Distribution and Abundance of Microsatellites in the Yeast Genome Can Be Explained by a Balance Between Slippage Events and Point Mutations

Semyon Kruglyak, Richard Durrett,† Malcolm D. Schug,†[1] and Charles F. Aquadro‡*

*Department of Mathematics, University of Southern California; and †Department of Mathematics and ‡Department of Molecular Biology and Genetics, Cornell University

We fit a Markov chain model of microsatellite evolution introduced by Kruglyak et al. to data on all di-, tri-, and tetranucleotide repeats in the yeast genome. Our results suggest that many features of the distribution of abundance and length of microsatellites can be explained by this simple model, which incorporates a competition between slippage events and base pair substitutions, with no need to invoke selection or constraints on the lengths. Our results provide some new information on slippage rates for individual repeat motifs, which suggest that AT-rich trinucleotide repeats have higher slippage rates. As our model predicts, we found that many repeats were adjacent to shorter repeats of the same motif. However, we also found a significant tendency of microsatellites of different motifs to cluster.

Introduction

Microsatellites are tandem repeats of short units (1–5 nt) of DNA. In humans, triplet repeats are involved in at least a dozen diseases, including Huntington's disease, fragile X syndrome, and myotonic dystrophy (see, e.g., Ashley and Warren 1995). However, the majority of such repeats occur in noncoding regions, and presumably most do not have significant selective consequences.

Microsatellite loci have a high degree of variability that is due to the high rate of mutations that alter their length. For this reason, they have been useful in testing for paternity (Ashworth et al. 1998; Foster et al. 1998), looking for disease genes (Bloutin et al. 1998; Mein et al. 1998), and studying the evolutionary history of humans (Goldstein et al. 1995; Reich and Goldstein 1998; Underhill et al. 1998; Pritchard et al. 1999; Ruiz-Linares et al. 1999). For more references, see the survey by Goldstein and Pollock (1997).

For many of these applications, one needs estimates of mutation rates and a realistic model of microsatellite evolution. Since the primary mechanism leading to changes in microsatellite length is polymerase slippage, and most changes are by ± 1 repeat unit (Brinkman et al. 1998), it has been common to use the stepwise mutation model of Ohta and Kimura (1973), in which microsatellites change by ± 1 unit at a rate μ, independent of their length.

The stepwise mutation model, however, has the drawbacks that lengths may become negative and the collection of repeat lengths in a sample does not have a stationary distribution. The problem with negative lengths is easy to fix: one simply forbids transitions to values less than 1. To address the absence of a stationary distribution, several researchers have imposed an upper limit on microsatellite lengths (Bell and Jurka 1997; Feldman et al. 1997) or introduced a drift toward length 0 (Stephan and Kim 1998) or toward a preferred length (Garza, Slatkin, and Freimer 1995; Zhivotovsky, Feldman, and Grisheckin 1997). However, it is not clear what biological mechanisms would be responsible for these effects.

Kruglyak et al. (1998) introduced a model in which a perfect repeat of length ℓ changes in length by ± 1 unit due to slippage at a rate $b(\ell - 1)$, while point mutations change the length from ℓ to j at rate a for each $1 \leq j < \ell$. To avoid an absorbing state at 1, they assumed that repeats of length 1 grow to size 2 at rate c due to base pair substitutions. Note that in contrast to the stepwise mutation model, slippage events occur at a rate proportional to the length (minus 1) of the repeat. In this model, an equilibrium distribution of microsatellites results from a balance between slippage events and point mutations. In Kruglyak et al. (1998), the model was fit to 1 Mb of sequence data from each of four organisms (humans, mice, fruit flies, and yeast), and these fits were used to estimate polymerase slippage rates. The rate estimates found were in good agreement with experimental results, which, in the case of yeast, were from Henderson and Petes (1992).

In the 2 years since our initial investigation, Petes and his coworkers have amassed a considerable amount of experimental results concerning the dynamics of microsatellite mutations in yeast. They have studied the dependence of mutation rates on the length of the microsatellite (Wierdl, Dominska, and Petes 1997) and how interruptions in the repeat lower mutation rates (Petes, Grenwell, and Dominska 1997). They have examined the influence on micro- and minisatellites of mutations in genes responsible for mismatch repair, DNA polymerase δ, and a nuclease involved in Okazaki fragment processing (Sia et al. 1997; Kokosa et al. 1998, 1999). Recently, they have considered the tendency of various families of triplet repeats to form secondary structures that escape DNA repair (see Moore et al. 1999).

In addition to experimental work mentioned above, the entire 12-Mb yeast (*Saccharomyces cerevisiae*) genome has now been sequenced (see Clayton et al. 1997).

[1] Present address: Department of Biology, University of North Carolina at Greensboro.

Key words: microsatellite, slippage rates, *Saccharomyces cerevisiae.*

Address for correspondence and reprints: Richard Durrett, Department of Mathematics, Cornell University, Ithaca, New York 14853. E-mail: rtd1@cornell.edu.

Mol. Biol. Evol. 17(8):1210–1219. 2000

For these reasons, we decided to take an in-depth look at the microsatellites in the yeast genome. Field and Wills (1998) previously examined microsatellites in the yeast genome, concentrating primarily on mononucleotide repeats and on comparison with the patterns found in eight microbial genomes. In contrast, the focus of our investigation is to compare the predictions of the model of the Kruglyak et al. (1998) model with the observed patterns of variation in yeast, so our research is complementary to their study and to the experimental work that seeks to determine the exact mechanisms involved in microsatellite mutations.

Materials and Methods

Our materials are the sequence of the 12-Mb yeast genome which can be found at (http://genome-www.stanford.edu/Sacchromyces). Programs written in the C programming language were used to count the number of microsatellites of various motifs and to determine the spacing between microsatellites. Once this was done, the methods described in Kruglyak et al. (1998) were used to estimate slippage rates. As indicated in equations (1) and (2) of Kruglyak et al. (1998), the Markov chain model has a stationary distribution in which repeats of length i have a probability $\pi(i)$ that satisfies

$$c\pi(1) = b\pi(2) + a\sum_{j=2}^{\infty}\pi(j) \quad (1)$$

$$b(i-1)\pi(i) = bi\pi(i+1) + ia\sum_{j=i+1}^{\infty}\pi(j) \quad \text{for } i \geq 2. \quad (2)$$

The transition rates in the Markov chain model are derived from the experiment of picking a pair of nucleotides at random and counting the number of times this motif is repeated as we scan the sequence to the right. Because of this counting scheme, each perfect repeat of length n in the yeast genome gives rise to one each of length $n-1$, $n-2$, ... as the starting position is varied. Thus, if $p(i)$ is the fraction of repeats of length exactly i, we let $q(i) = \Sigma_{j=i}^{\infty} p(i)$ be the fraction of repeats of length $\geq i$, then we fit the model to $q(i)$.

At this point, we make a minor change in the methods of Kruglyak et al. (1998). In the case of dinucleotide repeats, we fit our model to those with ≥ 6 repeat units rather that using a cutoff of length ≥ 5 as before. To explain the motivation for the change, note that if we assumed for simplicity that all bases were equally likely, then the number of dinucleotide repeats of length 5 we would expect to see in the yeast genome would be $(12 \times 10^6) \times 4^{-8} \times (1 - 4^{-2}) = 171$, while 368, or only slightly more than twice as many, are observed. Rose and Falush (1998) did this calculation more carefully using the observed frequency of various dinucleotides in the yeast genome. Since the observed frequencies of pairs of nucleotides are significantly different from 1/16, this will lead to an even larger predicted value. As Rose and Falush (1998) argue, this suggests that slippage begins to play a significant role in shaping the distribution of dinucleotide repeats only when the number of repeat units is ≥ 6, so in fitting our model, we confine our attention to that range of values.

Having altered the cutoffs, the rest of the fitting was done as in Kruglyak et al. (1998). The constant c only enters into the first equation, so it is irrelevant to the conditional distribution of repeats of length ≥ 2. Dividing the second equation by b shows that when we fit to our data, only the ratio of polymerase slippage to point mutations, b/a, is important. To obtain a direct estimate of b, we use a point mutation rate of 1×10^{-8} per nucleotide per generation as we did in Kruglyak et al. (1998).

To fit the model, we must choose a factor K to convert the probabilities $\pi(i)$ into estimates of the number of repeats found. We then choose K and b/a to minimize $\Sigma_i |K\pi(i) - q(i)|$. Here, we minimize the sum of the absolute values of the differences rather than the sum of the squares of the differences, since we feel that the latter puts too much weight on the first few differences.

Results and Discussion

Slippage Rate Estimates

Table 1 gives the total number of di-, tri-, and tetranucleotide repeats with lengths of at least 2 in the yeast genome. Figures 1 and 2 show the fit of the model for these three cases. As explained in *Materials and Methods*, the model is fit to cumulative microsatellite counts. For example, in the first panel of figure 1, which shows the fit of the model to dinucleotides of length 6 or more, the first bar represents the number of microsatellites that have six or more repeat units, and the second bar represents the number that have seven or more repeat units. The second panel of figure 1 shows the fit of the model to dinucleotides with lengths of at least 5. Note that the first fit is considerably better than the second.

The first row in table 2 gives the estimates from the 1-Mb sample in Kruglyak et al. (1998). The best fit slippage rate from the entire 12-Mb yeast genome is given in the second row. Here, we give the rate of 9.24×10^{-7} per repeat unit per generation, which comes from fitting to dinucleotides with six or more repeat units, rather than the estimate of 7.0×10^{-7}, which comes from fitting to five or more. The third and fourth rows in table 2 give per-repeat and per-locus estimates for di-, tri-, and tetranucleotide repeats in *Drosophila melanogaster* from Schug et al. (1998). These results and those in the fifth row for humans, from Chakraborty et al. (1997), come from the direct estimate of slippage for dinucleotides from Goldstein et al. (1995) combined with an analysis of variance of population variability to infer the ratio of mutation rates of tri- and tetranucleotide rates in comparison with those of dinucleotides. Chakraborty et al. (1997) used the human dinucleotide rate of 5×10^{-4} from Goldstein et al. (1995). Readers who prefer to use the estimate of 2.1×10^{-3} for human dinucleotides from Brinkman et al. (1998) can adjust the predictions accordingly (multiply by 4.2). We have no

Table 1
Microsatellite Counts in the Yeast Genome

Repeat Units	Dinucleotides	Trinucleotides	Tetranucleotides
2	355,945	178,593	38,528
3	24,290	7,325	395
4	2,271	626	50
5	368	188	8
6	141	75	4
7	76	39	0
8	50	32	1
9	38	24	0
10	37	12	0
11	34	7	0
12	15	3	0
13	20	5	1
14	7	1	0
15	8	1	0
16	7	2	0
17	5	0	0
18	4	0	0
19	3	1	0
20	3	1	0
21	0	3	0
24	0	1	0
31	1	0	0
32	1	0	0
36	0	1	0

information about the average lengths of loci studied by Chakraborty et al. (1997), so we cannot compute per-repeat slippage rates for humans.

Our dinucleotide value of 9.24×10^{-7} per repeat unit per generation based on the entire yeast genome is very close to the estimate of 9.3×10^{-7} from the 1-Mb sample in Kruglyak, Durrett, Schug, and Aquadro (KDSA). It is also similar to the value found by Kokoska et al. (1998), who, using techniques described in Wierdl, Dominska, and Petes (1997), inserted a 33-bp dinucleotide repeat into the reading frame of a yeast gene. Alterations in length could then be detected by checking sensitivity to the drug 5-fluoro-oroate (5FOA). With this technique, Kokoska et al. (1998) estimated a mutation rate of 4.8×10^{-6} per locus. Dividing by 15.5 (the number of repeat units minus 1) leads to a per-repeat estimate of 3.09×10^{-7}. This differs from our estimate by a factor of 3. Turning things around, we can say that if our mutation rate estimate were $a = 3.3 \times 10^{-9}$ per nucleotide, then our slippage rate would match their experimental result.

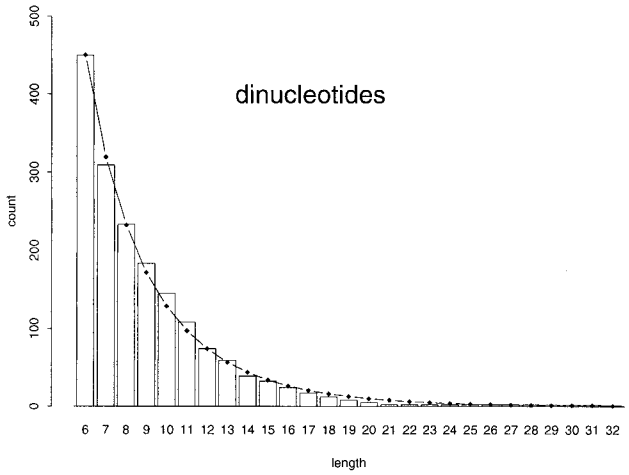
Some experimental mutation rates are much lower than this. Drake et al. (1998) give an estimate of 2.2×10^{-10} per nucleotide per generation. The source of this estimate is an investigation by Drake (1991) of three genes in yeast: URA3 (804 bp), SUP4 (75 bp exon and 14 bp intron), and CAN1 (258 bp of regulatory sequence, 1,773 bp open reading frame) and inferred mutation rates of 2.76×10^{-10}, 7.91×10^{-9}, and 1.73×10^{-10}. Whether the mutation screen used in these experiments accurately reflects the mutation rates in noncoding sequences is unknown. Of course, if a is 2.2×10^{-10}, then our slippage estimate of 9.24×10^{-7} per repeat unit per generation should reduced by a factor of 45 and becomes 2.03×10^{-9}, which is much lower than experimental rates.

Our new estimate for trinucleotide repeats is 2.5 times as large as the KDSA estimate based on a 1-Mb sample. The reason for the difference in the rate estimates can be seen by looking at the data. In the 1-Mb sample, KDSA found 44 trinucleotide repeats, with none of lengths greater than 10. In contrast, the entire yeast genome contains 396 trinucleotide repeats, with 26 of length 11–36. The difference in the sample could be due to bias: KDSA chose the largest contiguous pieces of sequence that could be found in a database in August 1997 and hence might have a greater tendency to come from coding regions. On the other hand, the absence of long repeats in our original 1-Mb sample may be simply due to random chance. Dividing the number found in the entire genome by 12 to account for the difference in sample sizes shows that one expects only about two repeats of lengths greater than 10 in a 1-Mb sample.

Experimental results of Kokoska et al. (1998) found a per-locus slippage rate of 8.4×10^{-6} for a tetranucleotide with 16 repeat units. Dividing by 15 gives a per-repeat-unit rate of 5.6×10^{-7}, versus our estimate of 1.68×10^{-7}. For Drosophila and humans, the per-repeat and per-locus slippage rate estimates were higher for tetranucleotides than for trinucleotides, while our per-repeat estimate for tetranucleotides in yeast was about four times as small as that for trinucleotides. Our estimate was based on only 14 repeats, so it seems that the small number of tetranucleotide repeats in the yeast genome did not allow us to obtain a reasonable estimate for their slippage rates.

Confidence Intervals via the Bootstrap

Having found point estimates for mutation rates, we wanted to compute confidence intervals for these estimates. To do this, we used the bootstrap procedure



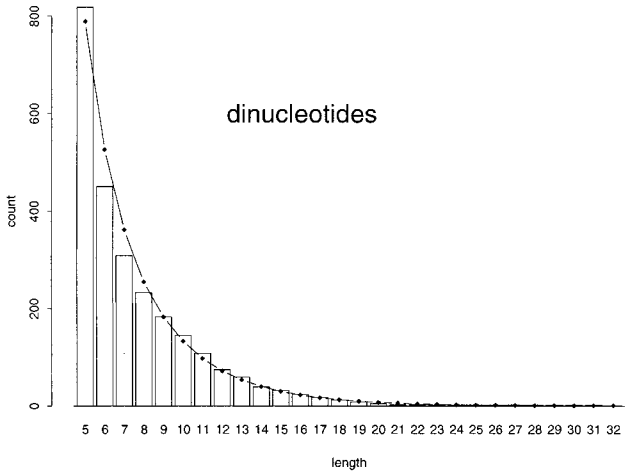
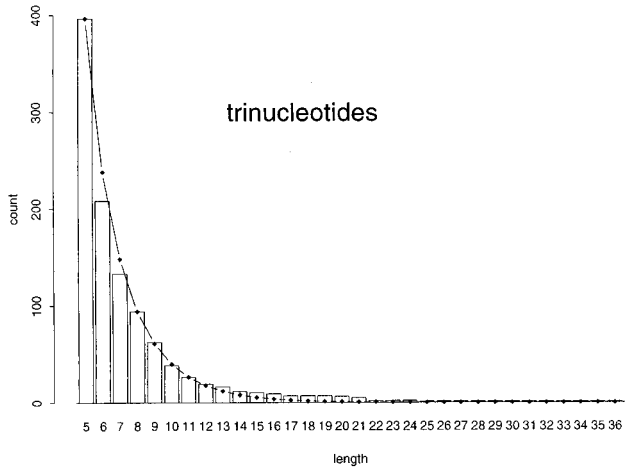


FIG. 1.—Two fits of the Kruglyak et al. (1998) model to all of the dinucleotide repeats in the yeast genome. The upper plot shows the fit to those with lengths of at least 6, and the lower plot shows the fit to those with lengths of at least 5.



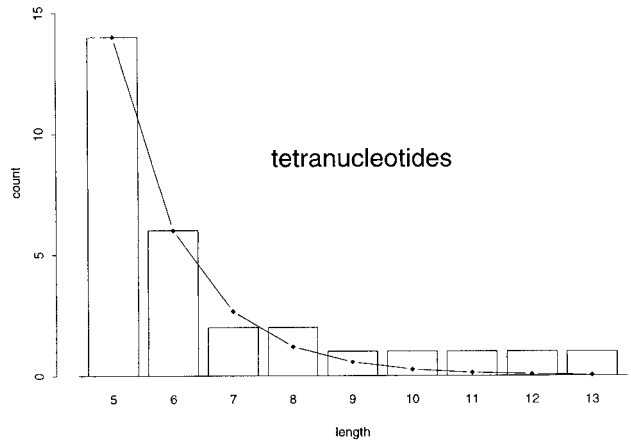


FIG. 2.—Fit of the Kruglyak et al. (1998) model to all of the tri- and tetranucleotide repeats in the yeast genome.

from statistics, which works as follows (see, e.g., Efron 1993). Suppose that we have a data set consisting of N points from some unknown distribution. The empirical distribution of the N sample points provides an estimate for the underlying distribution. If we sample from the data set with replacement, this simulates an independent and identically distributed sample from the empirical distribution. We then compute the estimate that results from this sample of N points. This procedure is repeated M times, for some large value M. The M values of the estimate we observe give an empirical distribution for the estimator. The quantiles of the distribution can then be used to compute the bootstrap confidence interval.

In order to find confidence intervals for the polymerase slippage rate, we applied the procedure to the dinucleotide repeats. The 450 dinucleotides with lengths of six or greater were sampled with replacement. Each resampling step generated a set of 450 repeats. The resampled values were used in the model fit to obtain a slippage rate estimate. The procedure was repeated 10,000 times. The empirical distribution of the dinucle-

Table 2
Slippage Rate Estimates

Organism	Rate Per	Dinucleotides	Trinucleotides	Tetranucleotides	Reference
Yeast.	Repeat	9.3×10^{-7}	2.0×10^{-7}	—	Kruglyak et al. (1998)
Yeast.	Repeat	9.24×10^{-7}	5.02×10^{-7}	1.68×10^{-7}	This study
Yeast.	Repeat	3.09×10^{-7}	—	5.6×10^{-7}	Kokosa et al. (1998)
Drosophila	Repeat	7.7×10^{-7}	2.7×10^{-7}	2.3×10^{-7}	Schug et al. (1998)
Drosophila	Locus	9.3×10^{-6}	1.5×10^{-6}	1.1×10^{-6}	Schug et al. (1998)
Human	Locus	5.0×10^{-4}	2.7×10^{-4}	3.1×10^{-4}	Chakraborty et al. (1997)

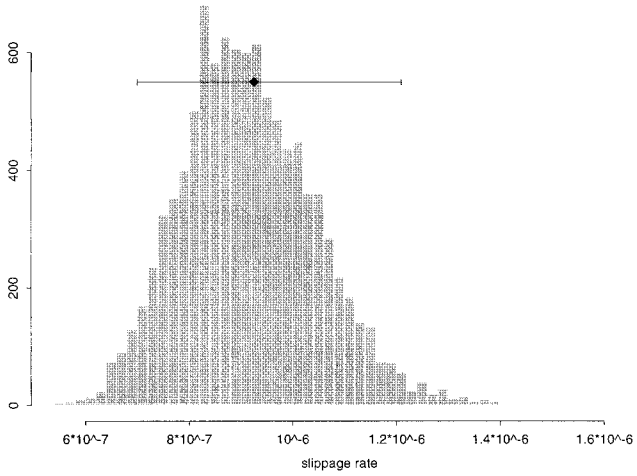


FIG. 3.—Histogram of estimates of per-repeat-unit dinucleotide slippage rates resulting from 10,000 applications of the bootstrap method applied to the entire yeast genome. The diamond and bars indicate mean and 95% confidence interval.

otide slippage rate is shown in figure 3. The diamond shows the slippage rate estimate 9.24×10^{-7} per repeat unit obtained from the entire yeast genome. The horizontal line denotes the 95% bootstrap confidence interval, which is $(7.00 \times 10^{-7}, 12.1 \times 10^{-7})$. The standard deviation of this distribution is $\sigma = 1.31 \times 10^{-7}$. If we let μ denote the mean of the empirical distribution, then the 95% confidence interval is from $\mu - 1.7\sigma$ to $\mu + 2.2\sigma$.

The bootstrap resampling scheme results in each observation being repeated a Poisson number of times with mean 1. Thus, many data points occur more than once, while e^{-1}, or approximately 36.78%, are omitted. An alternate resampling scheme we will call subsampling is more appropriate for comparison with the estimates obtained by KDSA. We divided the yeast genome into 120 pieces of 100 kb each. The microsatellite counts were determined for each piece. A fixed number of the pieces, e.g., 10, were chosen at random without replacement, and the microsatellite counts were combined. The model was fit to these data, and the best fit slippage parameter was determined. The 120 pieces were then resampled and 10 new ones were selected. The procedure was repeated 10,000 times. The empirical distribution was used to compute a 95% confidence interval as described above. We performed the same procedure selecting 20 pieces 10,000 times and then selecting 40 pieces 10,000 times. The results are given in table 3.

Our parameter estimation was based on minimizing the sum of the absolute errors between the model prediction and the data. This procedure fits into the theory estimating equations, which is a generalization of the work of Huber (1967) on maximum-likelihood estimation. A description of this theory can be found in appendix A.3 of Carroll, Ruppert, and Stefanski (1995). The theoretical result which is relevant here is that as $N \to \infty$, the standard deviation of our estimate will be asymptotically $c/\sqrt{N}$. Here, c is a constant that in principle can be inferred from the theory of estimating equations but for which there is not a simple formula.

Table 3
Confidence Intervals for Dinucleotide Per-Repeat-Unit Slippage Rates

Length	Average	SD	95% Confidence Interval
1 Mb . .	11×10^{-7}	5.6×10^{-7}	$(4.4 \times 10^{-7}, 25 \times 10^{-7})$
2 Mb . .	10×10^{-7}	3.1×10^{-7}	$(5.2 \times 10^{-7}, 17 \times 10^{-7})$
4 Mb . .	9.4×10^{-7}	1.9×10^{-7}	$(6.2 \times 10^{-7}, 13 \times 10^{-7})$

If the asymptotic result described in the previous paragraph applies in our situation, then increasing the sample size from 1 to 2 to 4 Mb would decrease the standard deviation by a factor of $\sqrt{2} = 1.414$ each time, while the observed ratios are $5.6/3.1 = 1.80$ and $3.1/1.9 = 1.63$. Since the entire yeast genome is approximately 12 Mb in length, we can extrapolate this trend to predict a standard deviation that is $1/\sqrt{3}$ times the one for 4 Mb of data, or $\sigma = 1.1 \times 10^{-7}$, compared with the direct bootstrap estimate of 1.3×10^{-7}.

Since the bootstrap and subsampling procedures give similar results, we only performed the bootstrap for trinucleotide repeats. The mean of the empirical distribution was 4.8×10^{-7} per repeat unit, and the 95% confidence interval was $(3.8 \times 10^{-7}, 6.0 \times 10^{-7})$. Since the standard deviation was 5.6×10^{-8}, the confidence interval is from $\mu - 1.8\sigma$ to $\mu + 2.1\sigma$. In contrast, KDSA obtained an estimate of 2.0×10^{-7} for the slippage rates of trinucleotide repeats based on a 1-Mb sample from the yeast genome. Although this estimate is low, it is not inconsistent with the new one. The confidence interval based on a 1-Mb sample is approximately $\sqrt{12}$ times as large as the one based on the entire yeast genome, or roughly $(1.3 \times 10^{-7}, 8.9 \times 10^{-7})$.

Repeat Motifs

To refine our understanding of mutational processes of microsatellites, we examined the dependence on the repeated sequence. There are 12 possible dinucleotide repeat motifs that are not mononucleotide repeats, and these fall naturally into four groups:

1. AC, GT; CA, TG
2. AG, CT; GA, TC
3. AT; TA
4. CG; GC.

To explain this classification, we note that an AC repeat gives rise to a GT repeat on the complementary strand (which has the opposite orientation), while an AC repeat contains a CA repeat of almost equal length and gives rise to a TG repeat on the complementary strand.

The first step in understanding the relative abundance of the repeat types is to examine the frequency of the various nucleotides in the 12-Mb yeast genome (A = 0.3090, C = 0.1917, G = 0.1013, T = 0.3078) and of the pairs themselves, which we take in the order above:

Table 4
Observed Numbers of Dinucleotide Repeats

Repeat Units	AT	AC	AG	CG
5	256	57	50	5
6	119	20	2	0
7	60	10	6	0
8	41	7	2	0
9	34	3	1	0
10	32	2	3	0
11	31	3	0	0
12	14	1	0	0
13	17	3	0	0
14	7	0	0	0
15	8	0	0	0
16	7	0	0	0
17	3	2	0	0
18	4	0	0	0
19	2	1	0	0
20	2	1	0	0
31	0	1	0	0
32	0	0	1	0
Total	637	111	65	5

1. 0.0528, 0.0526; 0.0650, 0.0646
2. 0.0585, 0.0582; 0.0624, 0.0623
3. 0.0894; 0.0732
4. 0.0391; 0.0375.

Table 4 gives the distribution of the long dinucleotide repeats (five or more repeats). There are not enough data to estimate slippage rates for GCs or AGs. Using methods described above, we performed a fit to the other two motifs to estimate their slippage rates. For ATs, we got 9.58×10^{-7} and a 95% bootstrap confidence interval of (7×10^{-7}, 13×10^{-7}). For ACs, we got 14.62×10^{-7}. This was 1.5 times as large as the AT rate, but the 95% confidence interval was huge: (3×10^{-7}, 76×10^{-7}). Thus, this yields no information about the relative sizes of the slippage rates.

Looking at table 4, one might naively guess that since AT repeats are more numerous and longer than AC repeats, the slippage rate for ATs must be larger than that for ACs. However, a closer look does not support this idea. The first and simplest observation is that AT repeats are roughly six times as numerous as AC repeats, and if one multiplies the second column in table 4 by 6, the distributions are similar. Indeed, the average length of AT repeats is 7.0, compared with 6.9 for AC repeats. We believe that the larger number of AT repeats is simply due to the fact that the higher AT content allows more AT microsatellites to form. However, we will have to investigate genomes larger than that of yeast to test this hypothesis.

Rose and Falush (1998) argue that a threshold size of five repeat units is needed for significant slippage to occur. Their computations predict that finding AT or TA repeated five times has a probability of 7.8×10^{-6}, while finding AC, GT, CA, or TG repeated five times has a probability of 3.09×10^{-6}. These probability estimates are crude, so it should not be surprising that the 2.5-to-1 ratio of probabilities is somewhat different from the 5.7-to-1 ratio between the observed numbers of AT and AC repeats. To complete the picture, we observe that there is a 3.23×10^{-6} probability of finding AG, CT, GA, or TC repeated five times, while the probability of finding CG or GC repeated five times is 1.65×10^{-7}. Adding up the four frequencies gives 14.3×10^{-6}. This would predict a density of dinucleotide repeats of roughly one per 70 kb, compared with the observed number of one every 14.7 kb.

There are 60 trinucleotide repeat motifs that are not mononucleotides. These divide naturally into 10 groups. To see this, note that a CTG repeat contains TGC and GCT repeats on the same strand and gives rise to CAG, GCA, and AGC repeats on the complementary strand, which is read in the opposite direction. We name each group of six repeats by writing the first in the class in alphabetical order, followed by its reverse complement. Thus, we refer to the example above as AGC/GCT. Table 5 gives the number of repeats of each motif of length 5, 6, 7, 8, 9, 10–17, and >18. The rows are ordered by the total of the lengths of all the microsatellites for a motif, and a line separates the motifs that have more than 50 repeats from those that have 12 or fewer.

For the repeats above the line there are enough data to use our approach to estimate slippage rates. Results are given in the last column. Note that the AAT/ATT family has a rate of 19.96×10^{-7}, compared with rate estimates of 2.14×10^{-7} to 5.00×10^{-7} for the other four repeat motifs for which we are able to obtain estimates. The small amount of data here is unlikely to give informative confidence intervals, so we did not compute them. The reader should note that in the AAT/

Table 5
Observed Numbers of Trinucleotide Repeats

		Motif Length							
Repeat	GC	5	6	7	8	9	10–17	18+	Rate Estimate
AAT/ATT	0	28	14	10	6	5	12	4	19.96×10^{-7}
AAC/GTT....	1	43	17	10	5	6	3	2	3.26×10^{-7}
AAG/CTT....	1	46	18	6	5	6	6		3.36×10^{-7}
ATC/GAT	1	24	9	4	4	4	5	1	5.00×10^{-7}
AGC/GCT....	2	29	9	4	8	2	2		2.14×10^{-7}
ACG/CGT....	2	5	3	1	2	0	1		
AGG/CCT....	2	8	2	1	0	1			
ACT/AGT....	1	3	2	2	1	0	1		
ACC/GGT....	2	2	1	1	1				
CCG/CGG....	3	0							

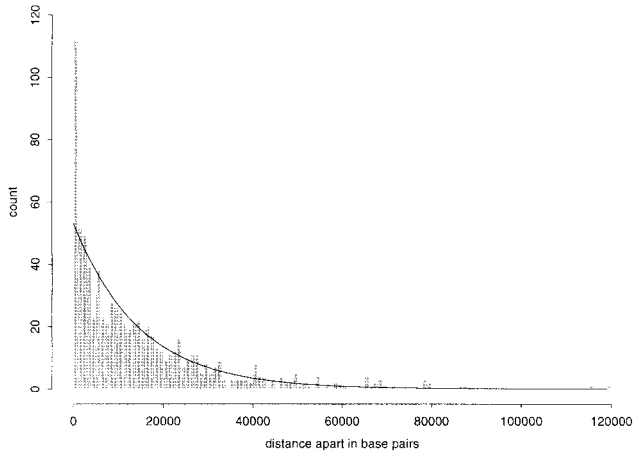


FIG. 4.—Spacing between dinucleotide repeats of length at least five in the yeast genome. The curve represents the exponential distribution with mean 14,700.

ATT family, there are none of the stronger GC bonds, while in three of the other four for which we have estimates, there is only one GC bond. One observation that does not fit this pattern is that in the yeast genome there are very few repeats of the ACT/AGT family but quite a few of the AGC/GCTs.

Young, Sloan, and Van Riper (2000) recently found that mono-, di-, and tetranucleotides are underrepresented and trinucleotides are overrepresented in coding sequence compared with the expectations of random placement. This indicates that selection also plays a role in shaping the distribution of trinucleotide repeats.

Spacings Between Microsatellites

Finally, we investigated the spacings between microsatellites in the yeast genomes. There are two reasons to examine spacing. First, if there is significant variation in the ratio of slippage rate to base pair substitution in the yeast genome, then the locations of the repeats would not follow a Poisson process and the spacings would show a departure from the exponential distribution. Second, our model predicts that the equilibrium distribution of repeat lengths results from a balance between slippage and nucleotide substitutions. Based on this, we would expect that there are many almost-adjacent pairs of perfect repeats that have the same repeat motif.

The average distance between the 818 dinucleotide repeats with lengths of 5 or greater was 12,000,000/818 $\approx$ 14,700 bp. If microsatellites were uniformly distributed across the genome, they would follow a Poisson process, and the distances between them would be exponentially distributed with a mean of 14.7 kb. Figure 4 gives the distribution of the spacing observed between microsatellites in the yeast genome. Since there are 818 dinucleotide repeats of five or more repeat units on 16 chromosomes, we have 802 data points.

Two deviations from the prediction of random spacing are apparent. First, there are two spacings of $>$100 kb. This is 6.8 times the mean of the exponential, so the probability of such a large spacing is $e^{-6.8} = 1/897$. We have 802 observations, so the number of gaps of this length we expect to see has approximately a Poisson distribution with mean 802/897 = 0.894. Recalling the formula for the Poisson distribution with mean λ,

$$P(X = k) = e^{-\lambda}\lambda^k/k!, \quad (3)$$

we see that the probability of two or more gaps this large is $1 - 1.894e^{-0.894} = 0.225$, so this is a fairly common event.

The second noticeable deviation from the exponential is the large number of microsatellites separated by 1,000 bp or less. Under the Poisson model, the spacings between mutations have an exponential distribution with mean 14,700, so the number we expect to find separated by 1,000 bp or less is

$$802 \times (1 - e^{-1{,}000/14{,}7000}) = 802 \times 0.0657 = 52.7, \quad (4)$$

while 112 were found. The actual number observed will have approximately a binomial distribution with 802 trials and a success probability of 0.0657, so the variance is 802(0.0657)(0.9343) = 49.2, and the standard deviation is 7.01. Thus, our observation is 8.5 standard deviations above the mean.

Looking closer at the small spacings, we see that there are 35 separated by $\leq$2 nt. For the exponential distribution, the probability that pair of repeats is separated by 0–2 bp is approximately $3/14{,}700 = 2 \times 10^{-4}$, so the number we expect to observe is Poisson with mean $802 \times 0.0002 = 0.16$. Since we observed 35 pairs, it is clear that the observed proximity of repeats does not occur by random chance.

As mentioned above, our model of microsatellite evolution predicts that this deviation will occur and be due to perfect repeats having been split into two by a point mutation. Table 6 gives the 35 close dinucleotide repeats. The last column gives the sequences which include both microsatellites. The first column gives the chromosome numbers, and second column gives the numbers of base pairs from the start of the chromosome to the start of the sequence given. The 18 patterns that can be explained by a long perfect repeat interrupted by a single substitution are listed first.

The 19th entry in the table is a pair of AT repeats separated by a GA. This configuration can only result from a perfect AT repeat by two substitutions, by an insertion and a deletion, or by a 2-bp insertion. The final 16 entries in the table are instances of two perfect repeats that are adjacent or, in one case, separated by 1 nt. Using equation (3) and discarding the factor $e^{-\lambda} < 1$, we see that the probability of this happening by random chance is smaller than $0.16^{16}/16! < 10^{-26}$, which suggests that there is some interaction between the repeats.

Writing CA as shorthand for a CA or AC repeat and adopting a similar convention for other dinucleotide motifs, it is remarkable that in eight cases we have a CA repeat followed by an AT repeat, and in eight cases we have an AT repeat followed by a TG, which, of course, on the complementary strand is a CA followed by an AT. Given the fact that repeats from the AT and

Table 6
Close Dinucleotides

Chromosome	Start	Motif
III	189,013	$(TA)^{7}TG(TA)^{5}$
IV	169,823	$(TA)^{5}TG(TA)^{10}$
VII	580,383	$(TA)^{5}TT(TA)^{5}$
XI	280,572	$(TA)^{5}GA(TA)^{8}$
XI	464,811	$(TA)^{10}AA(TA)^{7}$
XII	125,399	$(TA)^{10}CA(TA)^{5}$
XIV	605,717	$(TA)^{16}CA(TA)^{5}$
XVI	132,259	$(TA)^{5}CA(TA)^{6}$
IV	226,817	$(AT)^{17}GT(AT)^{6}$
VII	644,767	$(AT)^{7}GT(AT)^{9}$
XVI	139,583	$(AT)^{5}GT(AT)^{7}$
VI	96,023	$(AT)^{6}AG(AT)^{5}$
XII	563,515	$(AT)^{7}AG(AT)^{5}$
XII	873,639	$(AT)^{6}AA(AT)^{7}$
XIV	217,192	$(AT)^{6}AA(AT)^{6}$
XVI	126,531	$(AT)^{6}AC(AT)^{5}$
VII	270,274	$(AT)^{5}AC(AT)^{5}$
XIV	663,595	$(CT)^{7}TT(CT)^{6}$
X	177,609	$(AT)^{6}GA(AT)^{9}$
I	31,485	$(CA)^{11}(AT)^{8}$
IV	147,680	$(CA)^{8}(AT)^{5}$
XVI	631,272	$(CA)^{5}(AT)^{5}$
II	150,300	$(AC)^{6}(AT)^{7}$
II	650,206	$(AC)^{7}(AT)^{7}$
IV	1,229,204	$(AC)^{5}(AT)^{7}$
X	639,680	$(AC)^{19}(AT)^{6}$
XVI	165,813	$(AC)^{5}(AT)^{8}$
II	353,555	$(AT)^{20}(TG)^{13}$
VII	270,307	$(AT)^{5}(TG)^{6}$
XII	725,876	$(AT)^{5}(TG)^{5}$
XV	816,29	$(AT)^{6}(TG)^{9}$
XV	301,294	$(AT)^{6}(TG)^{8}$
IV	226,852	$(TA)^{6}(TG)^{6}$
IX	350,023	$(TA)^{7}(TG)^{7}$
XI	493,324	$(TA)^{6}C(TG)^{6}$

Table 7
Close Trinucleotides

Chromosome	Start	Motif
I	77,497	$(GAA)^{5}GGA(GAA)^{9}$
II	72,341	$(ACA)^{6}GCA(ACA)^{9}$
II	72,360	$(CAA)^{9}CAG(CAA)^{5}$
II	463,978	$(TGT)^{20}TGC(TGT)^{7}$
IV	1,290,973	$(AAT)^{5}AAC(AAT)^{5}$
V	81,195	$(TTA)^{5}TTG(TTA)^{8}$
V	83,183	$(TGT)^{5}TGC(TGT)^{5}$
VII	431,401	$(ATT)^{5}ACT(ATT)^{14}$
IX	105,318	$(ACA)^{5}ACG(ACA)^{5}$
XII	701,196	$(ATT)^{5}ACT(ATT)^{5}$
XII	701,213	$(TAT)^{5}TGT(TAT)^{6}$
XV	110,775	$(ATA)^{5}CTA(ATA)^{9}$
XV	768,667	$(ACG)^{6}ACA(ACG)^{5}$
XVI	536,703	$(TTC)^{16}CTC(TTC)^{7}$
III	281,952	$(CAA)^{6}AGG(CAA)^{5}$
II	780,386	$(CAA)^{5}(CAG)^{8}$
IV	390,421	$(TAT)^{5}(TTG)^{5}$
IV	747,489	$(ACA)^{7}(CAG)^{5}$
V	79,217	$(TCA)^{6}(TCT)^{7}$
VII	394,457	$(GAT)^{6}(GAC)^{6}$
IX	105,336	$(ACA)^{5}(AAT)^{9}$
IX	128,715	$(TCC)^{5}(TCT)^{6}$
IX	169,460	$(TGC)^{5}(TGT)^{7}$
X	149,844	$(TGT)^{5}(TGC)^{6}$
X	188,718	$(CTC)^{5}(TCT)^{6}$
XI	613,404	$(GAG)^{9}(GAA)^{5}$
XIII	169,465	$(TTA)^{5}(TTG)^{7}$
XV	110,791	$(TAA)^{9}(TAG)^{7}$
XVI	521,027	$(AAT)^{8}(TAC)^{8}$
XII	1,011,927	$(TGC)^{6}TCC(TGT)^{6}$

AC/GT families are the most frequent, it is perhaps not surprising that they form the most frequent adjacent pairs. On the other hand, the fact that the AC always precedes the AT in the 16 pairings is quite unusual. The probability that this would happen by chance is $2^{-15} = 3.05 \times 10^{-5}$, where we multiplied by 2 since we would have been equally surprised to always find the AT in front.

Table 7 gives the trinucleotide repeats that are separated by 0–3 bp. In this case, 14 patterns could be explained by a long perfect repeat interrupted by a single substitution, 1 requires three mutations, and 15 are adjacent repeats of different motifs. There are 396 trinucleotide repeats, so the average distance between them is 12,000,000/396 $\approx$ 30,303 bp. For the exponential distribution, the probability that a pair is separated by 0–3 bp is approximately 4/30,303 $\approx 1.33 \times 10^{-4}$, so the number we expect to observe is Poisson with mean $396 \times 0.000133 = 0.0526$. Finding 15 adjacent pairs of trinucleotide repeats is again very unlikely, suggesting some sort of interaction between the repeats. It is interesting to note that in each case after a shift of the reading frame of one triplet, the two repeat motifs differ by one base pair substitution. We would like to express our appreciation to Tom Petes for pointing this out.

In the discussion above, we concentrated on adjacent perfect repeats whose lengths are at least 5. If the KDSA model is correct, then one would expect that in many cases a perfect repeat has been split by a base pair substitution, leaving an adjacent fragment of a repeat. Our model, which keeps track of only the left perfect portion of the repeat after such mutations, does not allow us to make a quantitative prediction about the frequency of occurrence of interrupted repeats. However, we can test the qualitative prediction that they are more frequent than one would expect by chance.

To do this, we scanned through the yeast genome to find the number of times a perfect dinucleotide repeat of length ≥ 5 is followed or preceded by a repeat of length ≥ 2 with the same motif and the separation is 1 or 2 nt. Table 8 gives the results. For repeat lengths with 10 or more interrupted versions, we calculated the fraction of times the repeat was imperfect. Overall, 31.6% of the repeats were imperfect. However, this frequency results from the fact that 26% of repeats of length 5 were imperfect, versus 174/450 = 38.6% of those of length ≥ 6. To show that these frequencies are greater than expected by chance, we can observe that the most frequent nucleotide pair, AT, has a frequency of <0.09. Thus, given four places to look for a matching repeat motif (two places before and two places after), we will find two or more repeat units of the correct motif with probability $<4 \times 0.09^2 = 0.0324$. Thus, the adjacent partial repeats are at least nine times as frequent as we would expect from random chance.

Table 8
Imperfect Dinucleotide Repeats

	Total	2	3	4	5+	% Imperfect
5.....	368	58	16	10	12	26.0
6.....	141	27	11	8	7	37.5
7.....	76	14	5	4	5	36.8
8.....	50	12	2	2	2	36.0
9.....	38	13	2	0	2	44.7
10.....	37	9	3	1	4	45.9
11.....	34	6	5			32.3
12.....	15	1	3			
13.....	20	2	1			
14.....	7	2				
15.....	8	1	1			
16.....	7	1			1	
17.....	5	1	1		1	
18.....	4	1	1			
19.....	3	1				
20.....	1					
31.....	1					
32.....	1					
Total...	818	149	51	25	34	31.6

Conclusions

In this paper, we used the complete 12-Mb sequence of the yeast genome to investigate predictions the Kruglyak et al. (1998) model of microsatellite evolution. We first used the model to estimate slippage rates and the bootstrap from statistics to compute confidence intervals. Our results for di- and trinucleotides agreed well with earlier estimates. However, our result for tetranucleotides was considerably smaller than other estimates. This may be real or simply a result of the limited data for these repeats (only 14 repeats of length ≥ 5). We also examined the dependence of slippage rate on repeat motif. There were enough data to obtain estimates for two dinucleotide repeat families and five trinucleotide families. Although there were not enough data to draw any conclusions with 95% confidence, there were several trends apparent in the data. First, the number of repeat sequences found in the genome was strongly correlated with the number of A's and T's in the repeat sequence. One explanation is that the 62% AT content of the yeast genome provides a larger number of ''seed'' AT repeats of moderate length which grow into microsatellites.

There is not enough information in the sequence of the yeast genome to compare the slippage rates of AT and AC repeats. However, for trinucleotide repeats, the AAT/ATT family had a slippage rate of 19.96×10^{-7}, in contrast to the estimates of 3.26×10^{-7}, 3.36×10^{-7}, 5.00×10^{-7}, and 2.14×10^{-7} for the other four families for which we could obtain estimates: AAC/GTT, AAG/CTT, ATC/GAT, and AGC/GCT.

The recently completed Drosophila genome sequence should provide enough data to obtain a more detailed understanding of the dependence of slippage rates on motifs and variation of rates along and between chromosomes. A preliminary study of this kind has been done by Bachtrog et al. (1999). It is interesting to note that in the parts of the genome they examined, repeats of the CA type are more frequent than those of the AT type. This may reflect the fact that in some of the regions Bachtrog et al. (1999) examined, the AT content was close to 50% (see their table 1). As in our study of yeast, Bachtrog et al. (1999) found that AGs are less frequent than ACs, and ATs and GC repeats are almost nonexistent.

Our final objective was to examine spacings between dinucleotide repeats of length ≥ 5 and to compare them with the patterns that would result if they were scattered randomly through the yeast genome. Two large gaps of more than 100 kb without repeats were found, but probability calculations showed that this was not unusual. We found, as our model predicted, a significant tendency of repeats of the same motif to be adjacent and for perfect repeats to be part of longer imperfect repeats. However, we also found a surprising tendency for repeats of different motifs to cluster. Among dinucleotide repeats, there were 35 adjacent pairs of repeats, 18 of which could be explained by a longer repeat split by a single mutation, but there were also 16 pairs of adjacent repeats with different motifs. The consistent pattern of AC repeats preceding AT repeats suggests that there is some interaction between adjacent repeats, but more research will be needed to determine the mechanism.

Acknowledgments

We thank Tom Petes for a useful discussion of microsatellites in yeast. This work was supported by a postdoctoral fellowship from NSF grant DBI 9504393 to S.K., a National Research Service Award from NIH to M.D.S., NIH grant GM36431 to C.F.A., NIH grant GM36431-14S1 to C.F.A. and R.D., and NSF grant DMS9877066 to R.D.

LITERATURE CITED

ASHLEY, C. T., and S. T. WARREN. 1995. Trinucleotide repeat expansion and human disease. Nat. Genet. **13**:390–391.

ASHWORTH, D., M. BISHOP, K. CAMPBELL et al. (11 co-authors). 1998. DNA microsatellite analysis of Dolly. Nature **394**:329.

BACHTROG, D., S. WEISS, B. ZANGERL, G. BREM, and C. SCHLÖTTERER. 1999. Distribution of dinucleotide microsatellites in the *Drosophila melanogaster* genome. Mol. Biol. Evol. **16**:602–610.

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.

BLOUTIN, J. L., B. A. DOMBROWSKI, S. W. NATH et al. (28 co-authors). 1998. Schizophrenia susceptibility loci on chromosomes 13q32 and 8p21. Nat. Genet. **20**:70–73.

BRINKMAN, B., M. KLINTSCHAR, F. NEUHUBER, J. HÜHNE, and B. ROLF. 1998. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am. J. Hum. Genet. **62**:1408–1413.

CARROLL, R. J., D. RUPPERT, and L. A. STEFANSKI. 1995. Measurement error in nonlinear models. Chapman and Hall, London.

CHAKRABORTY, R., M. KIMMEL, D. N. STIVERS, L. J. DAVISON, and R. DEKA. 1997. Relative mutation rates at di-, tri-, and

tetranucleotide microsatellite loci. Proc. Natl. Acad. Sci. USA **94**:1041–1046.

CLAYTON, R. A., O. WHITE, K. A. KETCHUM, and J. C. VENTER. 1997. The first genome from the third domain of life. Nature **387**:459–462.

DRAKE, J. W. 1991. A constant rate of spontaneous mutation in DNA based microbes. Proc. Natl. Acad. Sci. USA **88**: 7160–7164.

DRAKE, J. W., B. CHARLESWORTH, D. CHARLESWORTH, and J. F. CROW. 1998. Rates of spontaneous mutation. Genetics **148**:1667–1686.

EFRON, B. 1993. An introduction to the bootstrap. Chapman and Hall, N.Y.

FELDMAN, M. W., A. BERGMAN, D. D. POLLOCK, and D. B. GOLDSTEIN. 1997. Microsatellite genetic distances with range constraints: analytic description and problems of estimation. Genetics **145**:207–216.

FIELD, D., and C. WILLS. 1998. Abundant microsatellite polymorphism in *Saccharomyces cerevisiae*, and the different distribution in eight prokaryotes and *S. cerevisiae*, result from strong mutation pressures and a variety of selective forces. Proc. Natl. Acad. Sci. USA **95**:1647–1652.

FOSTER, E. A., M. A. JOBLING, P. G. TAYLOR, P. DONNELLY, P. DE KNIJFF, R. MIEREMET, T. ZERJAL, and C. TYLER-SMITH. 1998. Jefferson fathered slave's last child. Nature **396**:27–28.

GARZA, J. C., M. SLATKIN, and N. B. FREIMER. 1995. Microsatellite allele frequencies in humans and chimpanzees with implications for constraints on allele size. Mol. Biol. Evol. **12**:207–216.

GOLDSTEIN, D. B., and D. D. POLLOCK. 1997. Launching microsatellites: a review of mutation processes and methods of phylogenetic inference. J. Hered. **88**:335–342.

GOLDSTEIN, D. B., A. RUIZ LINARES, L. L. CAVALLI-SZFORZA, and M. W. FELDMAN. 1995. Genetic absolute dating based on microsatellites and the origin of modern humans. Proc. Natl. Acad. Sci. USA **92**:6723–6727.

HENDERSON, S. T., and T. D. PETES. 1992. Instability of simple sequence DNA in *Saccharomyces cerevisiae.* Mol. Cell. Biol. **12**:2749–2757.

HUBER, P. J. 1967. The behavior of maximum likelihood estimates under non-standard conditions. 5th Berkeley Symp. Prob. Stat. **1**:221–223.

KOKOSKA, R. J., L. STEFANOVIC, A. B. BUERMEYER, R. M. LISKAY, and T. D. PETES. 1999. A mutation of the yeast gene encoding PCNA destabilizes both microsatellite and minisatellite DNA sequences. Genetics **151**:511–519.

KOKOSKA, R. J., L. STEFANOVIC, H. T. TRAN, M. A. RESNICK, D. A. GORDNIN, and T. D. PETES. 1998. Destabilization of yeast micro- and minisatellite DNA sequences by mutations affecting a nuclease involved in Ozaki fragment processing (*rad27*) and DNA polymerase δ (*pol3-t*). Mol. Cell. Biol. **18**:2779–2788.

KRUGLYAK, S., R. DURRETT, M. SCHUG, and C. F. AQUADRO. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc. Natl. Acad. Sci. USA **95**:10774–10778.

MEIN, C. A., L. ESPOSITO, M. G. DUNN et al. (15 co-authors). 1998. A search for type 1 diabetes susceptibility genes in families from the United Kingdom. Nat. Genet. **19**:297–300.

MOORE, H., P. W. GREENWELL, C. P. LIU, N. ARNHEIM, and T. D. PETES. 1999. Triplet repeats form secondary structures that escape DNA repair in yeast. Proc. Natl. Acad. Sci. USA **96**:1504–1509.

OHTA, T., and M. KIMURA. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22**:201–204.

PETES, T. D., P. W. GREENWELL, and M. DOMINSKA. 1997. Stabilization of microsatellite sequences by variant repeats in the yeast *Saccharomyces cerevsiae.* Genetics **146**:491–498.

PRITCHARD, J. K., M. T. SEIELSTAD, A. PEREZ-LEZAUN, and M. W. FELDMAN. 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol. Biol. Evol. **16**:1791–1798.

REICH, D. E., and D. B. GOLDSTEIN. 1998. Genetic evidence for a Paleolithic human population expansion in Africa. Proc. Natl. Acad. Sci. USA **95**:8119–8123.

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

RUIZ-LINARES, A., D. ORTIZ-BARRIENTOS, M. FIGUEROU et al. (12 co-authors). 1999. Microsatellites provide evidence for Y chromosome diversity among founders of the New World. Proc. Natl. Acad. Sci. USA **96**:6312–6317.

SCHUG, M. D., C. M. HUTTER, K. A. WETTERSTRAND, M. S. GAUDETTE, T. F. C. MACKAY, and C. F. AQUADRO. 1998. The mutation rates of di-, tri-, and tetranucleotide repeats in *Drosophila melanogaster*. Mol. Biol. Evol. **15**:1751–1760.

SIA, E. A., R. J. KOKOSKA, M. DOMINSKA, P. GREENWELL, and T. D. PETES. 1997. Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. Mol. Cell. Biol. **17**:2851–2858.

STEPHAN, W., and Y. KIM. 1998. Persistence of microsatellite arrays in finite populations. Mol. Biol. Evol. **15**:1332–1336.

UNDERHILL, P. A., L. JIN, R. ZEMANS, P. J. OFENER, and L. L. CAVALLI-SFORZA. 1996. A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. Proc. Nat. Acad. Sci. USA **93**:196–200.

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

YOUNG, E. T., J. S. SLOAN, and K. VAN RIPER. 2000. Trinucleotide repeats are clustered in regulatory genes in *Saccharomyces cerevisiae.* Genetics **154**:1053–1068.

ZHIVOTOVSKY, L. A., M. W. FELDMAN, and S. A. GRISHECKIN. 1997. Biased mutations and microsatellite variation. Mol. Biol. Evol. **14**:926–933.

WOLFGANG STEPHAN, reviewing editor

Accepted April 13, 2000