Microsatellite Allele Frequencies in Humans and Chimpanzees, with Implications for Constraints on Allele Size

John Carlos Garza,* † Montgomery Slatkin,* and Nelson B. Freimer†

* Department of Integrative Biology, University of California, Berkeley; and †Neurogenetics Laboratory, Center for Neurobiology and Psychiatry, Department of Psychiatry, University of California, San Francisco

The distributions of allele sizes at eight simple-sequence repeat (SSR) or microsatellite loci in chimpanzees are found and compared with the distributions previously obtained from several human populations. At several loci, the differences in average allele size between chimpanzees and humans are sufficiently small that there might be a constraint on the evolution of average allele size. Furthermore, a model that allows for a bias in the mutation process shows that for some loci a weak bias can account for the observations. Several alleles at one of the loci (Mfd 59) were sequenced. Differences between alleles of different lengths were found to be more complex than previously assumed. An 8-base-pair deletion was present in the nonvariable region of the chimpanzee locus. This locus contains a previously unrecognized repeated region, which is imperfect in humans and perfect in chimpanzees. The apparently greater opportunity for mutation conferred by the two perfect repeat regions in chimpanzees is reflected in the higher variance in repeat number at Mfd 59 in chimpanzees than in humans. These data indicate that interspecific differences in allele length are not always attributable to simple changes in the number of repeats.

Introduction

Simple-sequence repeat (SSR) or microsatellite loci are abundant and relatively evenly spaced in all complex eukaryotic genomes examined to date. Estimates of their abundance range as high as one every 6 kb (Beckmann and Weber 1992). As a result of this and their extreme polymorphism, they have become important in many areas of biology. They are the primary markers being used in the large-scale mapping efforts now underway for many vertebrate genomes and have been successfully used to map a number of human disease-causing genes. Microsatellites are also used increasingly in population biology for assessment of parentage and as indicators of gene flow and hybridization (Bruford and Wayne 1993). The discovery that these repeated regions are conserved between species (Moore et al. 1991; Schlotterer et al. 1991; Stallings et al. 1991) has led to attempts to use them to reconstruct phylogenies, using distance measures calculated from allele frequency data and pairwise differences between individuals (Bowcock et al. 1994).

In spite of their increasing use in many areas of biology, the mutational process that generates polymorphism at microsatellite loci is poorly understood. Valdes et al. (1993) and Shriver et al. (1993) used computer simulations to show that allele frequency distributions of dinucleotide repeats in humans were consistent with the stepwise mutation model introduced by Ohta and Kimura (1973), in which mutations increase or decrease allele size by one repeat unit. Di Rienzo et al. (1994) examined data from a human population and showed that a two-phase model of mutation, in which most mutations change allele size by one repeat unit but the rest change allele size by several repeat units, better fit the observed distributions for 8 of 10 microsatellite loci studied.

These models assume no constraints on allele size, and as a consequence they predict that the difference in the average numbers of repeats in reproductively isolated populations tends to increase without bound at a rate proportional to their time of separation (Goldstein et al. 1995; Slatkin 1995). However, the question of limits on allele size has been raised recently because of a number of different observations, including sequence database searches (Stallings et al. 1991), a lack of divergence at the interspecific level (Bowcock et al. 1994), and the observation that most microsatellite repeat regions are relatively small (Beckmann and Weber 1992). Although a few large SSRs are found in telomeric regions (Wilkie and Higgs 1992), the vast majority of microsatellite alleles are <50 repeat units in length. The question of what controls allele length at microsatellite loci has be-
come important, in light of the discovery that departure from the normal range of allele lengths at certain trinucleotide loci is associated with several genetic diseases in humans (Nelson 1993).

In an earlier study, Di Rienzo et al. (1994) collected genotypic data from three distinct human populations at 10 microsatellite loci and found striking similarities in allele frequency distributions and population statistics. In most comparisons the mean number of repeats at a locus in the three populations differed by <1 repeat unit, and in no case did the means differ by >3 repeat units between any two populations (data not shown). Mutation rates at most microsatellite loci have been estimated to be $10^{-3}$–$10^{-1}$ per generation (Weber and Wong 1993). With the high mutation rates of microsatellites and the relative isolation of the populations examined by Di Rienzo et al. (1994), one would expect that mutation and genetic drift would have led to substantial differentiation between the three populations. The observed similarities could, however, potentially be explained by gene flow between the populations in historic times. Alternatively, the similarities could be due to evolutionary processes such as selection or biased mutation that constrain the divergence of average repeat number in different populations.

To distinguish between these two possibilities, gene flow and constraints on divergence, we examined 10 di-nucleotide repeat loci in chimpanzees (*Pan troglodytes*), a hominoid population that has been isolated from the lineage leading to *Homo sapiens* for ≥4 Myr (Horai et al. 1992). These are the same 10 loci studied previously by Di Rienzo et al. (1994). If gene flow is the only process homogenizing the average number of repeats in humans at these loci, then we would expect that the average repeat numbers would be quite different in the two species.

We will show that the differences in average repeat number at several loci in humans and chimpanzees are less than what would be expected if these loci were subject to a neutral mutational process. We show that this conclusion is valid even if unbiased gene conversion is acting to replace one copy of the locus by another copy. We also show that a weakly biased mutational process would be sufficient to explain the observed similarities, and we describe how much bias would be required to explain our observations.

The differences between alleles at microsatellite loci have commonly been assumed to be due entirely to changes in the number of repeat units (Tautz 1989; Weber and May 1989). The models used to explain the mutational process at SSRs in the studies described above have rested on this assumption, as have a number of distance measures that have recently been constructed (Goldstein et al. 1995; Slatkin 1995). It is particularly important that this assumption be met if microsatellites are to be used in interspecific studies. We have examined this assumption by sequencing alleles from one of the loci with the largest divergence in mean repeat number between the chimpanzee and human samples. Our preliminary results show that mutational dynamics at microsatellite loci can be more complicated than previously assumed and that differences between alleles, at least at the interspecific level, are not always due solely to changes in the number of repeat units in the primary repeat region.

**Material and Methods**

Blood samples from 13 unrelated chimpanzees and 33 chimpanzees with variable levels of relatedness were obtained from Yerkes Primate Center (Emory University, Atlanta), along with complete pedigrees for all individuals. Total genomic DNA was extracted according to standard methods. Densities of DNA dilutions used for the polymerase chain reaction (PCR) were standardized to 8 µg/µl by spectrophotometry, and 3 µl (24 µg) of these dilutions were used for each 10-µl PCR.

All individuals were genotyped at the 10 dinucleotide repeat loci selected by Di Rienzo et al. (1994), but only unrelated individuals and chromosomes were used in our analysis. These loci were originally chosen to provide a large range of variance in repeat number and mutation rate. Primers were purchased from Research Genetics (Huntsville, Alabama) and end-labeled using γ-32P. PCR was performed as described elsewhere (Di Rienzo et al. 1994). For Mfd 5 and Mfd 33 we used the published human sequences to design alternative primers sets. These primers were synthesized using an IBI oligonucleotide synthesizer. PCR products were assayed on 5% acrylamide gels and were visualized by autoradiography. All PCR and sequencing reactions were performed in a Perkin-Elmer 9600 PCR machine, without overlaying mineral oil. Genotype determinations were made by two people independently and checked for concordance.

Alleles of Mfd 59 were sequenced using the γ-32P cycle sequencing protocol provided by Promega in the fmol DNA sequencing kit, but with the following modifications: before sequencing, all templates were subjected to 30 cycles of PCR in a 40-µl volume with the following profile: initial denaturation at 94°C for 3 min; denaturation at 92°C for 30 s; annealing at 57°C for 30 s; and extension at 72°C for 30 s; and a final extension at 72°C for 3 min. Alleles were separated on a 3% agarose gel (Nusieve, low-melt) and purified using the Promega MagicPreps system. Two microliters of the eluted DNA were used as the template for 20-cycle sequencing reactions. Sequencing reactions were visualized by autoradiography on 8% acrylamide gels. The sequence of the
human reference allele was extracted from GenBank (accession number x54584). The primers used for amplification and sequencing were described by Weber et al. (1990).

All statistics were calculated using repeat numbers derived from allele size as determined from migration of PCR products on an acrylamide gel relative to a size standard and not by directly sequencing alleles. Calculations of allele frequencies in the chimpanzee sample used only unrelated individuals and chromosomes (identifiable by pedigree information) and for the human sample used the combined data from the three human populations described by Di Rienzo et al. (1994).

**Results**

Eight of the 10 microsatellite loci studied by Di Rienzo et al. (1994) could be amplified in the chimpanzee by using the original primers obtained from human sequences. For the two loci that did not amplify, Mfd 5 and Mfd 33 (on human chromosomes 19 and 22, respectively), we designed additional primers that included ~50 additional bases on each side of the repeat. No combination of the alternative and/or standard primers yielded amplified PCR products.

The mean and variance in repeat number, as well as other summary statistics for the combined human populations and the chimpanzee sample, are displayed in table 1. Although the shapes of the allele frequency distributions are not always the same (fig. 1), the average repeat number and the range in allele size are surprisingly similar in most cases. The average number of repeats in the different human populations is almost identical in most cases. Between species, half of the loci have mean repeat numbers that differ by <3 repeat units and the other half by an average of 10.5 repeat units (table 1).

We investigated the basis of variation between alleles at the locus Mfd 59. The sequence of a human allele which was 4 bp smaller in size (72 bp) than the reference allele was determined (fig. 2). It differed from the reference allele not by the simple absence of two repeats in the primary CA repeat region but by the absence of four CA repeats and the presence of two additional AT repeats at the boundary of the CA region in an AT-rich region not previously recognized as being variable.

We also determined the sequence of four alleles of Mfd 59 in the chimpanzee (fig. 2). The major difference between alleles at this locus in chimpanzees and humans is that the perfect AT repeat of the chimpanzee loci is an AT-rich region in humans that differs from perfection by three transversions. In contrast, only one transitional difference was found between species in the rest of this locus (data not shown). Another difference at this locus between the two species is that the region containing five CA repeats in the human is reduced to one CA repeat in the chimpanzee. This region appears to be invariant in both species. Within the chimpanzee, variation at this locus is attributable to differences in both the CA and the AT region with three different numbers of CA and AT repeats present in four alleles. Thus, for example,
the addition of six AT repeats and the deletion of three CA repeats (fig. 2). The change in the basis of variation between the two species, from a simple to a compound repeat structure, is accompanied by a four-fold increase in variance in repeat number in chimpanzees (see table 1).

**Theory**

We can predict the difference in the average repeat number of a microsatellite locus in two isolated species, under the assumption that the locus of interest is neutral. We begin with the case in which there is no constraint on repeat number and then discuss a model in which there is a constraint in repeat number caused by a bias in the mutation process. In our analysis we will consider the combined effects of mutation, genetic drift, and gene conversion in two species that have been reproductively isolated since time \( t \) in the past. We assume that, since isolation, one species has been of constant size \( N_1 \) and the other of constant size \( N_2 \). Before isolation, the ancestral species was of size \( N_0 \). The theory is a slight generalization of that described by Goldstein et al. (1995) and Slatkin (1995), in that it allows for different population sizes and unbiased gene conversion.

Assume that we are concerned with a particular microsatellite locus and let \( a_{ij} \) be the number of repeat units in the \( i \)th copy of this locus in species \( j \), where \( i = 1, \ldots, 2N_j \) and \( j = 1 \) or 2. We will be concerned with

**Fig. 1.**—Allele frequency distribution of two dinucleotide microsatellite loci in humans and chimpanzees. a, Mfd 3, the locus with the smallest difference in average allele size between the species. b, Mfd 75, the locus with the largest difference in average allele size between the two species.

the difference between chimp alleles 1 and 2, which differ in size by 6 bp, is not due to the addition of three repeat units in one of the repeat regions. Rather, it results from

**Fig. 2.**—Sequences of six alleles of Mfd 59. The indicated lengths include only the repeated region and not the flanking or primer regions. The total length of the amplified product for human allele 1 was 183 bp. The length of flanking sequence was variable in all alleles sequenced. The sequence of human allele 1 was extracted from GenBank. The sequences of human allele 2 and four chimp alleles were determined as described in the text. The sequence is presented 5' to 3'.
the evolution of the average repeat number in generation $t$, $\bar{a}_i(t)$, and the variance in repeat number, $\sigma^2_i(t)$. In our model, we will predict those values, but we cannot do so exactly, because they depend on the stochastic effects of mutation, drift, and gene conversion. Instead, we can predict the values expected in a large number of replicates of the same process. Let $E[.]$ denote the expected value across replicates. For our purposes, $E[(\bar{a}_1 - \bar{a}_2)^2]$ is of most interest because it will tell us how large a difference in average allele size we would expect to see in two isolated species.

We can express the mean and variance in the two populations as

$$\bar{a}_j = \frac{1}{2N_j} \sum_{i=1}^{2N_j} a_{ij}, \quad (1)$$

and

$$\sigma^2_j = \frac{1}{2N_j} \sum_{i=1}^{2N_j} (a_{ij} - \bar{a}_j)^2. \quad (2)$$

In the derivations, it is easier to use the average of the squared differences in repeat number in each species,

$$S_j = \frac{1}{2N_j(2N_j - 1)} \sum_{i \neq j'} (a_{ij} - a_{ij'})^2, \quad (3)$$

and the average squared difference in allele sizes in the two species

$$D = \frac{1}{4N_1N_2} \sum_{i=1}^{2N_1} \sum_{i'=1}^{2N_2} (a_{1i} - a_{1i'})^2. \quad (4)$$

Note that

$$S_j = \frac{4N_j}{2N_j - 1} \sigma^2_j, \quad (5)$$

and

$$D = \sigma^2_1 + \sigma^2_2 + (\bar{a}_1 - \bar{a}_2)^2. \quad (6)$$

Thus, to calculate $E[(\bar{a}_1 - \bar{a}_2)^2]$ we need the expectations of $S_j$ and $D$.

For the case in which there is no constraint on allele size we will assume that the mutation rate is $\mu$ and that the distribution of changes in allele size given that a mutation occurs is drawn from a distribution with mean 0 and variance $\sigma^2_m$. The "one-step" model used by Valdes et al. (1993) and Shriver et al. (1993) and the "two-phase" model introduced by Di Rienzo et al. (1994) are special cases of this more general model. If we consider two copies of the locus, the difference in repeat number is the sum of the changes under the mutation process that have occurred since those two copies had a most recent common ancestor, i.e., the coalescence time, $t$. Under our hypothesis about mutation, the average number of mutations separating two copies is $2\mu t$, and hence the expected value of the squared difference in allele size is $2\mu \sigma^2_m$. This argument works because the mutation process is reversible in a mathematical sense, which means that the increment under mutation is the same whether the process is regarded as going forward or backward in time. This is not true when the average change in allele size under mutation depends on the current allele size, as in the model of biased mutation we consider later.

To compute the expected values of $S_1$, $S_2$, and $D$, we need the expected values of the appropriate coalescence times. That is,

$$E[S_j] = 2\mu \bar{t}_j \sigma^2_m \quad (7)$$

and

$$E[D] = 2\mu \bar{t}_{12} \sigma^2_m, \quad (8)$$

where $\bar{t}_j$ is the expected coalescence time of two copies of the locus drawn at random from species $j$ and $\bar{t}_{12}$ is the expected coalescence time of one copy from species 1 and the other from species 2.

If we assume that $\tau$, the time since the separation of the species is much larger than both $2N_1$ and $2N_2$ (which is probably true for both humans and chimpanzees), then the coalescence of two copies of the locus in either species is almost certain to occur before $\tau$. We assume that gene conversion occurs with probability $g$ per generation, and, when gene conversion occurs, a randomly chosen copy of the locus is replaced by another randomly chosen copy. This is unbiased gene conversion that produces identity in state. For our purposes, it is equivalent to a coalescence of two copies. Because of genetic drift, a coalescent event occurs with probability $1/(2N_j)$ per generation (Hudson 1990). The additional effect of gene conversion is to increase that probability by $g/(2N_j)$, and hence the net probability of coalescence is $(1 + g)/(2N_j)$, and the expected time to coalescence of two copies in each species is

$$\bar{t}_j = \frac{2N_j}{1 + g}. \quad (9)$$

provided $g \ll 1$. As Nagylaki (1983) noted, gene conversion of this type slightly reduces the effective population size.

To find $\bar{t}_{12}$, we use the fact that coalescence cannot occur before $\tau$ in the past and before that time the two
copies are in a randomly mating population of size $N_0$

Hence

$$\tilde{t}_{12} = \tau + \frac{2N_0}{1 + g}.$$ (10)

Combining these results, we find

$$E[\sigma_j^2] = \frac{(2N_j - 1)\mu\sigma_m^2}{1 + g},$$ (11)

$$E[D] = \left[\tau + \frac{2N_0}{1 + g}\right]2\mu\sigma_m^2,$$ (12)

and hence

$$E[(\tilde{a}_1 - \tilde{a}_2)^2] \approx \mu \tau \sigma_m^2 [1 + O(N/\tau)],$$ (13)

where $O(N/\tau)$ indicates terms that are of the same order of magnitude as the population sizes divided by the time of separation of the two species. For human and chimpanzees those terms are quite small and can be ignored.

Thus equation (13) tells us that, for large separation times, the expected difference in the square of the average repeat numbers increases linearly with time at a rate proportional to the net increase in variance per generation by mutation. This conclusion is independent of whether gene conversion occurs. Goldstein et al. (1995) obtained the same result for the one-step model of mutation when all populations had the same size.

We can test the predictions of this model using (13). To be conservative, we assume a separation time, $\tau$, of $2 \times 10^3$ generations (i.e., 4 Myr with 20 years per generation). Under the assumption that $\tau \gg 2N$, the distribution of $\tilde{a}_1 - \tilde{a}_2$ is approximately normally distributed with mean 0 and variance given by (13), because the mean values are essentially undergoing a random walk. In that case, the probability that $\tilde{a}_1 - \tilde{a}_2$ differ in absolute value by less than an amount $\Delta$ is

$$\Pr(|\tilde{a}_1 - \tilde{a}_2| < \Delta) = \frac{1}{\sqrt{2\pi}\sigma^2} \int_{-\Delta}^{\Delta} \exp\left[\frac{-x^2}{2\sigma^2}\right]dx = \operatorname{erf}\left(\frac{\Delta}{\sqrt{2}\sigma}\right),$$

where $\sigma^2 = 2\mu\sigma_m^2$ and erf(.) is the error function. Although we do not know $\mu\sigma_m^2$, we can assume $\sigma_m^2 = 1$, which is again conservative, and consider results for different mutation rates. We can use the fact that $\operatorname{erf}(x) < 0.05$ if $x < 0.0224$, to say that values of $\Delta < 0.0444\sqrt{\mu}\tau$ or $\Delta < 19.85\sqrt{\mu}$ would be significantly too small (at the 5% level). Thus if $\mu = 10^{-3}$, the threshold value of $\Delta$ is 0.628, and if $\mu = 10^{-4}$, the threshold value of $\Delta$ is 0.199.

For $\mu = 10^{-3}$, the value of $D$ is significantly too small only for Mfd 3 ($\Delta = 0.306$). Several other probability values are relatively small but not significantly so: they are between 0.134 (for Mfd 38) to 0.597 (for Mfd 75). For $\mu = 10^{-4}$, none of the probabilities indicates significance, although the value for Mfd 3 is only 0.054, and several >0.9. With a larger divergence time, $\tau$, these probabilities would be smaller. Without knowing the mutation rates at these loci, we cannot be certain that the differences in average allele size are inconsistent with the model that assumes no constraints on allele size. If mutation rates are of the order of magnitude of $10^{-3}$, as suggested by the data of Weber and Wong (1993), then several of these loci have differences in allele size that are much smaller than expected, suggesting that there is some constraint on allele size. If the mutation rates for these loci are much smaller, then there is no reason to assume a constraint.

One possible mechanism for constraining allele size is that mutation is biased in such a way that small alleles tend to increase in size while large alleles tend to decrease in size. We model a bias in the mutation process by assuming that there is some "target" size and that the expected change under mutation depends on the difference from that target size. Because we are concerned here with differences in allele size, the target size is unimportant and can be set to 0 without loss of generality, thus simplifying the resulting equations. Thus we assume that, if an allele of size $a_i$ mutates, the expected size of the descendent allele is $-\beta a_i$, and the variance in the change in size $\sigma_i^2$. This model of biased mutation is not reversible in the sense used before. The parameter $\beta$ is the magnitude of the bias.

We again need the expected squared difference in size of two copies of a locus separated by a coalescence time $t$. First, consider one copy of the locus and assume that at some time $t$ in the past it was of size $a_0$. We can find the distribution of sizes by noting that our mutation model is now equivalent to an Ornstein-Uhlenbeck stochastic process (Karlin and Taylor 1981, p. 170), although in the present model, the state space is discrete, not continuous. The general theory for this kind of process tells us that the distribution of sizes of a single descendent copy is normally distributed with a mean $a_0e^{-\mu t}$ and variance $\mu \sigma_m^2(1 - e^{-2\mu t})/2\beta$. This result can also be obtained directly by writing the forward diffusion equation and solving it. If two copies are descended from a common ancestor at time $t$, they represent two independent realizations of this process. Hence their expected difference in size is 0 and the expected squared difference in size is

$$E[(a - a')^2] = \frac{2\mu \sigma_m^2(1 - \mu e^{2\mu t})}{2\beta},$$ (14)
where \( a \) and \( a' \) represent the sizes of two different copies.

We can use this result to obtain the average squared difference in size in each species. In a randomly mating population, the distribution of coalescence times has an exponential distribution with mean \( 2N_t \), so the expected value of \( S_t \) is obtained by averaging (14) over all coalescence times:

\[
E[S_t] = \frac{4N_t \sigma_{m_s}^2}{1 + 4N_t \beta},
\]

(15)

assuming, as before, that \( N_1 \) and \( N_2 \) are much less than \( \tau \). Thus the variance in each species is smaller than in the absence of the bias by a factor \( 1/(1 + 4N_t \beta) \).

We can find \( E[D] \) in the same way. The only difference is that the distribution of coalescence times is 0 for \( t \leq \tau \) and then an exponential for \( t > \tau \), which implies

\[
E[D] = \frac{m \sigma_{m_s}^2}{\beta} \left[ 1 - e^{-2b \tau} \frac{1}{1 + 4N_0 \beta} \right].
\]

(16)

Clearly, (16) reduces to (12) if \( \beta \) goes to 0 while \( \tau \) is held constant. We have ignored the possibility of gene conversion, but it could be incorporated by dividing all the population sizes by \( 1 + g \).

We can combine these results to obtain

\[
E[(\bar{a}_t - \bar{a}_s)^2] = E[D] - E[\sigma_{m_s}^2] - E[\sigma_{m_r}^2].
\]

(17)

There appears to be no useful simplification of this expression. But if all the population sizes are equal \((N_0 = N_1 = N_2 = N)\) and \( \beta \tau \gg 1 \), we find

\[
E[\bar{a}_t - \bar{a}_s] = \frac{\mu \sigma_{m_s}^2}{\beta (1 + 4N_0 \beta)} = \frac{E[\sigma_{m_s}^2]}{2N_0 \beta}.
\]

(17)

where \( E[\sigma_{m_s}^2] \) is the variance in each of the populations at equilibrium (which are equal because of the assumption that both population sizes are equal). In the general case, \( E[(\bar{a}_t - \bar{a}_s)^2] \) approaches a constant value, which is a weighted average of the equilibrium variances of allele sizes, as \( \tau \) increases.

For this mutation model to account for the observations presented above, the value of \( 2N_0 \beta \) would have to be relatively large. For example, the largest difference in average allele size is 11.8 repeat units for locus Mfd 75. If we assume that \( E[(\bar{a}_t - \bar{a}_s)^2] \) is the square of this value, then (17) suggests that, for the parameter values described above, the smallest bias that would explain this observation would be \( 2N_0 \beta = 0.033 \). We obtained this value of \( \beta \) by dividing 4.65, which is an estimate of the variances in repeat number, by \((11.83)^2\). In this case, the value of \( \beta \) is substantially less than \( \mu \), the mutation rate, indicating that a bias of this magnitude has little effect. A larger value of the variance within species would give a correspondingly larger value of \( 2N_0 \beta \). Other loci, with smaller differences in the average allele sizes would require larger biases. A very small difference in average allele size, such as is found in Mfd 3 would require a value of \( 2N_0 \beta \) on the order of 66 (setting \( \bar{a}_1 - \bar{a}_2 = 0.3 \) and \( \sigma_{m_s}^2 = 6 \), which implies that the value of \( \beta \) is much larger than the mutation rate.

**Discussion**

Our results indicate that constraints exist on repeat number at microsatellite loci. Tachida and Izuka (1992) used a linear model of birth and death processes to examine persistence times of selectively neutral repetitive elements and found that the observations for one out of six elements were inconsistent with the results of their model. Deka et al. (1994) found that the genetic distance between humans and chimpanzees, calculated using microsatellites, was much smaller than expected, indicating a constraint on allele size. The observation that average repeat number at human loci remains relatively low in spite of an apparent upward bias in directly observed mutational events (Weber and Wong 1993) confirms that some sort of regulation must be taking place. If it was not, average repeat number would increase or decrease indefinitely.

One can imagine three sorts of mechanisms for regulating repeat number, gene conversion, biased mutation, or selection on the loci themselves. Gene conversion has recently been shown to be an important mechanism for generating new alleles at minisatellite loci (Jeffreys et al. 1994), but its importance in microsatellite evolution is not yet known. In any case, our results suggest that unbiased gene conversion of the type modeled here does not help to explain the discrepancy between our observations and the predictions for a neutral locus. The reason is that gene conversion simply replaces one copy of the locus by another but it does not retard the overall drift of the average repeat number.

As shown by our theory, a biased mutation process could explain the results. However, the only bias in mutation so far observed has been a predominance of upward mutations. The strength of the bias necessarily depends on how similar average repeat numbers are. For loci at which average repeat numbers do not differ substantially despite several million years of separation between the two species, our theory suggests that the bias has to be substantial. We have modeled a symmetric process in which the same type of bias affects large and small alleles, but it may be that instead there are occasional large downward mutations that because of their rarity are not easily observed. Strand slippage by DNA polymerase is thought to be the major mechanism of microsatellite evolution (Levinson and Gutman 1987),
but larger mutations may be due to a different mutational mechanism, such as unequal crossing-over.

Selection may also be an important factor regulating repeat number. Although a general function for microsatellite loci has not been described, certain functions have been observed for some microsatellites and could be the basis of selection on particular loci. Hamada et al. (1984) have shown that dinucleotide repeats can affect regulation of transcription of at least one gene in mammalian cells. This activity was independent of location and is consistent with the observation that microsatellites seem to be evenly spread throughout the genome and not concentrated in traditional areas of transcription regulators (Stallings et al. 1991). How widespread a phenomenon this is, however, is not yet known. However, Mfd 3, the only locus which lies within a gene and which could be amplified in both species, also has the smallest difference in average allele size, an observation which is consistent with such an explanation. Mfd 5 also lies within a gene but failed to amplify in chimpanzees.

It has also been suggested that microsatellites play a role in recombination (Pardue et al. 1987). It is easy to imagine how conserved regions of repetitive sequence could facilitate chromosomal alignment during meiosis, but the evidence is contradictory. While certain microsatellites are indeed less abundant in regions of the *Drosophila* genome that do not undergo recombination, they are in general more abundant in the mouse genome than in the human genome, in spite of the fact that there is less recombination in the mouse genome (Stallings et al. 1991).

A possibility for a more general selective constraint on these loci, but one that is not yet well understood, is that perhaps microsatellites play some role in regulating large-scale chromosomal structure. CA repeats have been shown to have Z-DNA forming potential (Hamada et al. 1982; Nordheim and Rich 1983). The function of such DNA has not yet been completely elucidated, but it may facilitate packaging during chromosomal condensation in meiosis. This speculation is supported by the fact that microsatellites are, in general, absent from bacteria that do not undergo such chromosomal packaging (Gross and Garrard 1986).

Microsatellite loci have previously been shown to be conserved in both position and sequence between closely related mammalian species. Our finding that 8/10 microsatellite loci are conserved between chimps and humans is consistent with the results of Bowcock et al. (1994), who found that many human microsatellite loci are also present in the great apes. Stallings et al. (1991), found 100% homology in the positions of the two GT loci for which informative flanking sequence was available in both humans and chimpanzees, indicating that for the most part microsatellite loci are conserved in closely related species. The two loci that did not amplify are on different chromosomes, and it is possible that these loci are no longer present in the chimpanzee. Attempts to design primers that encompassed a larger region of the locus yielded no amplified product, thus suggesting that the regions have been deleted in the chimpanzee genome.

The fact that some GT repeats are present and conserved in position in species as divergent as human and rat also indicates that some of these loci are ancient in origin (Stallings et al. 1991). This homology has raised the possibility that microsatellites might be used in studies of phylogeny. While it may be true that within a species they are useful for tree building (Bowcock et al. 1994), our results suggest that they should not be used for interspecific phylogenetic reconstruction. The interspecific differences in the structure of the repeat locus we have examined indicates that, in many cases, alleles of the same size in different species may not be identical by descent and instead may differ by an unknown number of mutational events. The existence of a constrained range of allele size coupled with the high mutation rates of microsatellite loci underscores this problem. The discovery in the chimpanzee of the deletion of 4 CA repeat units distal to the AT repeat region of Mfd 59 reduces the mean difference in number of repeats in the variable repeat region and emphasizes the similarity in allele frequency distributions. However, it also indicates that a simplistic model of the evolution of microsatellites may be inadequate at the interspecific level.

Although the number of sequences examined in our study is small, it is clear that the mutational dynamics of a microsatellite locus can differ between species. This raises the possibility that alleles of the same size in different species may be non-identical by descent and also differ significantly in sequence-level structure. Other loci need to be examined to see how widespread a phenomenon this is. Although a definite example of such non-homology has not yet been shown, the presumably independent mutational processes of the different repeat regions in compound microsatellite loci should allow this to be done quite easily. Thus compound microsatellites should be avoided in studies where homology is crucial.

The sequence data presented here indicate that imperfections in the repeated region may cause a decrease in the mutation rate of microsatellite loci. The locus Mfd 59 in chimpanzees contains a perfect AT repeat region whereas in humans the same region contains three transversional imperfections. This change in structure is accompanied by an almost fivefold increase in variance in repeat number (34.4 in chimps vs. 7.3 in humans).
This is consistent with the work of Chung et al. (1993), who found that the expanded microsatellite associated with spinocerebellar ataxia type 1 had a perfect configuration in 30 affected chromosomes, whereas imperfections were present in 98% of the unaffected chromosomes. It may be that imperfections in repeated regions are a mechanism for constraining the mutation process, and thus allele length, at some microsatellite loci.

The fact that primers designed for humans amplify polymorphic loci in chimpanzees as well and that the polymorphism is inherited in a Mendelian fashion (data not shown) confirms the possibility that microsatellites might be used for the genetic management of captive chimpanzee colonies and zoo populations, as has been suggested by Morin et al. (1993). Determination of paternity and unequal genetic contribution by certain males can be a problem in such colonies. The microsatellite loci described here are sufficiently polymorphic to be used as multilocus DNA fingerprints capable of distinguishing all members of a colony. In addition, these loci can be amplified and characterized from a single plucked or shed hair (Morin et al. 1993), which should further facilitate their use in the management of chimpanzee colonies.

Conclusion

We have shown that average allele sizes at several microsatellite loci in chimpanzees and humans are sufficiently similar that there may be some constraint on the evolution of average allele size. We show that a weakly biased mutation process is sufficient to account for the observed differences in allele size and demonstrate how to estimate the extent of bias. Furthermore, the sequences of several copies of one of the loci show the potential complexity of the mutation process when considered at the interspecific level. Together, these results imply that allele sizes of microsatellite loci are unlikely to provide a way to infer phylogenetic relationships. Instead, differences among species can illustrate the kinds of rare mutational events that as yet do not appear to be important when examining distributions of allele sizes within species.

Acknowledgments

We thank Amy Peterson for technical help. This research was supported in part by grants from the National Institutes of Health to M.S. and to N.B.F. J.C.G. was supported by a Ford Foundation Predoctoral Fellowship.

LITERATURE CITED


Naoyuki Takahata, reviewing editor

Received September 21, 1994

Accepted January 24, 1995