selected populations

Species and locus	$2n$[a]	$\theta_{[S]}$[b]	$\hat{V}_{\hat{\theta}_{[S]}}$[c]	$\hat{H}$
B. acutorostrata				
GATA028	138	6.3	1.3	0.80
Subarray 0	9	0.56		
Subarray 1	1	0.0		
Subarray 2	121	5.7		
GATA098	138	3.1	0.14	0.74
GATA417	138	7.3	0.76	0.86
Subarray 0	69	1.1		
Subarray 3	69	6.2		
B. musculus				
GATA028	178	7.4	1.6	0.77
GATA098	178	7.6	2.1	0.73
ACCC392	178	430	2500	0.71
GATA417	178	13	1.3	0.86
B. physalus				
TAA023	194	7.1	0.24	0.74
GATA028	194	23	13	0.89
Subarray 0	130	20		
Subarray 1	56	2.7		
Subarray 2	6	0.0		
Subarray 3	2	0.0		
GATA053	194	28	3.8	0.85
Subarray 0	29	18		
Subarray 1	147	10		
Subarray 3	18	0.0		
GATA098	194	4.2	0.14	0.80
GGAA520	194	9.8	0.90	0.89
Subarray 0	163	7.0		
Subarray 1	31	2.8		
M. novaeangliae				
GATA028	1192	29	1.0	0.48
Subarray 0	1002	0.89		
Subarray 3	190	28		
TAA031	1192	30	6.3	0.82
Subarray 0	218	26		
Subarray 2	974	3.9		
GATA053	1192	12	0.10	0.82
GATA098	1192	25	0.34	0.68
GATA417	1192	9.2	0.079	0.87
Subarray 0	467	4.0		
Subarray 2	21	0.38		
Subarray 3	704	4.8		
GGAA520	1192	48	2.5	0.81
Subarray 0	1099	46		
Subarray 1	54	2.1		
Subarray 2	1	0.0		
Subarray 3	38	0.0		

[a] Number of chromosomes in sample.
[b] Equation 3.
[c] The sampling variance of $\theta_{[S]}$ estimated from 10,000 bootstrap samples.

confined to a single species, locus, or population, but was detected across several species and loci, arguing that imperfect mutations are relatively common phenomena.

The imperfect mutations, which we interpreted as partial-repeat slippage, could also be indels not generated by single-strand slippage. However, single-strand slippage appears to be the most likely mutational mechanism for generating the imperfect mutations observed at locus GATA028 for the following reasons:

The nucleotide sequences of the alleles at locus GATA028 contained as many nucleotides from the flanking regions as from the microsatellite array (data not presented); however, neither indels nor any nucleotide substitutions were observed in the flanking regions.

All the inferred partial-repeat changes were located within a stretch of perfect repeats where single-strand slippage is presumably the main mutational mechanism.

The apparent positive correlation of $\theta_{[I]}$ with $\theta_{[S]}$.

The two imperfect repeats generated from these mutations consisted of partial GATA repeats (GAT or TA).

The sequence data presented by Estoup *et al.* (1995) and Angers and Bernatchez (1997) also suggest partial-repeat slippage mutations, although at an interspecific level.

As suggested for minisatellites (Monckton *et al.* 1994), the mutation in the repeat array could also be influenced by elements in the flanking regions; however, the present data do not allow for the testing of such a possibility.

Constraints on allele size as a result of partial-repeat mutations: While multirepeat mutations have been presented earlier, partial-repeat mutations within the microsatellite array are not commonly reported. Imperfections in the repeat array of an allele appear to reduce (Weber 1990) or completely halt the rate of single-strand slippage mutations (Jin *et al.* 1996). Hence, imperfections in the repeat array may, in part, provide a mechanism that would counteract the expansion in overall allele length caused by a mutational bias toward a gain of repeats as reported recently by Amos and Rubinsztein (1996b) and Primmer *et al.* (1996).

A number of deleterious diseases, *e.g.*, Huntington's disease (Duyao *et al.* 1993), have been shown to be caused by a rapid increase in the number of repeats at specific microsatellite loci, and, thus, selection could also hinder expansion of microsatellite loci. However, as many microsatellite loci are situated in noncoding DNA sequences, selection does not appear to be the sole mechanism preventing a continuous expansion.

Partial-repeat mutations may partly counteract continuous expansion of the repeat number at neutral microsatellite loci by generating imperfections in the microsatellite array. We tested the effects of partial-repeat mutations on the overall number of repeats by simulations. We assumed a biased (toward gain of repeats) single-step mutation model with an equal probability of a partial-repeat mutation per repeat in the microsatellite array. The occurrence of a partial-repeat mutation in a microsatellite array changed the rate of single-step

TABLE 8

Intraspecific estimates of *R*

Species and locus	$2n^a$	$\hat{\theta}_{[I]}{}^b$	$\hat{V}_{\hat{\theta}_{[I]}}{}^c$	$\hat{\theta}_{[S]}{}^d$	$\hat{R}^e$
B. acutorostrata	138				
GATA028		0.12	0.0014	6.3	
GATA098		0.00	—	3.1	
GATA417		0.74	0.00021	7.3	
Mean		0.37		5.6	0.065
B. musculus	178				
GATA028		0.00	—	7.4	
GATA098		0.00	—	7.6	
ACCC392		0.00	—	430	
GATA417		0.00	—	13	
Mean					0.00
B. physalus	194				
TAA023		0.00	—	7.1	
GATA028		0.65	0.0056	23	
GATA053		0.49	0.0062	28	
GATA098		0.00	—	4.2	
GGAA520		0.27	0.0025	9.8	
Mean		0.28		14	0.020
M. novaeangliae	1192				
TAA031		0.32	0.00047	30	
GATA028		0.27	0.00041	29	
GATA053		0.00	—	12	
GATA098		0.00	—	25	
GATA417		0.74	0.00040	9.2	
GGAA520		0.13	0.00019	48	
Mean		0.24		24	0.010

[a] Number of chromosomes in the sample.
[b] Equation 4.
[c] The sampling variance estimated from 10,000 bootstrap samples.
[d] From Table 7.
[e] Equation 1.

mutations from $\theta_{[S]}$ to zero. The prediction of such a model is that alleles with a high number of repeats on average are more prone to partial-repeat mutations than alleles with fewer repeats, which in turn will reduce the rate of single-strand slippage (in this case to zero). The proposed mechanism is consistent with the observation that some loci contain alleles with a large number of perfect repeats (Rico *et al.* 1994), or that some species are fixed for alleles with an imperfect repeat array.

A limited number of simulations, under the model proposed above using forward simulations with multinomial resampling of alleles over discrete generations and constant population size, did indeed confirm the predictions of the model (Figure 1). The presence of partial-repeat mutations reduced the increase in mean allele length relative to the absence of partial-repeat mutations. The number of simulations conducted was very limited and assumed that a partial-repeat mutation completely halted the rate of single-step mutations, which our own data indicate is not necessarily the case. A more thorough assessment is warranted over a wide range of parameter values before any firm conclusions can be drawn. However, this result indicates that a relatively minor extension of the main mutational mechanism at microsatellite loci could provide an explanation for the absence of continuous expansion of microsatellite loci, which is a logical consequence of the empirical data suggesting a mutational bias toward gain of repeats at microsatellite loci.

Consequence for detection and estimation of divergence: Our study revealed that approximately half of the imperfect mutations were multirepeat changes. This estimate is likely to be an underestimate because of the approach used in this study (see above). Amos and Rubinsztein (1996b) as well as Primmer *et al.* (1996) identified germ-line mutations, each at single microsatellite loci, and detected multirepeat mutations at a frequency of 0 and 18%, respectively. Di Rienzo *et al.* (1994) observed allele distributions at 8 out of 10 dinucleotide repeat microsatellite loci that were consistent with the occurrence of multirepeat mutations when compared to the null expectations under a strict single-step mutation model. Nielsen and Palsbøll (1999) estimated the frequency of multirepeat mutations at 9 microsatellite loci with perfect arrays in different baleen

TABLE 9

Estimates of *R* from additional fin whale populations (*B. physalus*) at loci GATA028, GATA053, and GGAA520

		Locus								
		GATA028			GATA053			GGAA520			
Locality	$2n^a$	$\hat{H}^b$	$\hat{\theta}_{[I]}{}^c$	$\hat{\theta}_{[S]}{}^d$	$\hat{H}^b$	$\hat{\theta}_{[I]}{}^c$	$\hat{\theta}_{[S]}{}^d$	$\hat{H}^b$	$\hat{\theta}_{[I]}{}^c$	$\hat{\theta}_{[S]}{}^d$	$\hat{R}^e$
Mediterranean Sea	128	0.89	0.73	26	0.74	0.25	25	0.90	0.43	12	0.022
Sea of Cortez	102	0.72	0.56	3.0	0.10	0.061	6.6	0.29	NA	2.4	0.065

NA, not available.
[a] Number of chromosomes in sample.
[b] Estimated heterozygosity.
[c] Equation 4.
[d] Equations 2 and 3.
[e] Equation 1.

TABLE 10

Estimates of $\theta_{[S]}$ and R at locus GATA028 directly from the sequence data

Species and population	$2n^a$	$\hat{H}^b$	$\hat{\theta}_{[I]}{}^c$	$\hat{V}_{\hat{\theta}_{[I]}}{}^d$	$\hat{\theta}_{[S]}{}^e$	$\hat{V}_{\hat{\theta}_{[S]}}{}^d$	$\hat{R}^f$
B. physalus							
Gulf of St. Lawrence	194	0.89	0.65		14 (23)		0.046
Mediterranean	128	0.89	0.73		17 (26)		0.043
Sea of Cortez	102	0.72	0.56		1.3 (3.0)		0.43
M. novaeangliae							
West Indies 1992	1192	0.48	0.27		11 (29)		0.025

[a] Number of chromosomes in sample.
[b] Heterozygosity.
[c] Equation 4.
[d] The sampling variance estimated from 10,000 bootstrap samples.
[e] Estimated using Equation 2 for perfect GATA repeats only directly from the sequence data, ignoring non-single-step mutations. Numbers in parentheses are the estimates of $\hat{\theta}_{[S]}$ obtained by adding separate estimates at each subarray (Equation 3).
[f] Equation 1.

whale populations using a maximum likelihood procedure (Nielsen 1997). They found significant deviations from the null expectations under a strict single-step mutation model, consistent with multirepeat mutations at 2 loci. The estimates of the frequency of multirepeat mutations most compatible with the observed data were 0.05 and 0.29, respectively.

The results from the above-mentioned studies as well as the present studies indicate that multirepeat mutations occur at a high proportion of loci. Multirepeat mutations will change the sample mean and increase the sample variance several repeat units in a single mutational event. In the present study, we observed two instances where one subarray was completely absent from one or several population samples (locus GATA028 and locus GGAA520, *B. physalus*, data not shown), which, of course, will affect the linear relationship between the microsatellite-specific statistics and divergence time (Goldstein *et al.* 1995a,b; Slatkin 1995). The increase in variance of the microsatellite-specific statistics caused by multirepeat mutations may have a considerable impact on the accuracy of studies of natural populations, which are typically based on analyses of a relatively modest number of loci (Zhivotovsky and Feldman 1995), and may explain why some population genetic studies find a poor correlation between geographic and interpopulation genetic distances (*e.g.*, Valsecchi *et al.* 1997).

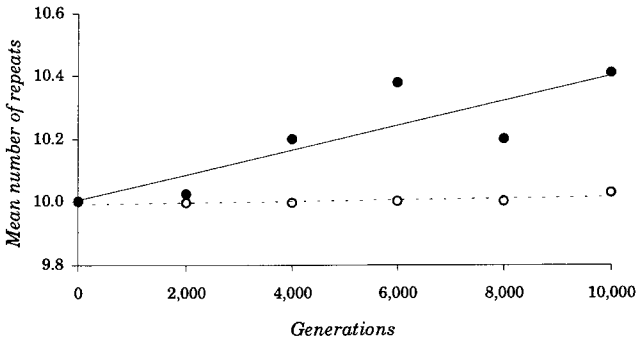

Figure 1.—Estimates of mean allele length with or without partial-repeat mutations. Estimates of mean number of repeats per allele after 2000–10,000 generations in a population of 1000 chromosomes and an initial allele size of 10 repeats at generation 0. A total of 100 simulations were conducted per estimate. The single-step mutation rate ($\theta_{[S]}$) was set at 10^{-4}, and the probability of a gain was set at 0.7. (○) A probability of 0.005 of a partial-repeat mutation per repeat, which reduced $\theta_{[S]}$ to zero. (●) Simulations under similar conditions, but with no partial-repeat mutations.

The partial-repeat mutations detected in the current study have an impact on the accuracy of divergence estimates obtained from statistics based on the number of alleles, such as Weir's θ (Weir 1990). As our results have shown, the number of alleles are correlated with the number of subarrays and, thus, partial-repeat mutations will increase the variance of such statistics.

The results from this and other studies (see above) show that the sequence changes observed at microsatellite loci do not follow a simple pattern, which presumably increases the variance of the current statistics proposed for estimating divergence from microsatellite data. Most studies of natural populations rely on the analysis of a relatively modest number of microsatellite loci, and, thus, the increase in variance is of concern and needs to be addressed. It may be that microsatellite loci with imperfect allele arrays, such as those described in the present study, constitute a useful class of loci, which possesses the high rate of mutation that is characteristic of microsatellite loci, but with an elevated number of alleles relative to perfect loci.

We thank the following institutions for donating samples: Allied Whale, Center for Coastal Studies, Cetacean Research Group at Memorial University, Department of Animal Biology at Barcelona University, Department of Marine Biology at University of Baja California, Fisheries Research Institute at Tromsø University, Greenland Natural Resources Institute, the Marine Research Institutes in Iceland and Norway, Mingan Island Cetacean Study, Inc., and Tethys. The majority

of the humpback whale samples was collected during the international collaborative project YoNAH (Years of the North Atlantic Humpback whale). In addition, we thank T. H. Andersen, T. P. Feddersen, C. Færch-Jensen, A. H. Larsen, K. B. Pedersen, D. Poulsen, P. Raahauge, R. Sponer, and E. Widén for technical assistance. This work was greatly improved by the valuable comments and suggestions from R. R. Hudson. M. Slatkin also provided useful comments on earlier drafts. We also owe thanks to one anonymous reviewer, who pointed out the possible effect of population growth to our estimations, and R. Nielsen for advice. R. R. Hudson and P. Arctander are thanked for their support. This project was in part funded by the Commission for Scientific Research in Greenland, the European Union Biotechnology Program (grant to P. Arctander), the Greenland Home Rule, the International Whaling Commission, the Natural Science Research Council (Denmark), World Wildlife Foundation (Denmark), and the Åge V. Jensen Charity Foundation.

LITERATURE CITED

Amos, W., and A. R. Hoelzel, 1991 Long-term preservation of whale skin for DNA analysis. Rep. Int. Whaling Comm. Spec. Issue **13:** 99–104.

Amos, W., and D. C. Rubinsztein, 1996a Microsatellites are subject to directional evolution. Nat. Genet. **12:** 13–14.

Amos, W., and D. C. Rubinsztein, 1996b Microsatellites show mutational bias and heterozygote instability. Nat. Genet. **13:** 390–391.

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

Bérubé, M., and P. Palsbøll, 1996a Erratum of identification of sex in cetaceans by multiplexing with three ZFX and ZFY specific primers. Mol. Ecol. **5:** 602.

Bérubé, M., and P. J. Palsbøll, 1996b Identification of sex in cetaceans by multiplexing with three *ZFX* and *ZFY* specific primers. Mol. Ecol. **5:** 283–287.

Bérubé, M., A. Aguilar, D. Dendanto, F. Larsen, G. Notarbartolo-di-Sciara *et al.*, 1998 Population genetic structure of North Atlantic, Mediterranean Sea and Sea of Cortez fin whales, *Balaenoptera physalus* (Linnaeus, 1758): analysis of mitochondrial and nuclear loci. Mol. Ecol. **7:** 585–600.

Chakraborty, R., and K. M. Weiss, 1991 Genetic variation of the mitochondrial DNA genome in American Indians is at mutation-drift equilibrium. Am. J. Phys. Anthropol. **86:** 497–506.

Chakraborty, R., M. Kimmel, D. N. Stivers, L. J. Davison and R. Deka, 1997 Relative mutation rate at di-, tri-, and tetranucleotide microsatellite loci. Proc. Natl. Acad. Sci. USA **94:** 1041–1046.

Clapham, P. J., P. J. Palsbøll and D. K. Mattila, 1993 High-energy behaviors in humpback whales as a source of sloughed skin for molecular analysis. Mar. Mamm. Sci. **9:** 213–220.

Di Rienzo, A., A. C. Peterson, J. C. Garza, A. M. Valdes, M. Slatkin *et al.*, 1994 Mutational processes of simple-sequence repeat loci in human populations. Proc. Natl. Acad. Sci. USA **91:** 3166–3170.

Duyao, M., C. Ambrose, R. Myers, A. Novelletto, F. Persichetti *et al.*, 1993 Trinucleotide repeat length instability and age of onset in Huntington's disease. Nat. Genet. **4:** 387–392.

Ellegren, H., C. R. Primmer and B. C. Sheldon, 1995 Microsatellite 'evolution': directional bias? Nat. Genet. **11:** 360–362.

Estoup, A., C. Tailiez, J.-M. Cornuet and M. Solignac, 1995 Size homoplasy and mutational processes of interrupted microsatellites in two bee species, *Apis mellifera* and *Bombus terrestris* (Apidae). Mol. Biol. Evol. **12:** 1074–1084.

Garza, J. C., and N. B. Freimer, 1996 Homoplasy for size at microsatellite loci in humans and chimpanzees. Genome Res. **6:** 211–217.

Garza, J. C., M. Slatkin and N. B. Freimer, 1995 Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. Mol. Biol. Evol. **12:** 594–603.

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

Goldstein, D. B., A. Ruiz Linares, L. L. Cavalli-Sforza and M. W. Feldman, 1995b Genetic absolute dating based upon microsatellites and the origin of modern humans. Proc. Natl. Acad. Sci. USA **92:** 6723–6727.

Grimaldi, M.-C., and B. Crouau-Roy, 1997 Microsatellite allelic homoplasmy due to variable flanking sequences. J. Mol. Evol. **44:** 336–340.

Gyllensten, U. B., and H. A. Erlich, 1988 Generation of single-stranded DNA by the polymerase reaction and its application to direct sequencing of the *HLA-DQA* locus. Proc. Natl. Acad. Sci. USA **85:** 7652–7656.

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxfords Surveys in Evolutionary Biology*, edited by D. J. Futuyma and J. Antonovics. Oxford University Press, Oxford.

Hudson, R. R., and N. L. Kaplan, 1986 On the divergence of alleles in nested subsamples from finite populations. Genetics **113:** 1057–1076.

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

Kimmel, M., and R. Chakraborty, 1996 Measures of variation at DNA repeat loci under a general stepwise mutation model. Theor. Popul. Biol. **50:** 345–367.

Kimmel, M., R. Chakraborty, D. N. Stivers and R. Deka, 1996 Dynamics of repeat polymorphisms under a forward-backward mutation model: within- and between-population variability at microsatellite loci. Genetics **143:** 549–555.

Kimmel, M., R. Chakraborty, J. P. King, M. Bamshad, W. S. Watkins *et al.*, 1998 Signatures of population expansion in microsatellite repeat data. Genetics **148:** 1921–1930.

Kuhner, M. K., J. Yamato and J. Felsenstein, 1998 Maximum likelihood estimation of population growth rates based on the coalescent. Genetics **149:** 429–434.

Levinson, G., and G. A. Gutman, 1987a High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. Nucleic Acids Res. **15:** 5323–5339.

Levinson, G., and G. A. Gutman, 1987b Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. **4:** 203–221.

Mahtani, M. M., and H. F. Willard, 1993 A polymorphic X-linked tetranucleotide repeat locus displaying a high rate of new mutation: implications for mechanisms of mutation at short tandem repeat loci. Hum. Mol. Genet. **2:** 431–437.

Maniatis, T., E. F. Fritsch and J. Sambrook, 1982 *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Messier, W., S.-H. Li and C.-B. Steward, 1996 The birth of microsatellites. Nature **381:** 483.

Monckton, D. G., R. Neumann, T. Guram, N. Fretwell, K. Tamaki *et al.*, 1994 Minisatellite mutation rate variation associated with a flanking DNA sequence polymorphism. Nat. Genet. **8:** 162–170.

Moran, P. A. P., 1975 Wandering distribution and the electrophoretic profile. Theor. Popul. Biol. **8:** 318–330.

Nielsen, R., 1997 A likelihood approach to population samples of microsatellite alleles. Genetics **146:** 711–716.

Nielsen, R., and P. J. Palsbøll, 1999 Single-locus tests of microsatellite evolution: multi-step mutations and constraints on allele size. Mol. Phylogenet. Evol. (in press).

Orti, G., D. E. Pearse and J. C. Avise, 1997 Phylogenetic assessment of length variation at a microsatellite locus. Proc. Natl. Acad. Sci. USA **94:** 10745–10749.

Palsbøll, P. J., F. Larsen and E. Sigurd Hansen, 1991 Sampling of skin biopsies from free-ranging large cetaceans in West Greenland: development of new biopsy tips and bolt designs. Rep. Int. Whaling Comm. Spec. Issue **13:** 71–79.

Palsbøll, P. J., P. J. Clapham, D. K. Mattila, F. Larsen, R. Sears *et al.*, 1995 Distribution of mtDNA haplotypes in North Atlantic humpback whales: the influence of behaviour on population structure. Marine Ecol. Prog. Ser. **116:** 1–10.

Palsbøll, P. J., J. Allen, M. Bérubé, P. J. Clapham, T. P. Feddersen *et al.*, 1997a Genetic tagging of humpback whales. Nature **388:** 676–679.

Palsbøll, P. J., M. Bérubé, A. H. Larsen and H. Jørgensen, 1997b Primers for the amplification of tri- and tetramer microsatellite loci in cetaceans. Mol. Ecol. **6:** 893–895.

Primmer, C. R., N. Saino and A. P. Møller, 1996 Directional evolu-

tion in germline microsatellite mutations. Nat. Genet. **13:** 391–393.

Rico, C., I. Rico and G. Hewitt, 1994 An optimized method for isolating and sequencing large (CA/GT)n (n > 40) microsatellites from genomic DNA. Mol. Ecol. **3:** 181–182.

Rubinsztein, D. C., W. Amos, J. Leggo, S. Goodburn, S. Jain *et al.*, 1995 Microsatellite evolution—evidence for directionality and variation in rate between species. Nat. Genet. **10:** 337–343.

Schlötterer, C., and D. Tautz, 1992 Slippage synthesis of simple sequence DNA. Nucleic Acids Res. **20:** 211–215.

Shriver, M. D., L. Jin, R. Chakraborty and E. Boerwinkle, 1993 VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. Genetics **134:** 983–993.

Shriver, M. D., L. Jin, E. Boerwinkle, R. Deka, R. E. Ferrell *et al.*, 1995 A novel measure of genetic distance for highly polymorphic tandem repeat loci. Mol. Biol. Evol. **12:** 914–920.

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

Slatkin, M., and R. R. Hudson, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555–562.

Talbot, C. C., D. Avramopoulos, S. Gerken, A. Chakravarti, J. A. Armour *et al.*, 1995 The tetranucleotide repeat polymorphism D21S1245 demonstrates hypermutability in germline and somatic cells. Hum. Mol. Genet. **4:** 1193–1199.

Tautz, D., 1989 Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res. **17:** 6463–6471.

Valdes, A. M., M. Slatkin and N. B. Freimer, 1993 Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics **133:** 737–749.

Valsecchi, E., P. Palsbøll, P. Hale, D. Glockner-Ferrari, M. Ferrari *et al.*, 1997 Microsatellite genetic distances between oceanic populations of the humpback whale (*Megaptera novaeangliae*). Mol. Biol. Evol. **14:** 355–362.

Weber, J. L., 1990 Informativeness of human $(dC\text{-}dA)_n \cdot (dG\text{-}dT)_n$ polymorphisms. Genomics **7:** 524–530.

Weber, J. L., and P. E. May, 1989 Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am. J. Hum. Genet. **44:** 388–396.

Weber, J. L., and C. Wong, 1993 Mutation of human short tandem repeats. Hum. Mol. Genet. **2:** 1123–1128.

Weir, B. S., 1990 *Genetic Data Analysis. Methods for Discrete Population Genetic Data.* Sinauer Associates, Sunderland, MA.

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

Communicating editor: S. Yokoyama