The Relationship Between Microsatellite Slippage Mutation Rate and the Number of Repeat Units

*Yinglei Lai**[1] *and Fengzhu Sun*†

*Department of Mathematics and †Program in Molecular and Computational Biology, Department of Biological Sciences, University of Southern California

Microsatellite markers are widely used for genetic studies, but the relationship between microsatellite slippage mutation rate and the number of repeat units remains unclear. In this study, microsatellite distributions in the human genome are collected from public sequence databases. We observe that there is a threshold size for slippage mutations. We consider a model of microsatellite mutation consisting of point mutations and single stepwise slippage mutations. From two sets of equations based on two stochastic processes and equilibrium assumptions, we estimate microsatellite slippage mutation rates without assuming any relationship between microsatellite slippage mutation rate and the number of repeat units. We use the least squares method with constraints to estimate expansion and contraction mutation rates. The estimated slippage mutation rate increases exponentially as the number of repeat units increases. When slippage mutations happen, expansion occurs more frequently for short microsatellites and contraction occurs more frequently for long microsatellites. Our results agree with the length-dependent mutation pattern observed from experimental data, and they explain the scarcity of long microsatellites.

Introduction

Microsatellites are tandem repeats of DNA units. There are usually 1–6 bp (base pairs) for a repeat unit (motif). Microsatellites are highly abundant in eukaryotic genomes and can be genotyped using polymerase chain reaction (Weber and May 1989; Weber and Wong 1993). Microsatellites are highly polymorphic and are widely used as genetic markers in studies of disease mutations (Ashley and Warren 1995; Leeflang et al. 1999), tumor and cancer research (Sturzeneker et al. 2000), genetic mapping (Kong et al. 2002), population genetics (Rosenberg et al. 2002), and linkage and association studies (Ott 1999). Microsatellites are subject to mutations durng evolution. Besides point mutations, polymerase template slippage mutations may occur to change the number of repeat units in a microsatellite locus (Schlötterer and Tautz 1992; Viguera, Canceill, and Erlich 2001). An important problem is to understand microsatellites slippage mutation mechanisms.

Based on phylogenetic analysis, Messier, Li, and Stewart (1996) suggested a minimum number of repeat units for slippage mutations. Using a simple mathematical model, Rose and Falush (1998) demonstrated the existence of a minimum threshold size for slippage mutations by studying the ratio between the observed frequency and the expected frequency of microsatellites. The estimated threshold size was about eight nucleotides long irrespective of different motifs for mononucleotides, dinucleotides, and tetranucleotides. The study suggested more complicated mechanisms for microsatellite slippage mutations (Rose and Falush 1998). However, Pupko and Graur (1999) debated the existence of threshold sizes for slippage mutations.

In experimental studies for human microsatellite mutations in vivo, high mutation rates from about 10^{-4} to 10^{-2} per locus per generation were observed. Besides single step mutational events, some multiple steps mutational events were also observed. Zhang et al. (1994) observed that longer trinucleotide repeats had much higher mutation rates than short ones and that contractions occurred more frequently than expansions. Xu et al. (2000) observed more mutations and contractions for longer tetranucleotide repeats. Bacon, Dunlop, and Farrington (2001) observed high mutation rates for mononucleotides. Huang et al. (2002) observed that the mutation rate increased and the probability of expansion given mutation occurrence decreased as the number of repeat units increased for dinucleotides. Length-dependent mutation patterns of microsatellites were also observed from different organisms, such as flies (Harr and Schlötterer 2000) and yeast (Wierdl, Dominska, and Petes 1997). In all those experiments, the numbers of observed mutations were not large enough to give clear patterns for the relationship between microsatellite slippage mutation rate and the number of repeat units.

With the whole genome sequence available, it is possible to collect a large volume of data for microsatellite distributions. The equilibrium assumption assumes that the observed distributions of this generation are the same as those of the next generation. Together with the equilibrium assumption, it is possible to estimate microsatellite mutation rates. Bell and Jurka (1997) first proposed such an approach and applied it to some genome sequences. Kruglyak et al. (1998, 2000) extended such an idea and proposed a novel estimation method. Sibly, Whittaker, and Talbort (2001) further generalized it with a maximum likelihood estimation method. Those studies were based on the symmetric single stepwise model that assumes the expansion rate to be the same as the contraction rate. A recent study by Sibly et al. (2003) found that the symmetric single stepwise model for microsatellite slippage mutations cannot explain the observed human sequence data. In a recent study by Calabrese and Durrett (2003), they found that it was difficult to model microsatellite slippage mutations using simple functions. They observed a bias toward contraction for long microsatellites

[1] Present address: Department of Epidemiology and Public Health, Yale University School of Medicine.

Key words: microsatellites, Markov processes, branching processes.

E-mail: fsun@hto.usc.edu.

Mol. Biol. Evol. 20(12):2123–2131. 2003
DOI: 10.1093/molbev/msg228
Molecular Biology and Evolution, Vol. 20, No. 12,

Table 1
An Example for Collecting Tetranucleotides Classified by Their Length

	Repeat 2	Repeat 3	Repeat 4	Repeat 5
all repeats	1 (7,14)	1 (19,30)	1 (45,60)	1 (61, 80)
perfect repeats	1 (7,14)	0	1 (45,60)	1 (61, 80)

NOTE.—The results for collecting tetranucleotides from the sequence ''TTTAAA*ATGT ATGT*ATCT*AT GTATGTATGT* TTTAAACACA CACA*ATGGAT GGATGGATGG CAGGCAGGCA GGCAGGCAGG* TA'' with a space after every 10 nucleotide bases. ''repeat *i*'' means the number of tetranucleotides with the number of repeat units equal to *i*. The numbers in parentheses are the start and end positions of the collected tetranucleotides.

by assuming a quadratic model or piecewise linear model for slippage mutation rates. Most of the previous approaches were based on the single stepwise mutation model. This simplified model can reflect microsatellite mutation mechanisms because single-step mutational events were the major mutational events observed in experiments. In previous studies (Bell and Jurka 1997; Kruglyak et al. 1998, 2000; Sibly, Whittaker, and Talbort 2001; Calabrese and Durrett 2003; Sibly et al. 2003), a constant, linear, or quadratic relationship between microsatellite slippage mutation rate and the number of repeat units was assumed. Such assumptions are not strongly supported by the experimental results (Zhang et al. 1994; Xu et al. 2000; Bacon, Dunlop, and Farrington 2001; Huang et al. 2002).

In this study, we propose a novel method using two sets of equations based on two stochastic processes to estimate microsatellite slippage mutation rates. This study differs from previous studies by introducing a new multitype branching process in addition to the stationary Markov process proposed before (Bell and Jurka 1997; Kruglyak et al. 1998, 2000; Sibly, Whittaker, and Talbort 2001; Calabrese and Durrett 2003; Sibly et al. 2003). The distributions from the two processes make it possible to estimate microsatellite slippage mutation rates without assuming any relationship between microsatellite slippage mutation rate and the number of repeat units. We apply our method to the sequence data from the human genome. We also develop a novel method for estimating the threshold size for slippage mutations. In the following paragraphs, we first explain our method for data collection and the mathematical model; we then present estimation results.

Materials and Methods

In this section, we first describe how the data are collected from public sequence database. Then, we introduce two stochastic processes to model the collected data. Based on the equilibrium assumption that the observed distributions of this generation are the same as those of the next generation, two sets of equations are derived for estimation purposes. Next, we introduce a novel method for estimating threshold size for microsatellite slippage mutation. Finally, we give the details of our estimation method.

Data Collection

We downloaded the human genome sequence from the National Center for Biotechnology Information database *ftp://ftp.ncbi.nih.gov/genbank/genomes/H_sapiens/OLD/*(updated on September 06, 2001). We collected mono-, di-, tri-, tetra-, penta-, and hexa- nucleotides in two different schemes. The first scheme is simply to collect *all repeats* that are microsatellites without interruptions among the repeats. The second scheme is to collect *perfect repeats* (Sibly, Whittaker, and Talbort 2001), such that there are no interruptions among the repeats and the left flanking region (up to $2l$ nucleotides) does not contain the same motifs when microsatellites (of motif with l nucleotide bases) are collected. Mononucleotides were excluded when di-, tri-, tetra-, penta-, and hexa- nucleotides were collected; dinucleotides were excluded when tetra- and hexa- nucleotides were collected; trinucleotides were excluded when hexanucleotides were collected. For a fixed motif of l nucleotide bases, microsatellites with the number of repeat units greater than 1 were collected in the above manner. The number of microsatellites with one repeat unit was roughly calculated by [(total number of counted nucleotides) $-\sum_{i>1} l \times i \times$ (number of microsatellites with i repeat units)]/l. All the human chromosomes were processed in such a manner. Table 1 gives an example of the two schemes.

Mathematical Models and Equations

We study two models for microsatellite mutations. For *all repeats*, we use a multi-type branching process. For *perfect repeats*, we use a Markov process as proposed in previous studies (Bell and Jurka 1997; Kruglyak et al. 1998, 2000; Sibly, Whittaker, and Talbort 2001; Calabrese and Durrett 2003; Sibly et al. 2003). Both processes are discrete time stochastic processes with finite integer states $\{1, 2, \ldots, N\}$ corresponding to the number of repeat units of microsatellites. To guarantee the existence of equilibrium distributions, we assume that the number of states N is finite. In practice, N could be an integer greater than or equal to the length of the longest observed microsatellite. In both models, we consider two types of mutations: point mutations and slippage mutations. Because single-nucleotide substitutions are the most common type of point mutations, we only consider single-nucleotide substitutions for point mutations in our models. Because the number of nucleotides in a microsatellite locus is small, we assume that there is at most one point mutation to happen for one generation. Let a be the point mutation rate per repeat unit per generation, and let e_k and c_k be the expansion slippage mutation rate and contraction slippage mutation rate, respectively. In the following models, we assume that $a > 0$; $e_k > 0, 1 \le k \le N-1$ and $c_k \ge 0, 2 \le k \le N$.

Modeling all repeats

For *all repeats*, we consider the following stochastic process:

1. After one generation, by microsatellite slippage mutations, any state k can change to state $k + 1$ with

probability e_k for $1 \le k \le N-1$ and state $k-1$ with probability c_k for $2 \le k \le N$.

2. After one generation, by point mutations, one microsatellite will break, and the breaking point will be another microsatellite motif. So with probability a per repeat unit, a microsatellite of state k, $2 \le k \le N$ can generate one of state 1, one of state h, and one of state $k-h-1$ where $1 \le h \le k-1$, if the point mutation does not happen at the boundary. Otherwise, we will have one of state 1, and one of state $k-1$.

Let $Z_n = (Z_{n1}, Z_{n2}, \ldots, Z_{nN})$, $n = 0, 1, 2, \ldots$ be the corresponding stochastic process, where Z_{nk} is the number of microsatellites with k repeat units after n generations. $\{Z_n\}$ forms a multi-type branching process. The first moments matrix (Harris 1963) is given by

$$M = \begin{pmatrix} 1-e_1 & e_1 & 0 & \ldots & & & & & & & \\ 4a+c_2 & 1-2a-e_2-c_2 & e_2 & 0 & \ldots & & & & & & \\ \ldots & & & & & & & & & & \\ (k+2)a & 2a & 2a & \ldots & 2a+c_k & 1-ka-e_k-c_k & e_k & 0 & \ldots & & \\ \ldots & & & & & & & & & & \\ (N+2)a & 2a & 2a & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & 2a+c_N & 1-ka-c_N \end{pmatrix}.$$

The jth entry in the ith row in the matrix M represents the expected number of microsatellites with j repeat units in the next generation when there is only one microsatellite with i repeat units in this generation. We explain in more detail about the expected number generated by point mutations. For a microsatellite with k repeat units, there are k possible repeat units in which a point mutation may occur. The breaking point will generate another microsatellite motif, which is of state 1. If the point mutation happens at the second position or the penultimate position, we will have a microsatellite of state 1 at the boundary. Therefore, the expected number of microsatellites of state 1 is $(k+2)a$.

From the *Perron-Frobenius Theorem* (Harris 1963), the *Perron-Frobenius eigenvalue* of M is greater than 1, and we denote it as $1+\lambda$. We denote the left *Perron-Frobenius eigenvector* of M as $\mathbf{p} = (p_1, p_2, \ldots, p_N)$. From the theory of multi-type branching processes (Harris 1963; Athreya and Ney 1972; Athreya and Vidyashankar 1995), we have $\lim_{n\to\infty} Z_n/|Z_n| = \mathbf{p}$. Here $|\cdot|$ means to sum over all the entries of a vector. Therefore, the distribution of *all repeats* will converge to $\mathbf{p}$. From equation $\mathbf{p}M = (1+\lambda)\mathbf{p}$, we have

$$p_k e_k - p_{k+1} c_{k+1} = a r_{k+1} + 2a(w_2 + w_3 + \cdots + w_{k+1}) - \lambda v_k, \quad (1)$$

for all $0 < k < N$, where $r_k = \sum_{i=k}^{N} i p_i$, $w_k = \sum_{i=k}^{N} p_i$ and $v_k = \sum_{i=1}^{k} p_i$.

Modeling perfect repeats

For *perfect repeats*, we consider the following Markov process proposed in previous studies (Bell and Jurka 1997; Kruglyak et al. 1998, 2000; Sibly, Whittaker, and Talbort 2001; Calabrese and Durrett 2003; Sibly et al. 2003).

1. After one generation, by microsatellite slippage mutations, any state k can change to state $k+1$ with probability e_k for $1 \le k \le N-1$ and state $k-1$ with probability c_k for $2 \le k \le N$.
2. After one generation, by point mutations, any state k, $2 \le k \le N$ can change to some state h, $1 \le h \le k-1$. There are k positions that a point mutation can happen equally likely with probability a per repeat unit. If a point mutation happens at the boundary, we will collect a microsatellite with repeat $k-1$. Therefore by point mutation, any state k can change to state h, $1 \le h \le k-2$ with probability a, and state $k-1$ with probability $2a$.

Let X_n, $n = 0, 1, 2, \ldots$ be the corresponding stochastic process. $\{X_n\}$ forms a Markov process. The transition matrix is given by

$$P = \begin{pmatrix} 1-e_1 & e_1 & 0 & \ldots & & & & & & & \\ 2a+c_2 & 1-2a-e_2-c_2 & e_2 & 0 & \ldots & & & & & & \\ \ldots & & & & & & & & & & \\ a & a & a & \ldots & 2a+c_k & 1-ka-e_k-c_k & e_k & 0 & \ldots & & \\ \ldots & & & & & & & & & & \\ a & a & a & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & 2a+c_N & 1-ka-c_N \end{pmatrix}.$$

From the theory of Markov process, there is a stationary distribution $\mathbf{q} = (q_1, q_2, \ldots, q_N)$ with $\mathbf{q}P = \mathbf{q}$, which is equivalent to

$$q_k e_k - q_{k+1} c_{k+1} = a(k+1)u_{k+1} - a u_{k+2}, \qquad (2)$$

for all $0 < k < N$, where $u_k = \sum_{j=k}^{N} q_j$.

Two Sets of Equations

Note that a is a nuisance parameter in both models. We can only estimate the relative expansion slippage rates and contraction slippage rates compared to the point mutation rate. We divide both sides of equations (1) and (2) by a and denote λ, e_k, and c_k for the previous λ/a, e_k/a and c_k/a, respectively. We have the following two fundamental equations.

$$\begin{cases} p_k e_k - p_{k+1} c_{k+1} = r_{k+1} + 2(w_2 + w_3 + \cdots + w_{k+1}) - \lambda v_k, \\ q_k e_k - q_{k+1} c_{k+1} = (k+1)u_{k+1} - u_{k+2}. \end{cases} \qquad (3)$$

Compared to microsatellites slippage mutation rates, point mutation rates are relatively small. The difference between the matrices M and P is of the level of point mutation rate a, which is very small. Therefore, we expect only slight differences between the two distributions $\mathbf{p}$ and $\mathbf{q}$ when they are normalized.

For convenience, the above point mutation rate a is the point mutation rate for the whole motif. We will apply our estimation method to sequence data of mono-, di-, tri-, tetra-, penta-, and hexa- nucleotides. Therefore, a is different for microsatellites with motifs of different numbers of nucleotide bases. The estimation results are the relative ratios between the slippage mutation rate and point mutation rate. To keep the estimation results comparable, we will multiply the estimated slippage mutation rates by the motif length l.

Threshold Size

We define microsatellite slippage threshold size T as the number of repeat units such that $c_k = 0$, $2 \le k \le T$ and $c_k > 0$, for $k > T$. Under this threshold size T, there are almost no slippage mutations; Above T, microsatellites slippage mutations will dominate point mutations.

For the observed distributions $\{p_k\}$ for *all repeats* and $\{q_k\}$ for *perfect repeats*, we consider their sequential ratios $\{p_{k+1}/p_k\}$ and $\{q_{k+1}/q_k\}$. A null hypothesis is that there is no microsatellite slippage mutation and that microsatellites are generated by random arrangement by different nucleotides (Pupko and Graur 1999; Rose and Falush 1998). Under this hypothesis, $\{p_k\}$ and $\{q_k\}$ should follow a geometric distribution. Therefore, we expect that the sequential ratios are all of relatively low and constant level.

If the sequential ratios can keep a relatively low and constant level up to L, then the observed fractions of states up to $L+1$ can be explained by the above null hypothesis. This implies that there is almost no slippage mutation from $L+2$ to $L+1$. Therefore, we can estimate the threshold size T by $L+2$.

Estimating Slippage Mutation Rates

When the number of repeat units is below T, microsatellite slippage mutation rates are small and can be regarded as 0. In the following paragraphs, we will examine only slippage mutation rates of microsatellites with a number of repeat units greater than T. Statistically, the estimated results will be reliable only when we have a large number of observations. Therefore, we estimate slippage mutation rates of microsatellites with a number of repeat units ranging from $T+1$ to $H-1$, where H is the minimum number of repeat units for which either the observed number of *all repeats* or *perfect repeats* with H repeat units is less than 100.

The estimated threshold size for microsatellites slippage mutation is useful for computing the *Perron-Frobenius eigenvalue* $1 + \lambda$. On the threshold size T, we set the contraction slippage mutation rate $c_T = 0$. Then λ and e_{T-1} can be obtained by directly solving equations (3) for $k = T-1$. With λ available, we can estimate e_k and c_{k+1} using equations (3) for $k \ge T$. Owing to random variation of the observations, some of solved values for e_k and c_{k+1} are negative. It was observed from experiments (Zhang et al. 1994; Xu et al. 2000; Huang et al. 2002) that the contraction slippage mutation rate increased with the number of repeat units. We thus use the following strategy to guarantee non-negative solutions: If the direct solutions e_k and c_{k+1} from equations (3) are all non-negative, we will accept them. Otherwise, we set $c_{k+1} = c_k$ and compute e_k using the least squares method for equations (3). The confidence intervals for our estimated slippage mutation rates can be obtained using the bootstrap method (Effron 1979).

Results

Microsatellites Frequencies

Data were collected for 22 autosomes, chromosome X, and chromosome Y. The observed distributions from different chromosomes had similar patterns (data not shown), indicating that the microsatellite mutation mechanism is similar for different chromosomes. Therefore, we combined the distributions for all the chromosomes together as the observed distribution.

Figure 1 shows the observed frequency in logarithm scale for *all repeats* $\{p_k\}$ and *perfect repeats* $\{q_k\}$ (see *Materials and Methods* for details). We observe that mononucleotides are the most abundant microsatellites in the human genome, followed by dinucleotides, trinucleotides, etc. Microsatellites can contain a large number of repeat units, with the observation of more than 65 for mononucleotides and more than 49 for dinucleotides. Overall, microsatellite frequencies decrease exponentially as the number of repeat units increases. But the shape of the frequency distribution is not regular, with different slopes in different intervals of the number of repeat units. Around repeat 36 for mononucleotides, repeat 10 for tetranucleotides, we observe ''humps.'' The complicated shape of microsatellite frequency distributions indicates that the microsatellite mutation mechanism is complicated.

The Threshold Size

Figure 2 shows the observed sequential ratios of *all repeats* and *perfect repeats* for microsatellites in the human genome (See *Materials and Methods* for definition of sequential ratio). We observe that the sequential ratios keep a relatively low and constant level up to 7, 2, 2, 2, 2, and 2 for mono-, di-, tri-, tetra-, penta-, and hexa-nucleotides, respectively, and then the sequential ratios suddenly jump and maintain a high and fluctuating level. Those observations provide evidence for the existence of the threshold size for microsatellite slippage mutations. Based on our criterion for estimating the threshold size (see *Materials and Methods* for details), the estimated threshold size T is 9, 4, 4, 4, 4, or 4 for mono-, di-, tri-, tetra-, penta-, or hexa- nucleotides, respectively. The results are also given in table 2.

Estimating Slippage Mutation Rates

Figure 3 shows the total estimated slippage mutation rates $\{e_k + c_k\}$ in logarithm scale, and figure 4 shows the estimated expansion ratio $\{e_k/(e_k + c_k)\}$ together with confidence intervals. Our estimation results show that microsatellites with different motif lengths have similar mutation mechanisms. There is an exponentially increasing trend for the estimated slippage rate $\{e_k + c_k\}$, and a decreasing trend for the estimated expansion ratio $\{e_k/(e_k + c_k)\}$. Our results are consistent with the estimated level of mutation rates from experimental studies (Zhang et al. 1994; Xu et al. 2000; Bacon, Dunlop, and Farrington 2001; Huang et al. 2002) in which higher mutability and more contractions than expansions were observed at longer microsatellite loci. The point mutation rate per nucleotide per generation is of the level of $a = 10^{-8}$ (Li 1997). Using this quantity, the estimated total slippage rates from $T + 1$ to $H - 1$ are as given in table 3. The estimated average mutation rates based on experimental studies are 1.94×10^{-4} for dinucleotides (Huang et al. 2002) and 1.8×10^{-3} for tetranucleotides (Xu et al. 2000), both within the above intervals from our estimation. In the experiments with mononucleotides (Bacon, Dunlop, and Farrington 2001), only *BAT-40* loci were studied. Most *BAT-40* loci contain about 40 poly-(A/T) repeat units, and the estimated average mutation rate is 6.95×10^{-2}. Our estimated slippage mutation rate for mononucleotides with 40 repeat units is 8.8×10^{-3} with confidence level (1.8×10^{-3}, 1.6×10^{-1}). In experiments with trinucleotides (Zhang et al. 1994), the estimated average mutation rate is 1.3×10^{-2} for 20–22 repeat alleles and 4.4 times higher for 28–31 repeat alleles. Those numbers of repeat units in the experiment of Zhang et al. are not in our estimation range (5–17). But if we extend the trend of our estimated slippage mutation rates for trinucleotides in figure 3, we expect to have the same level of slippage mutation rate as reported in experimental studies. Interestingly, in figure 3 all of our estimated values including the confidence intervals, are less than 1 when the point mutation rate is set at $a = 10^{-8}$.

As shown in figure 3 and figure 4, the patterns of the estimated slippage mutation rates and expansion ratios are quite complicated. Overall, we can roughly use a line to fit the estimated slippage mutation rates in logarithm scale, which implies an exponential relationship between the slippage mutation rate and the number of repeat units. The trend of the estimated expansion ratios looks like it is decreasing exponentially.

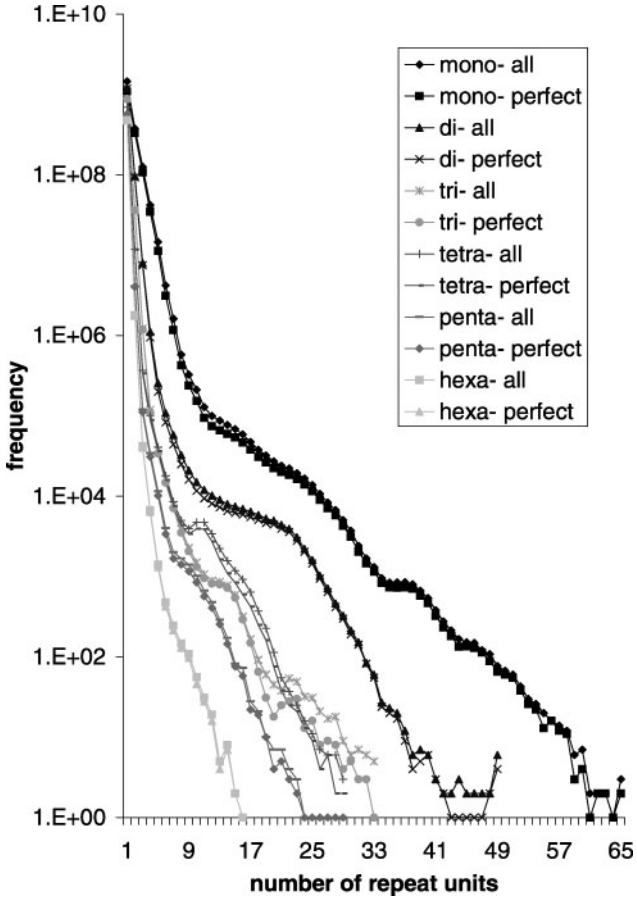

FIG. 1.—The observed frequencies for mono-, di-, tri-, tetra-, penta-, and hexa- nucleotides for the human genome. In the figure, ''all'' denotes observed frequencies of *all repeats* $\{p_k\}$, ''perfect'' denotes observed frequencies of *perfect repeats* $\{q_k\}$. The X-axis denotes the number of repeat units, and the Y-axis denotes the value of frequency in logarithm scale.

We obtained 95% confidence intervals using the bootstrap method. Those confidence intervals for our estimations are shown in figure 3 and figure 4. Because the number of microsatellites decreases rapidly as the number of repeat units increases, the interval becomes wider as the number of repeat units increases.

Discussion

A slippage mutation threshold size was estimated by a previous *in silico* study for the yeast *Saccharomyces cerevisiae* (Rose and Falush 1998), where the authors claimed that a minimum threshold size of about 8 nucleotide bases is necessary for slippage mutations. A similar threshold size was observed in previous studies of microsatellite mutation during polymerase chain reaction (Lai et al. 2003; Lai and Sun 2003; Shinde et al. 2003). In

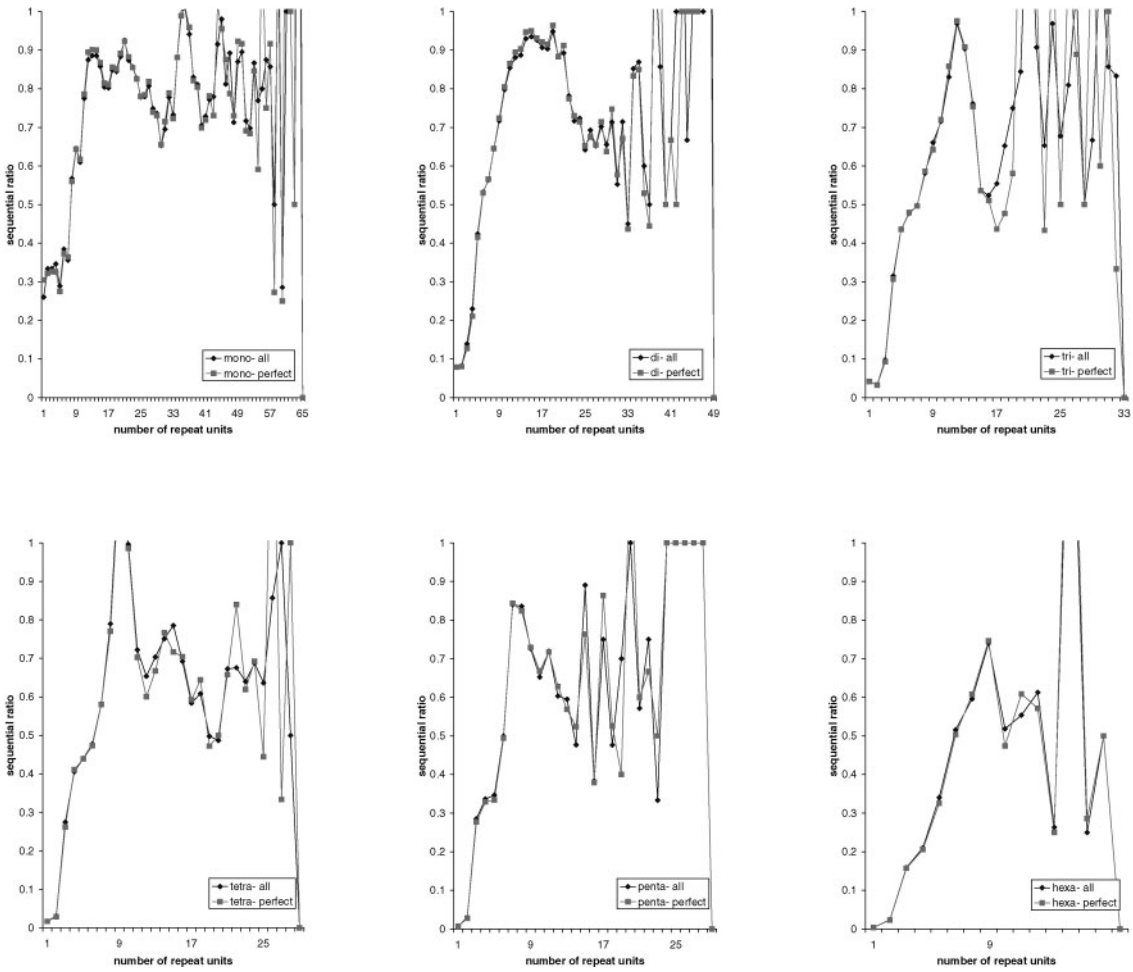

FIG. 2.—The observed sequential ratios: (the first row from left to right) mono-, di-, tri-nucleotides; (the second row from left to right) tetra-, penta-, and hexa- nucleotides. In the figure, ''all'' denotes observed sequential ratios of *all repeats* $\{p_{k+1}/p_k\}$; ''perfect'' denotes observed sequential ratios of *perfect repeats* $\{q_{k+1}/q_k\}$. The X-axis denotes the number of repeat units, and the Y-axis denotes the value of ratio. We set the range for Y-axis from 0 to 1 because most sequential ratios are smaller than 1.

the present study, we also observed evidence of a slippage mutation threshold. The results from those studies suggest common features of microsatellite mutation mechanism both in vivo and in vitro.

Using two sets of equations based on a multi-type branching process and a Markov process, we estimated mutation rates of microsatellites in the human genome without assuming any relationship between microsatellite slippage mutation rate and the number of repeat units. The novelty of this study is the introduction of a multi-type branching process. In previous studies involving only the Markov process, some relationship between the microsatellite slippage mutation rate and the number of repeat units has to be assumed. Our method can also be applied to estimate microsatellite mutation mechanisms for other organisms when large amounts of genome sequence data are available. It is possible to compare microsatellite mutation mechanisms among different organisms.

We observed an exponentially increasing trend for the estimated slippage mutation rates and a decreasing trend for the estimated slippage expansion ratios. The total slippage mutation rate may differ up to $10^3 \sim 10^4$-fold for different numbers of repeat units. Our estimation results are consistent with experimental studies (Zhang et al. 1994; Xu et al. 2000; Bacon, Dunlop, and Farrington 2001; Huang et al. 2002) and computational studies (Calabrese and Durrett 2003). Long microsatellites are highly unstable and likely to mutate. When slippage mutations happen, expansions occur more frequently if the number of repeat units is small, and contractions occur more frequently if the number of repeat units is large. When mutations happen, long microsatellites are likely to mutate to shorter ones; short microsatellites are likely to mutate to longer ones. The scarcity of large numbers of repeat units in a microsatellite locus can be explained by

Table 2
The Estimated Threshold Size

mono-	di-	tri-	tetra-	penta-	hexa-
9	4	4	4	4	4

NOTE.—The estimated threshold size for microsatellite slippage mutations in the table denotes the number of repeat units (motif) of a microsatellite.

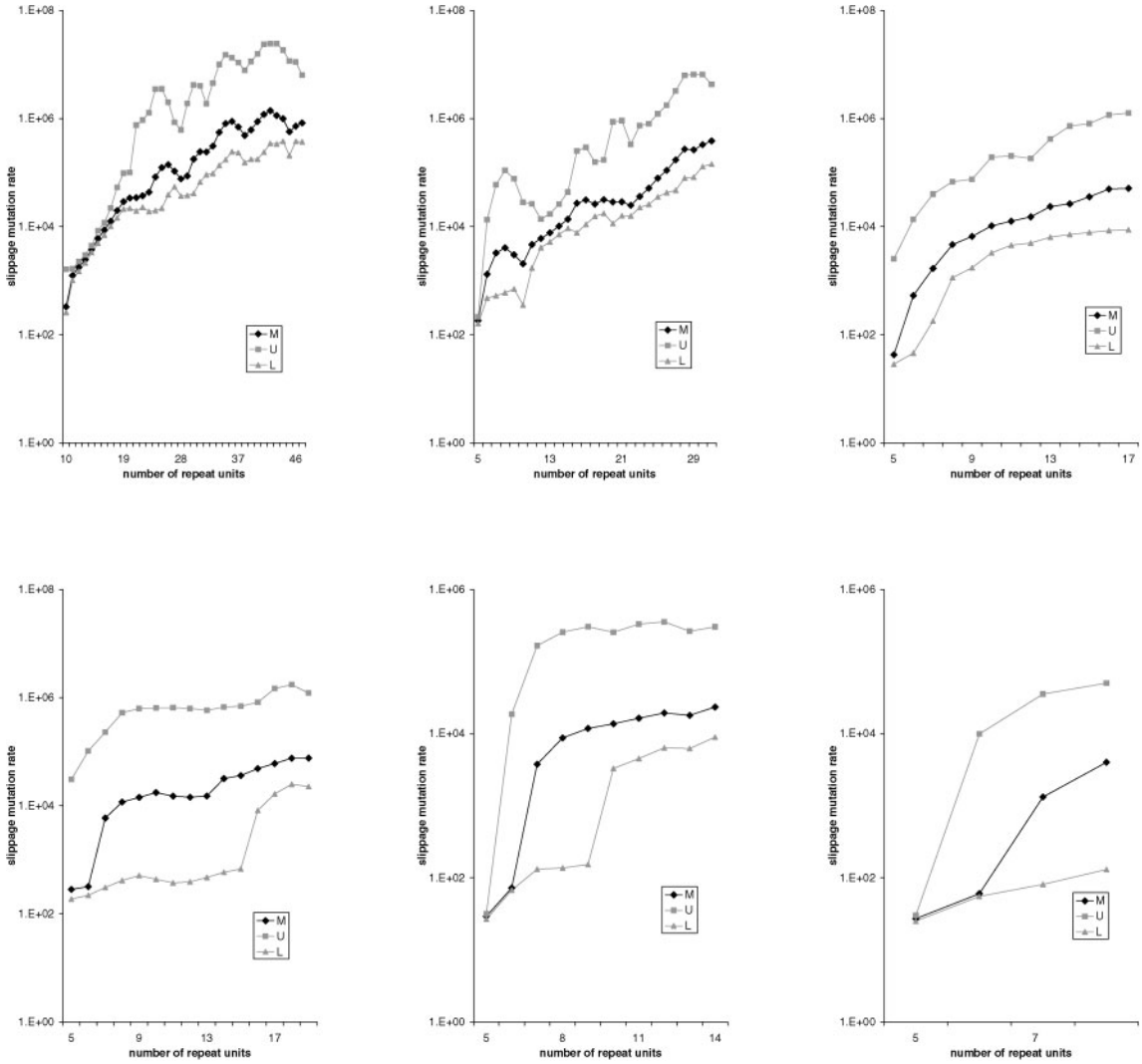

FIG. 3.—The estimated results for the total slippage mutation rates for different microsatellites: (the first row from left to right) mono-, di-, tri-nucleotides; (the second row from left to right) tetra-, penta-, and hexa- nucleotides. The X-axis denotes the number of repeat units, and the Y-axis denotes the ratio in logarithm scale of the estimated total slippage mutation rate compared to the point mutation rate $(c_k + e_k)/a$. M, U, and L denote the median, upper 2.5%, and lower 2.5% quantiles from the bootstrap estimation.

the high mutation rate and downward mutation bias when the number of repeat units is large.

As Calabrese and Durrett (2003) have pointed out, it is difficult to describe microsatellite slippage mutation rates using simple functions. We observe complicated patterns in our estimated results, which suggests that the microsatellite slippage mutation mechanism is complicated.

It is possible that genetic characteristics of local sequences influence the microsatellites mutation mechanism. Calabrese and Durrett (2003) applied comparative studies to show that local dinucleotide distributions were not significantly different for the regions with different local recombination rates, proximity to genes, local *GC* contents, location on the chromosome, and proximity to *Alu* repeats. Such results support the approach to estimating microsatellite slippage mutation rates using whole genome sequence data.

There are several limitations to our approach. One is that we grouped all the motifs with the same length together in this study. Different motifs may have different mutation mechanisms, and their mutation mechanisms need to be studied separately when enough data become available. In the present study, we assumed that the distribution of the number of *perfect repeats* and *all repeats* had achieved equilibrium, a common assumption

Table 3
The Range of the Estimated Slippage Mutation Rates

Microsatellites	Repeats Range	Slippage Mutation Rates Range
mono-	[10, 47]	$[10^{-6}, 10^{-2}]$
di-	[5, 31]	$[10^{-6}, 10^{-2}]$
tri-	[5, 17]	$[10^{-6}, 10^{-3}]$
tetra-	[5, 19]	$[10^{-6}, 10^{-3}]$
penta-	[5, 14]	$[10^{-7}, 10^{-4}]$
hexa-	[5, 8]	$[10^{-7}, 10^{-4}]$

NOTE.—The range of our estimated slippage mutation rates when point mutation rate a is set at 10^{-8} per nucleotide per generation.

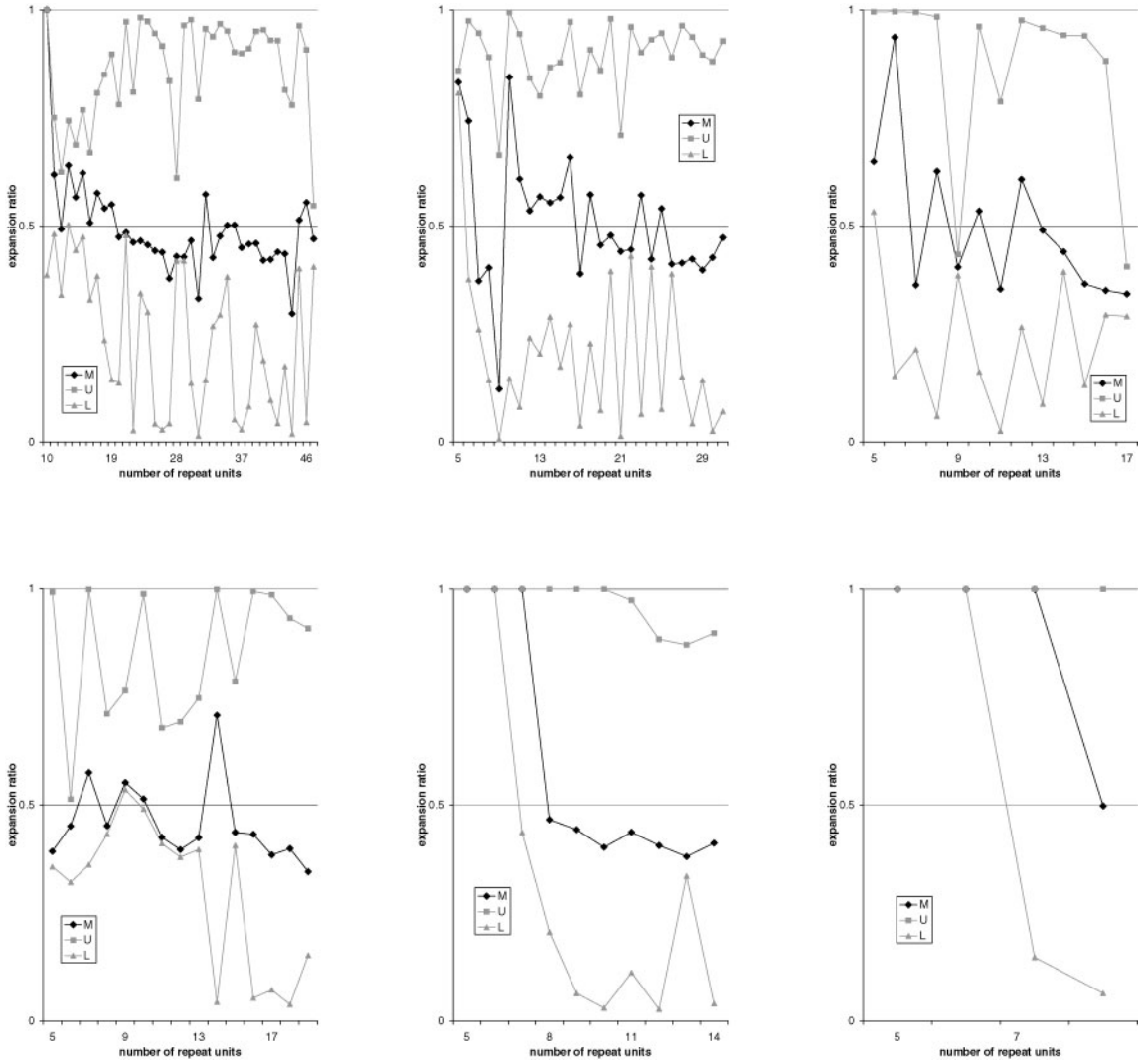

FIG. 4.—The estimated results for the expansion ratios for different microsatellites: (the first row from left to right) mono-, di-, tri- nucleotides; (the second row from left to right) tetra-, penta-, and hexa- nucleotides.The X-axis denotes the number of repeat units, and the Y-axis denotes the value of expansion ratio $e_k/(c_k + e_k)$. M, U, and L denote the median, upper 2.5%, and lower 2.5% quantiles from the bootstrap estimation.

in almost all the studies of similar type. An important question is how to test if the distributions have achieved equilibrium. These questions need to be considered in future studies.

Acknowledgments

We thank Dr. Peter Calabrese for helpful discussions and suggestions that improve the presentation of the manuscript. We also appreciate valuable suggestions and comments from two anonymous reviewers. This research was supported in part by National Institutes of Health grant DK 53392.

Literature Cited

Ashley, C. T., and S. T. Warren. 1995. Trinucleotide repeat expansion and human disease. Annu. Rev. Genet. **29**:703–728.

Athreya, K. B., and P. E. Ney. 1972. Branching processes. Springer-Verlag, Berlin.

Athreya, K. B., and A. N. Vidyashankar. 1995. Large deviation rates for branching processes. II. The multitype case. Ann. Appl. Probab. **5**:566–576.

Bacon, A., M. G. Dunlop, and S. M. Farrington. 2001. Hypermutability at a poly(A/T) tract in the human germline. Nucleic Acids Res. **29**:4405–4413.

Bell, G. I., and J. Jurka. 1997. The length distribution of perfect dimer repetitive DNA is consistent with its evolution by an unbiased single-step mutation process. J. Mol. Evol. **44**:414–421.

Calabrese, P., and R. Durrett. 2003. Dinucleotides repeats in the Drosophila and Human genome have complex, length-dependent mutation processes. Mol. Biol. Evol. **20**:715–725.

Effron, B. 1979. Bootstrap method: another look at the Jackknife. Ann. Stat. **7**:1–26.

Harr, B., and C. Schlötterer. 2000. Long microsatellite alleles in *Drosophila melanogaster* have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. Genetics **155**:1213–1220.

Harris, T. E. 1963. The theory of branching processes. Springer-Verlag, Berlin.

Huang, Q. Y., F. H. Xu, H. Shen, H. Y. Deng, Y. J. Liu, Y. Z. Liu, J. L. Li, R. R. Recker, and H. W. Deng. 2002. Mutation patterns at dinuleotide microsatellite loci in humans. Am. J. Hum. Genet. **70**:625–634.

Kong, A., D. F. Gudbjartsson, J. Sainz et al. (16 co-authors). 2002. A high-resolution recombination map of the human genome. Nat. Genet. **31**:241–247.

Kruglyak, S., R. T. Durrett, M. D. Schug, and C. F. Aquadro 1998. Equilibrium distribution of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc. Natl. Acad. Sci. USA **95**:10774–10778.

———. 2000. Distribution and abundance of microsatellites in the yeast genome can be explained by a balance between slippage events and point mutations. Mol. Biol. Evol. **17**:1210–1219.

Lai, Y., D. Shinde, N. Arnheim, and F. Z. Sun. 2003. The mutation process of microsatellites during the polymerase chain reaction. J. Comp. Biol. **10**:143–155.

Lai, Y., and F. Z. Sun. 2003. Microsatellite mutations during the polymerase chain reaction: mean field approximations and their applications. J. Theor. Biol. **224**:127–137.

Leeflang, E. P., S. Tavaré, P. Marjoram, C. O. S. Neal, J. Srinidhi, M. E. MacDonald, M. Young, N. S. Wexler, J. F. Gusella, and N. Arnheim. 1999. Analysis of germline mutation spectra at the Huntington's disease locus supports a mitotic mutation mechanism. Hum. Mol. Genet. **8**:173–183.

Li, W. H. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.

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

Ott, J. 1999. Analysis of human genetic linkage. The Johns Hopkins University Press, Baltimore and London.

Pupko, T., and D. Graur. 1999. Evolution of microsatellites in the yeast *Saccharomyces cerevisiae*: role of length and number of repeated units. J. Mol. Evol. **48**:313–316.

Rose, O., and D. Falush. 1998. A threshold size for microsatellite expansion. Mol. Biol. Evol. **15**:613–615.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman. 2002. Genetic structure of human populations. Science **298**:2381–2385.

Shinde, D., Y. Lai, F. Z. Sun, and N. Arnheim. 2003. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: $(CA/GT)_n$ and $(A/T)_n$ microsatellites. Nucleic Acids Res. **31**:974–980.

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

Sibly, R. M., J. C. Whittaker, and M. Talbort. 2001. A maximum-likelihood approach to fitting equilibrium models of microsatellite evolution. Mol. Biol. Evol. **18**:413–417.

Sibly, R. M., A. Meade, N. Boxall, M. J. Wilkinson, D. W. Corne, and J. C. Whittaker. 2003. The structure of interrupted human AC microsatellites. Mol. Biol. Evol. **20**:453–459.

Sturzeneker, R., R. A. U. Bevilacqua, L. A. Haddad, A. J. G. Simpson, and S. D. J. Pena. 2000. Microsatellite instability in tumors as a model to study the process of microsatellite mutations. Hum. Mol. Genet. **9**:347–352.

Viguera, E., D. Canceill, and S. D. Ehrlich. 2001. Replication slippage involves DNA polymerase pausing and dissociation. EMBO J. **20**:2587–2595.

Weber, J., and P. 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.

Wierdl, M., M. Dominska, and T. D. Petes. 1997. Microsatellite instability in yeast: dependence on the length of the microsatellite. Genetics **146**:769–779.

Xu X., M. Peng, Z. Fang, and X. Xu. 2000. The direction of microsatellite mutations is dependent upon allele length. Nat. Genet. **24**:396–399.

Zhang, L., E. P. Leeflang, J. Yu, and N. Arnheim. 1994. Studying human mutations by sperm typing: instability of CAG trinucleotide repeats in the human androgen receptor gene. Nat. Genet. **7**:531–535.

Adam Eyre-Walker, Associate Editor

Accepted July 23, 2003