Microsatellite lengths change over evolutionary time through a process of replication slippage. A recently proposed model of this process holds that the expansionary tendencies of slippage mutation are balanced by point mutations breaking longer microsatellites into smaller units and that this process gives rise to the observed frequency distributions of uninterrupted microsatellite lengths. We refer to this as the slippage/point-mutation theory. Here we derive the theory’s predictions for interrupted microsatellites comprising regions of perfect repeats, labeled segments, separated by dinucleotide interruptions containing point mutations. These predictions are tested by reference to the frequency distributions of segments of AC microsatellite in the human genome, and several predictions are shown not to be supported by the data, as follows. The estimated slippage rates are relatively low for the first four repeats, and then rise initially linearly with length, in accordance with previous work. However, contrary to expectation and the experimental evidence, the inferred slippage rates decline in segments above 10 repeats. Point mutation rates are also found to be higher within microsatellites than elsewhere. The theory provides an excellent fit to the frequency distribution of peripheral segment lengths but fails to explain why internal segments are shorter. Furthermore, there are fewer microsatellites with many segments than predicted. The frequencies of interrupted microsatellites decline geometrically with microsatellite size measured in number of segments, so that for each additional segment, the number of microsatellites is 33.6% less. Overall we conclude that the detailed structure of interrupted microsatellites cannot be reconciled with the existing slippage/point-mutation theory of microsatellite evolution, and we suggest that microsatellites are stabilized by processes acting on interior rather than on peripheral segments.

Introduction

Microsatellite DNA sequences consist of repeats of short motifs 1 to 6 bp sequences long. The number of such repeats in a microsatellite is referred to as the length of the microsatellite, and lengths change over evolutionary time, mainly through a process of replication slippage, although gene conversion through recombination may also play a minor role (Ellegren 2000). Slippage mutations usually consist of an increase or decrease of one repeat, although larger steps are also known, and the slippage mutation rate is quite high, typically $10^{-4}$ to $10^{-3}$ per haplotype per generation in mammals. Early models of the slippage mutation process did not adequately account for the observed distributions of microsatellite lengths, since they predicted indefinite expansions would sometimes occur, whereas in practice, observed lengths only very rarely exceed a few tens of repeats (Tautz 1993). To account for this discrepancy, two principal theories have been advanced. The first supposes that longer microsatellites experience more contractions than expansions, so that there is a length-dependent mutation bias (Xu et al. 2000).

The second theory—not mutually exclusive with the first—notes that microsatellites are as vulnerable to point mutations as the rest of the genome and suggests that the frequency distributions of microsatellite lengths represent a balance between the expansionary tendencies of slippage mutation and the contractions caused by point mutations breaking longer microsatellites into smaller units (Bell and Jurka 1997; Kruglyak et al. 1998). We refer to this as the slippage/point-mutation theory. Existing treatments of the slippage/point-mutation theory have calculated and evaluated predictions for peripheral segments of microsatellites, segment being used here to indicate a sequence of perfect repeats bounded at both ends by sequences that are not repeats. Here we extend the analysis to the case of interrupted AC microsatellites, consisting of segments separated by dinucleotides that are not AC (see Methods). Predictions as to how the variance of length changes with evolutionary time have been derived by Calabrese, Durrett, and Aquadro (2001) but do not overlap with the predictions derived and tested here.

To calculate the predictions of the slippage/point-mutation theory, we follow earlier treatments in four important respects. Firstly, we restrict attention to models in which microsatellite lengths change by slippage mutations by one repeat, which seems a reasonable simplification given the evidence stated above. Secondly, we suppose that the expansion rate at any length is the same as the contraction rate at that length. Without this assumption, the theory would not be clearly differentiated from the mutation bias theory described above. Thirdly, we assume that current distributions of microsatellite lengths represent an equilibrium between the expansionary tendencies of slippage mutation and the splitting effects of point mutations breaking microsatellites into smaller units. It seems reasonable to assume such evolutionary processes are at equilibrium given the long periods for which some microsatellite loci are known to have existed. Lastly, we assume that the point mutation rate does not vary within the genome, despite some evidence to the contrary (Wolfe, Sharp, and Li 1989; Santibanez-Koref, Gangeswaran, and Hancock 2002). For peripheral segments of microsatellites, it is possible to use the slippage/point-mutation theory to calculate analytically the equilibrium distributions of slippage models that specify the relationships...
between segment length, slippage rate, and the point mutation rate (Kruglyak et al. 1998; 2000; Sibly, Whittaker, and Talbot 2001). These equilibrium distributions can then be compared with observed distributions of lengths obtained from a genome search, and the best parameter values of the slippage models can be found using maximum-likelihood (Sibly, Whittaker, and Talbot 2001). These methods have also been used to compare nested linear models of the relationship between slippage rate and length. The results suggest that slippage rates increase with length for dinucleotide microsatellites in humans, mice, and fruit flies and that no or very little slippage occurs in very short segments comprising one to four repeats (Sibly, Whittaker, and Talbot 2001).

Here we derive and test the predictions of the slippage/point-mutation theory for the complete structure of interrupted microsatellites. No assumptions are made as to the form of the relationship between slippage rate and microsatellite length. Predictions are made using a combination of analytical methods and computer simulations. Predictions are tested using data from human AC microsatellites obtained from the human genome, taking one allele per locus.

Methods

Human Genome Data

A tabulation program, available at http://www.rubic.rdg.ac.uk/meade/ms, was used to record details of all AC dinucleotide microsatellites in the human genome (http://genome.cse.ucsc.edu, downloaded 27/3/01) that satisfied the following criteria: (1) The 10-base 5′ flank did not contain an AC or a CA dinucleotide. This criterion restricts attention to microsatellites for which equilibrium distributions can be calculated analytically (Sibly, Whittaker, and Talbot 2001). (2) The microsatellite included at least one sequence of five or more uninterrupted repeats. (3) The microsatellite was allowed to contain dinucleotide interruptions, as in ACACNNACACACACAC, where NN counts as an interruption if NN is not AC. Note that AC sequences following mononucleotide interruptions were not recorded since they cannot arise under the version of the slippage/point-mutation theory for the complete structure of interrupted microsatellites. No assumptions are made as to the form of the relationship between microsatellite length and slippage rate. The version of the slippage/point-mutation theory (Sibly, Whittaker, and Talbot 2001): Let

$$p_i s_i = p_{i+1}s_{i+1} + 2ai \sum_{j=i+1}^n p_j$$

for all \(i > 1\)  

$$p_i c = p_{i+1}s_{i+1} + 2a \sum_{j=2}^i p_j$$

In practice, following Sibly, Whittaker, and Talbot (2001), we estimated the ratios \(s/a\) and \(c\) was not estimated because the fitting procedure was conditioned on the absence of microsatellites smaller than five repeats in the data set. Under the equilibrium assumption the observed frequencies of microsatellite lengths have a multinomial distribution with parameters given by the above equilibrium distribution and the sample size \(n\) so the likelihood of the data can be written in terms of \(n\) and the \(p_i\) of the equilibrium distribution. Since the \(p_i\) are functions

Models

The version of the slippage/point-mutation theory implemented here follows earlier treatments in presuming that (1) slippage mutations cause an increase or decrease of one repeat; (2) the expansion and contraction rates are identical at any length; (3) the frequency distributions of microsatellite segment lengths in the genome are in equilibrium; and (4) the point mutation rate is invariant throughout the genome. The underlying framework is a discrete time Markov chain, the states of the chain being the positive integers, 1, 2, 3 . . . , each of which corresponds to the number of repeats at a microsatellite locus. Under the terms of the model, for each segment each generation three types of transition may occur:

1. The number of repeats may change by \(\pm 1\) unit because of slippage. For \(i \geq 2\) let \(s_i\) be the slippage mutation rate from state \(i\), defined here as the per generation probability that a segment of length \(i\) mutates by slippage to length \(i + 1\); \(s_i\) is also the probability of a slippage mutation to length \(i - 1\).
2. Point mutation may occur at any point within a segment, breaking it into two smaller segments. Point mutation causes a microsatellite of \(i\) repeat units to move to any of the states 1, 2 . . . \(i - 1\) at rate approximately equal to \(2a\), where \(a\) is the point mutation rate.
3. A transition from \(i = 1\) to \(i = 2\) due to specific base substitutions occurs at rate \(c\). This assumption is necessary to prevent \(i = 1\) being an absorbing state.

Analytical Calculation of the Equilibrium Distribution of 5′ Peripheral Segments

For peripheral segments, the equilibrium frequency distribution of a given slippage model can be calculated analytically (Sibly, Whittaker, and Talbot 2001). Here we show how the methods previously used to analyze straight-line models are readily extended to larger models that place few restrictions on the form of the relationship between microsatellite length and slippage rate. Note that when point mutation breaks a 5′ segment into two, the one at the 5′ side of the point mutation becomes the new 5′ segment. Letting \(p_i\) be the probability that a randomly selected 5′ segment is of length \(i\), \(i \geq 1\), the equilibrium values of \(p_i\) satisfy the following equations (Sibly, Whittaker, and Talbot 2001):

$$\sum_{i=1}^n p_i = 1$$

$$p_i s_i = p_{i+1}s_{i+1} + 2ai \sum_{j=i+1}^n p_j \text{ for all } i > 1$$

$$p_i c = p_{i+1}s_{i+1} + 2a \sum_{j=2}^i p_j$$
of the model parameters, this gives the likelihood as a function of the model parameters. Maximum likelihood methods are then used to estimate parameter values and their standard errors as previously described (Sibly, Whittaker, and Talbot 2001).

Computer Simulations

Since the available analytical methods only apply to first segments (Kruglyak et al. 1998, 2000; Sibly, Whittaker, and Talbot 2001), we employed computer simulations to find the equilibrium distributions of second and later segments. For this purpose we needed a sample of independently evolving microsatellites, which we obtained by following the evolution of a single microsatellite, and taking samples sufficiently sparsely that they were not autocorrelated. To see that it is sufficient to model the evolution of a single microsatellite, note that all current microsatellites at any locus are descended from a single ancestor, the most recent common ancestor (MRCA), and that sampling a single microsatellite from the population is equivalent to choosing at random from the possible descendants of the MRCA. The simulation model was run for $2.5 \times 10^{11}$ generations using the parameters, illustrated in figure 1 (right), that were derived from the frequency distribution of 5'-edge segments. Microsatellite characteristics were recorded only every 107 generations, to remove autocorrelation from the data set. Thus, although the simulation followed the evolution of a single microsatellite, by taking samples sufficiently sparsely that they were not autocorrelated, we obtained a sample equivalent to what would be obtained from independently evolving microsatellites.

Results and Discussion

The left panel of figure 1 shows the log frequency distribution of the lengths of peripheral 5'-edge segments of AC microsatellites in the human genome. Frequency declines with length until 10 repeats and is then approximately uniform with about $10^3$ microsatellites per genome for each length between 10 and 20 repeats, with a decline thereafter. The result of fitting a 17-parameter slippage/point-mutation model to these data using the analytical method is shown in the right panel of figure 1. Each parameter specifies a length-specific slippage rate, $s_i$, in multiples of the point mutation rate. One parameter was assigned to each microsatellite length between five and 20 repeats inclusive, and thereafter linear decline was assumed until a final plateau is reached at a length of 25 repeats (see fig. 1, right). Sixteen parameters were needed to specify the slippage rates at lengths between five and 20 repeats, and the final parameter specified the slippage rate at and above a length of 25 repeats. Parameter assignment was curtailed at 17 parameters because of the declining availability of data above 20 repeats.

The estimated slippage rates shown in figure 1 (right) are relatively low for the first four repeats, and then rise initially roughly linearly with length, as found previously (Rose and Falush 1998; Sibly, Whittaker, and Talbot 2001). However in contrast to the theoretical expectation that slippage rate increases linearly with length (Kruglyak et al. 1998), the more complex model used here suggests that slippage rate peaks at a length of 10 repeats and then declines, roughly linearly, until 20 repeats. The decline in slippage rates after 10 repeats is necessary, in the slippage/point-mutation model, to maintain the frequency-length distribution of figure 1 (left) near to horizontal between 10 and 20 repeats.

The frequency length distributions of the first five segments of interrupted AC microsatellites in the human genome, read in the 5' to 3' direction, are shown in figure 2. The distributions show some similarities, but the later segments are shorter. The equilibrium distributions of segments other than the first cannot be calculated analytically, and we employed simulation to discover the implications for later segments of the slippage model shown in figure 1 (right). The frequency distributions so obtained approximate equilibrium distributions, and are shown in the bottom row of figure 2. The frequency distributions of the different segments of the simulated distributions are very similar to each other, and there is none of the shortening of later segments seen in the genome. Furthermore, there are fewer later segments in the genome than in the simulation.

The frequencies of interrupted microsatellites that consist of 1, 2, 3 or more segments are shown in figure 3.
Frequencies decline geometrically with microsatellite size measured in number of segments, so that for each additional segment, the number of microsatellites is 33.6% less. The fit to a strict geometric decline is extremely good. The simulated microsatellites do not show this strict geometric decline, and the rate of decline is slower (fig. 3).

So far, segments have been counted simply according to their position from the 5' end. However, AC microsatellites are produced not only by polymerase replicating the sequence recorded in the database but also by polymerase traveling in the other direction on the complementary strand. This suggests treating all peripheral segments as equivalent, regardless of whether they occur at the 5' or 3' end, and classifying the interior segments according to their distance in segments from the periphery. When this is done (figs. 4 and 5) it appears, intriguingly, that peripheral segments are consistently the longest and that segment lengths decrease towards the center of the array.

General Discussion

Our results show that the detailed structure of these AC interrupted microsatellites are incompatible with the existing slippage/point-mutation theory of microsatellite evolution. The estimated slippage rates are relatively low for the first four repeats, and then rise initially roughly linearly with length (fig. 1, right), in accordance with previous work. Contrary to expectation and the experimental evidence, however, the inferred slippage rates decline in segment lengths above 10 repeats (Kruglyak et al. 1998; Brinkmann et al. 1998; Brohede et al. 2002; Huang et al. 2002). A further problem is that the slippage rates appear to be too low. The slippage rates shown in figure 1 (right) are given in multiples of the point mutation rate, and the point mutation rate in humans is about $10^{-9}$ per nucleotide per generation (Crow 1993). It can be deduced, therefore, that the slippage rates shown in figure 1 (right) appear too low by at least two orders of magnitude, since slippage at peak rates (fig. 1, right) would be around $2 \times 10^{-6}$ per meiosis per generation, whereas pedigree estimates suggest values around $5 \times 10^{-4}$ per meiosis per generation (Ellegren 2000). This discrepancy would however disappear if point mutation rates were higher within microsatellites than in other DNA sequences. Interestingly, it has recently been suggested that duplication of interruptions might occur through slippage (Harr, Zangerl, and Schlotterer 2000; Rolfsmeier and Lahue 2000). Such duplication of interruptions would give a similar effect to increased point mutation rate in microsatellites that contained at least one interruption, and so might explain why the ratio of slippage to point mutation rate appears so low in figure 1 (right).

The slippage/point-mutation theory also fails to adequately explain the frequency distributions of the lengths of the various segments (figs. 2 and 5). The theory predicts the various segments will have similar distributions, but in the genome, the later segments are shorter, counting in the 5' to 3' direction (fig. 2), and counting
inwards from the periphery, interior segments are shorter than those on the periphery (figs. 4 and 5). The fact that segment lengths decrease towards the center of interrupted microsatellites (fig. 4) suggests that microsatellites become wholly or partly stabilized by processes acting differentially on interior and peripheral segments. This suggests that stabilizing processes act more strongly in the interior than on the peripheral segments of microsatellites. There are several plausible mechanisms by which this feature might be achieved. Interruptions may stabilize the microsatellite (Petes, Greenwell, and Dominska 1997) and perhaps block expansions (Rolfsmeier, Dixon, and Lahue 2000; Rolfsmeier and Lahue 2000) or lead to segment shortening (Taylor, Durkin, and Breden 1999).

Stabilization of the interior segments of microsatellites would also explain the otherwise puzzling geometric decline in the frequency of microsatellites with number of interruptions shown in figure 3. The geometric decline suggests that there is a constant chance of occurrence of one further duplication, and from figure 3 this chance is 0.337. Thus, the chance of obtaining one duplication is 0.337, of obtaining two duplications is 0.337^2, of obtaining three duplications is 0.337^3, and so on. At first sight, it appears odd that the chance of obtaining one further duplication does not increase with the number of interruptions already in the microsatellite. If for instance the microsatellite contains five interruptions, one would have thought it five times more prone to slippage mutation than a microsatellite with one interruption. This can be explained, however, if internal interruptions are stabilized as discussed above, so that only peripheral interruptions are prone to duplication.

We conclude that the detailed structure of interrupted microsatellites is incompatible with the slippage/point-mutation theory in the simple form in which it has been tested here. The inferred pattern of mutation rates (fig. 1, right) is contrary to expectation, the ratio of slippage to point mutation rates is two orders of magnitude less than it should be, there is unexplained variation between the frequency distributions and sizes of the various segments (figs. 2, 4, and 5), and the frequency distribution of interruptions falls off faster than predicted (fig. 3). Reconciliation of some of these results with the theory might be possible by invoking slippage to duplicate/remove interruptions within microsatellites as described above, but it is uncertain whether this would have the desired quantitative effects.

In deriving predictions from the slippage/point-mutation theory it was assumed that the point mutation rate is invariant within a microsatellite locus. Could the slippage/point-mutation theory be rescued by relaxing this assumption while constraining the slippage mutation rate to increase with segment length in a more plausible manner? Variation of the point mutation rate with segment length might arise as a result of the action of slippage duplicating or stabilizing interruptions as described above. It can be incorporated into the models by making the parameter a a function of segment length j, so that equation (2) becomes

\[ p_i s_i = p_{i+1} s_{i+1} + 2i \sum_{j=i+1}^{\infty} a_j p_j \quad \text{for all } i > 1 \]  

with corresponding changes to equation (3). Note that the \( \Sigma \) term on RHS is necessarily +ve, so that if \( p_i = p_{i+1} \) for \( i \) between 10 and 20 as in figure 1 (left), then over this range of \( i \), \( s_{i+1} \) is necessarily less than \( s_i \). In other words, a decline in slippage rate with segment length is needed to produce the uniform frequencies of microsatellites of length between 10 and 20 repeats shown in figure 1 (left), regardless of how point mutation rate may change with segment length. We conclude that the slippage/point-mutation theory cannot be rescued simply by allowing the point mutation rate to vary with segment length.

Derivation of predictions also relied on the assumption that the point mutation rate is invariant throughout the genome. There is, however, some evidence to the contrary (Wolfe, Sharp, and Li 1989; Santibanez-Koref, Gangeswaran, and Hancock 2002). Recently, indeed, it has been pointed out that the slippage/point-mutation theory predicts a negative correlation between microsatellite length and the local point mutation rate, and a negative correlation has been reported between microsatellite lengths and the substitution rate in their flanking sequences, comparing orthologous loci in mouse and rat (Wolfe, Sharp, and Li 1989; Santibanez-Koref, Gangeswaran, and Hancock 2002). It was suggested on this basis that length differences are to some extent caused by differences in the local point mutation rate. It is therefore worth considering how the analysis that produced figure 1 (right) would be affected by local variation in point

**FIG. 4.—**Mean length (± SE) of segments in microsatellites consisting of (a) three, (b) four, and (c) five segments. Segments are numbered in the 5' to 3' direction. Length differences were compared using the Kruskal-Wallis test adjusted for ties: (a) \( H = 47.85, df = 2, P = 0.000 \); (b) \( H = 5.58, df = 3, P = 0.134 \); (c) \( H = 12.98, df = 4, P = 0.011 \).
mutation rate. The easiest way to think about this is to suppose that loci are classified by their local point mutation rates, the frequency of each such class being known. Each class then corresponds to a particular value of \( a \) in equations (2) and (3). The equilibrium distribution for each value of \( a \) can be calculated as before, and the frequencies of each microsatellite length can then be obtained by summing the class frequencies. Realization of this analysis in practice will have to await information as to the distribution of local point mutation rates.

Given that the results presented here are incompatible with the existing slippage/point-mutation theory, could they be explained by the alternative theory mentioned in the Introduction? According to this theory, longer microsatellites experience more contractions than expansions, so that there is there a length-dependant mutation bias (Xu et al. 2000). This “mutation-bias” theory can certainly produce frequency distributions like that shown in figure 1a, but when we have attempted to incorporate its effects, the parameter estimates failed to converge. Thus, we have not so far been able to combine the slippage/point-mutation and mutation-bias theories into a single theory with estimable parameters. Note also that such a theory will still have difficulties accounting for the observed differences between peripheral and interior segments (figs. 4 and 5), so further factors will still be needed to explain the stabilization of interior segments.

In conclusion, we have shown that the results presented here are incompatible with the existing slippage/point-mutation theory, and, in addition, important information has been gained about the detailed structure of interrupted microsatellites. We now know that segment lengths decrease towards the center of interrupted microsatellites (fig. 4), and that microsatellite frequency declines geometrically with microsatellite size measured in number of segments (fig. 3). An intriguing implication of these findings is that peripheral segments essentially behave as two isolated and perfect microsatellites, whereas neighboring segments heavily influence the growth of internal segments. The mechanistic explanations of these results probably involve stabilizing effects of interruptions, perhaps blocking expansions or enabling segment shortening, together with removal of interruptions through slippage.

Acknowledgments

D. Tautz and M. Beaumont and two referees provided valuable comments on the manuscript. A.M. and N.B. were supported by BBSRC studentships.

Literature Cited


Diethard Tantz, Associate Editor

Accepted November 26, 2002