DOI: 10.1534/genetics.103.022665

Microsatellite Mutation Models: Insights From a Comparison of Humans and Chimpanzees

Raazesh Sainudiin,*,[1] Richard T. Durrett,[†] Charles F. Aquadro[‡] and Rasmus Nielsen[§]

**Department of Statistical Science, Cornell University, Ithaca, New York 14853, [†]Department of Mathematics, Cornell University, Ithaca, New York 14853, [‡]Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853 and [§]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853*

Manuscript received September 27, 2003
Accepted for publication June 3, 2004

ABSTRACT

Using genomic data from homologous microsatellite loci of pure AC repeats in humans and chimpanzees, several models of microsatellite evolution are tested and compared using likelihood-ratio tests and the Akaike information criterion. A proportional-rate, linear-biased, one-phase model emerges as the best model. A focal length toward which the mutational and/or substitutional process is linearly biased is a crucial feature of microsatellite evolution. We find that two-phase models do not lead to a significantly better fit than their one-phase counterparts. The performance of models based on the fit of their stationary distributions to the empirical distribution of microsatellite lengths in the human genome is consistent with that based on the human-chimp comparison. Microsatellites interrupted by even a single point mutation exhibit a twofold decrease in their mutation rate when compared to pure AC repeats. In general, models that allow chimps to have a larger per-repeat unit slippage rate and/or a shorter focal length compared to humans give a better fit to the human-chimp data as well as the human genomic data.

MICROSATELLITES are tandem repeats of short DNA motifs between 2 and 5 bp. Their high length variability, genome-wide distribution, and abundance make them useful for evolutionary and population genetic inference in areas as diverse as molecular forensics, parentage testing, molecular anthropology, and conservation genetics and in studies of human evolutionary history (*e.g.*, Jarne and Lagoda 1996; Ellegren 2000b). Population genetic inferences may be sensitive to the assumed model of microsatellite evolution. Therefore, much focus has centered on the development of biologically realistic models. However, there has been relatively little focus on testing and comparing these models using real data.

The simplest popular model of microsatellite evolution is the classical stepwise mutation model (SMM) of Ohta and Kimura (1973) in which, upon a mutation, 1 repeat unit is either gained, resulting in an expansion, or lost, resulting in a contraction. However, mutations have been observed to change the repeat length by >1 unit (Xu *et al.* 2000; Harr *et al.* 2002; Huang *et al.* 2002). The two-phase model (TPM) of DiRienzo *et al.* (1994) addresses this by allowing mutations of 1 repeat unit (one-phase) with probability p and mutations of ≥ 1 unit(s) (two-phase) with probability $1 - p$, while the distribution of the lengths of multiunit mutations is geometric. In a simpler two-phase model of Fu and Chakraborty (1998) mutations of length ≥ 1 are geometrically distributed. Under the SMM and the TPM, a microsatellite is assumed to mutate at a constant rate, irrespective of its repeat length. Moreover, under these models there is no bias toward an expansion or a contraction, and thus the microsatellites are expected to grow or contract unconstrained over time. While constraining the range of repeat lengths through a model with reflecting boundaries (Nauta and Weissing 1996; Feldman *et al.* 1997) can circumvent this problem of unbounded growth, the biological reality of such a defined boundary is unclear.

Evidence for length-dependent effects on mutation rate (Ellegren 2000a), whereby longer microsatellites mutate more often than shorter ones, and the presence of point mutations in some repeats make the proportional slippage (PS) model of Kruglyak *et al.* (1998) and its extensions by Calabrese *et al.* (2001) attractive. In the symmetric PS model, an equilibrium distribution of repeat lengths exists through a balance between slippage events and point mutations (Kruglyak *et al.* 1998). Various mutational biases have been proposed, including an upward bias favoring expansions in humans (Amos *et al.* 1996) and barn swallows (Primmer *et al.* 1996), an excess of contractions in long microsatellites of yeast (Wierdl *et al.* 1997) and fruit fly (Harr and Schlötterer 2000), and the rate of contractions increasing exponentially with repeat length in humans (Xu *et al.* 2000). In the presence of a linear bias toward a target or focal length, as proposed by Garza *et al.* (1995) and

[1]*Corresponding author:* Department of Statistical Science, 301 Malott Hall, Cornell University, Ithaca, NY 14853.
E-mail: rs228@cornell.edu

further elaborated by Zhivotovsky *et al.* (1997), microsatellites below the focal length tend to expand, and those above it tend to contract. Other models emphasize mutational bias by allowing the probability of an expansion upon mutation to be independent of repeat length (Walsh 1987; Tachida and Iizuka 1992; Fu and Chakraborty 1998) or be dependent on it exponentially (Calabrese and Durrett 2003; Whittaker *et al.* 2003), quadratically, or piecewise linearly (Calabrese and Durrett 2003).

Thus, broadly speaking, there are at least three sets of qualitatively contrasting features in the existing models of microsatellite evolution. The first is one-phase *vs.* two-phase mutations. The second is mutation rate proportionality (the proportional dependence of mutation rate on repeat length) *vs.* rate equality. The final set of contrasting features is the presence or absence of mutational bias, whereby the probability of expansion upon mutation may depend on the repeat length of the mutating microsatellite in one form or another. We address only constant bias, where the probability that a mutation results in an expansion is constant for all alleles, and linear bias, where this probability varies linearly with repeat length.

We test the relative significance of these contrasting features, as embodied by variants of some popular models and their hybrids, with likelihood-ratio tests (LRTs) and the Akaike information criterion (AIC), using data from dinucleotide loci homologous between humans (*Homo sapiens*) and chimps (*Pan troglodytes*). Complications to the mutational process from variation in repeat motif as well as interruptions by point mutations are also explored. We address the question of longer repeat length in humans compared to chimps through a lineage-specific analysis.

MODELS

For mathematical convenience, most models of microsatellite evolution assume that the number of repeat units can be any positive integer. Our numerical computations are done with finite matrices, so we study these Markov chains on a truncated state space $\mathbb{S} = \{\kappa, \kappa + 1, \ldots, \Omega\}$. Truncation of the state space from above is biologically reasonable, as microsatellites are rarely longer than Ω (a few tens of repeat units), and that from below ensures that κ is greater than the threshold repeat length above which mutations in length that are characteristic of microsatellites occur (Rose and Falush 1998).

The data D for our study are a $2 \times N$ matrix of microsatellite allele lengths from N loci homologous in humans and chimps. We model the distribution of D by superimposing three Markov chains, $\mathbf{X}^{(a)}$, $\mathbf{X}^{(c)}$, and $\mathbf{X}^{(h)}$, on the ancestral, chimp, and human branches, respectively, of the two-taxa tree $\boldsymbol{\tau}$, as shown in Figure 1. In $\boldsymbol{\tau}$, each of the two terminal branch lengths, λ_c and λ_h, represents the product of mutation rate at allele κ and number of generations along the chimp and human lineages, respectively. We assume that the time to coalescence for a pair of homologous alleles, within the ancestral population, is negligible relative to the time since the human-chimp speciation.

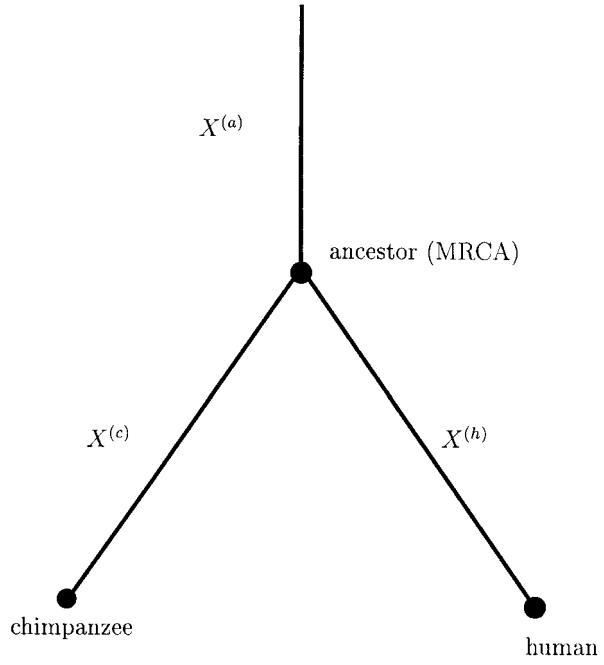

Figure 1.—Markov chains on the branch leading to the ancestor ($X^{(a)}$), chimpanzee ($X^{(c)}$), and human ($X^{(h)}$).

Let $\Theta^{(a)}$, $\Theta^{(c)}$, and $\Theta^{(h)}$ be parameters of the Markov chains $\mathbf{X}^{(a)}$, $\mathbf{X}^{(c)}$, and $\mathbf{X}^{(h)}$, with transition probability matrices $\mathbf{P}^{(a)}$, $\mathbf{P}^{(c)}$, and $\mathbf{P}^{(h)}$, respectively. For an ergodic continuous-time Markov chain, its transition probability matrix $\mathbf{P}(\lambda) := (P_{ij})_{i,j=\kappa}^{\Omega} = \exp\{\mathbf{Q}\lambda\}$, where $\mathbf{Q} := (q_{ij})_{i,j=\kappa}^{\Omega}$ is its infinitesimal generator or rate matrix. The stationary distribution of such a Markov chain, denoted by $\boldsymbol{\pi} = (\pi_\kappa, \pi_{\kappa+1}, \ldots, \pi_\Omega)$, is the unique probability distribution on $\mathbb{S}$ satisfying the matrix equation $\boldsymbol{\pi}\mathbf{Q} = \mathbf{0} = (0, 0, \ldots, 0)$ (see, *e.g.*, Brémaud 1999). Interest in $\mathbf{P}(\lambda)$ and $\boldsymbol{\pi}$ arises because they determine the likelihood function L_i in Equation 1.

Let $\boldsymbol{\pi}^{(a)}$ be the stationary distribution of the ancestral chain. Let $\boldsymbol{\Theta} := (\Theta^{(a)}, \Theta^{(c)}, \Theta^{(h)})$ and $\boldsymbol{\lambda} := (\lambda_c, \lambda_h)$. The likelihood, given homologous allele length data $D_i = (C_i, H_i)$ at locus i, is

$$L_i(\boldsymbol{\Theta}, \boldsymbol{\lambda} | D_i) := \sum_{j \in \mathbb{S}} \pi_j^{(a)} P_{j,C_i}^{(c)}(\lambda_c) P_{j,H_i}^{(h)}(\lambda_h). \qquad (1)$$

Since we do not know the ancestral state, the likelihood may be thought of as a weighted sum over all possible ancestral states, where the weights come from the stationary distribution of the ancestral chain. Assuming independence among the N loci, the likelihood, given the total data D, is obtained by multiplication.

$$L(\boldsymbol{\Theta}, \boldsymbol{\lambda}|D) := \prod_{i=1}^{N} L_i(\boldsymbol{\Theta}, \boldsymbol{\lambda}|D_i). \quad (2)$$

A general model within which all other models of interest are nested is defined below. We start by defining $\gamma(m, i, j)$, a truncated geometric distribution with success probability m, given by

$$\gamma(m, i, j) = \begin{cases} \dfrac{m(1-m)^{|i-j|-1}}{1-(1-m)^{\Omega-i}}, & \kappa \le i < j \le \Omega \\ \dfrac{m(1-m)^{|i-j|-1}}{1-(1-m)^{i-\kappa}}, & \kappa \le j < i \le \Omega. \end{cases}$$

Observe that for every allele i, $\sum_{j=i+1}^{\Omega} \gamma(m, i, j) = \sum_{j=\kappa}^{i-1} \gamma(m, i, j) = 1$.

A continuous-time Markov chain $\mathbf{X}$ on $\mathbb{S}$ is defined with an infinitesimal generator $\mathbf{Q}$ given by

$$q_{i,j} = \begin{cases} \beta(i, s)\alpha(u, v, i)(p + (1-p)\gamma(m, i, j)), & i = j-1 \\ \beta(i, s)\alpha(u, v, i)(1-p)\gamma(m, i, j), & i < j-1 \\ \beta(i, s)(1-\alpha(u, v, i))(p + (1-p)\gamma(m, i, j)), & i = j+1 \\ \beta(i, s)(1-\alpha(u, v, i))(1-p)\gamma(m, i, j), & i > j+1 \\ -\sum_{i \ne j} q_{i,j}, & i = j \end{cases} \quad (3)$$

where $\beta(i, s)$ is the mutation rate of allele i and $\alpha(u, v, i)$ is the probability that a mutation results in an expansion. When $p = 1$, any microsatellite allele mutates (*i.e.,* expands or contracts) by only 1 unit of repeat length, but when $p < 1$, it mutates by 1 or more unit(s) of length with probability $1 - p$ and by 1 unit of length with probability p. Given that an allele i undergoes a multistep mutation, the probability of expanding or contracting by k units is given by $\gamma(m, i, j)$. The functions α and β are defined as

$$\beta(i, s) = \mu(1 + (i - \kappa)s),$$

$$\alpha(u, v, i) = \max\{0, \min\{1, u - v(i - \kappa)\}\}.$$

The proportional dependence of mutation rate on repeat length is captured by the proportional rate parameter $s \in (-1/(\Omega - \kappa + 1), \infty)$ in $\beta(i, s)$. When $s = 0$, alleles of all lengths have the same mutation rate $\mu \in (0, \infty)$ of allele κ. Thus, s represents the strength of length dependence of the mutation rate. Observe that $1/\beta(i, s)$ is the average amount of time spent by a microsatellite locus in an allele of repeat length i (mean holding time in allele i).

In the function $\alpha(u, v, i)$, the constant bias parameter is $u \in [0, 1]$ and the linear bias parameter is $v \in (-\infty, +\infty)$. If $u = 0.5$ and $v = 0$, we have a symmetric unbiased mutational process in which the probability that a mutation is an expansion or a contraction is equal. If $v = 0$, then $\alpha(u, v, i) = u \in [0, 1]$ for any allele i, and we have a model with constant mutational bias. Furthermore, we have a linear mutational bias when $v \ne 0$. If $0.5 < u < 1$ and $(u - 0.5)/(\Omega - \kappa) < v < \infty$, we have a *focal* length $f = ((u - 0.5)/v) + \kappa$, where the probability of contraction equals that of expansion ($\alpha(u, v, f) = 0.5$), and toward which the mutational process is linearly biased. So, when $i < f$, the mutational bias is upward, toward f, since $\alpha(u, v, i) > 0.5$, and when $i > f$, the bias is downward, toward f, as $\alpha(u, v, i) < 0.5$.

The probability of a transition from allele i to j in t generations, with a mutation rate μ at κ, is given by $P_{i,j}(\lambda)$, where $\lambda = \mu t$. Note that for large values of λ, $P_{i,j}(\lambda)$ is approximately equal to π_j, the stationary distribution. Large values of λ model heavily saturated data obtained from a pair of highly diverged species whose repeat length distributions are approximately independent of each other and close to stationarity. Therefore, the extent of saturation in the observed data is reflected by the magnitude of the estimate of λ. The structure of the various submodels within the general model is described by the tree in Figure 2. The last column shows some of the common models in the literature that are closely related to some of these submodels. The model parameters that are fixed for a set of submodels are written above the branches leading to them.

The equal-rate unbiased one-phase model (EU1) is a truncated version of the SMM of Ohta and Kimura (1973). The equal-rate, constant-biased, one-phase model (EC1) embodies constant bias toward expansion in the mutation process by constraining $\alpha(u, 0, i) = u$ for any allele i. Observe that u does not vary with allele length in the EC1 model, as v, the linear bias parameter, is set at 0. Freeing v allows a linear mutational bias as embodied by the equal-rate, linear-biased, one-phase model (EL1), with a mutational bias toward the focal length f, akin in spirit to the mutation scheme introduced by Garza *et al.* (1995). Note that EL1 is related to the simplest version of the PLBias model of Calabrese and Durrett (2003). The equal-rate, one-phase models, EU1, EC1, and EL1, have s set to 0, making the mutation rate equal for all alleles ($\beta(i, s) = \mu$), unlike their proportional-rate, one-phase cousins, PU1, PC1, and PL1, respectively, which allow s to take values in $(-1/(\Omega - \kappa + 1), \infty)$. The PU1 model is related to PS\0M, a proportional slippage model without point mutations proposed by Calabrese *et al.* (2001). The PC1 model is similar to the models proposed by Walsh (1987) and Tachida and Iizuka (1992).

In all six models discussed so far, alleles mutate by only 1 unit of repeat length, since p and m are set at 1. When $p < 1$ and $m < 1$, we have an equal-rate, unbiased, two-phase model EU2*, the truncated version of the TPM of DiRienzo *et al.* (1994), which allows both single-step and multistep mutations instantaneously. However, in this two-phase model, the parameters p and m are nonidentifiable at the boundaries of interest ($p = 1$ or $m = 1$). We rectify this by a single-valued transformation prior to inference.

It is possible to obtain the six one-phase models from Equation 3 by setting p at 0 to allow mutations of length ≥ 1 and setting m at 1 to force the geometric distribution to put all its mass on one-step mutations. When $m < 1$, we have their two-phase cousins in the spirit of Fu and

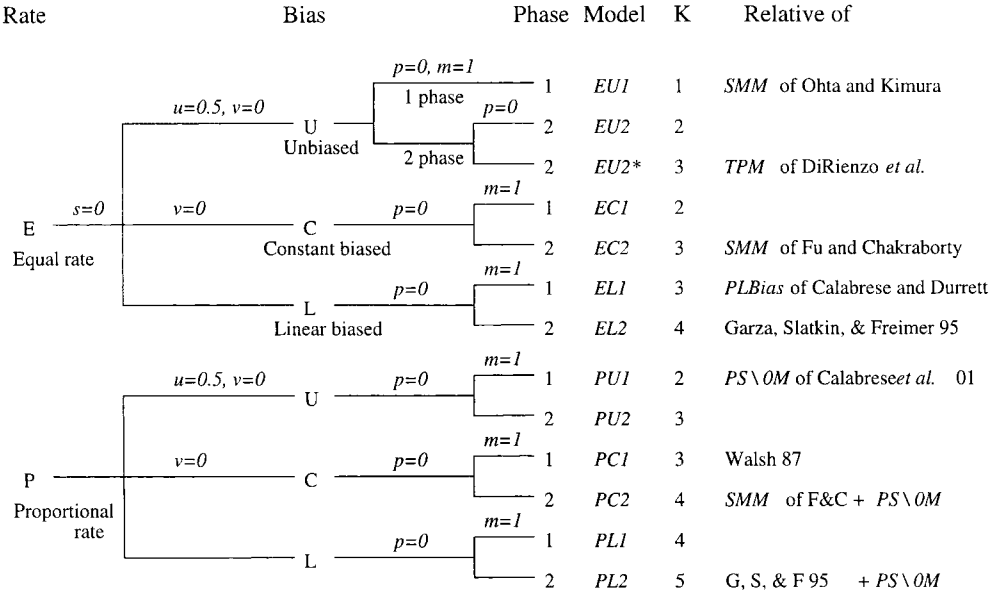

FIGURE 2.—Model description. *K* denotes the number of free parameters. The fixed parameter(s) for each set of models is/are shown above the branch leading to it.

CHAKRABORTY (1998), namely EU2, EC2, EL2, PU2, PC2, and PL2. These models capture the qualitative features of one-phase and two-phase in a simpler and identifiable manner. Observe that the equal-rate linear-biased two-phase model (EL2) is not unlike a model inspired by the mutation scheme of GARZA *et al.* (1995; see also ZHIVOTOVSKY *et al.* 1997), in which the mutation rate is independent of allele length and the bias is a linear function of repeat length with an attracting focal length. The most general model PL2 is related to a hybrid of PS\0M and the model due to GARZA *et al.* (1995).

DATA AND METHODS

To find the largest number of homologous loci in the pair of primates, while minimizing ascertainment bias and sequencing error, we first obtained 21.4 Mbp of the *P. troglodytes* (chimp) sequences in HTGS (high-throughput genomic sequence) (OUELLETTE and BOGUSKI 1997) phase 3, available by March 4, 2003, through the *Entrez* retrieval system of NCBI (http://www.ncbi.nlm.nih.gov/entrez/). The sequences in HTGS phases 0, 1, and 2 were excluded to minimize sequencing error. For all analyses in this study we set the lower bound $\kappa = 10$. Chimp microsatellites of dinucleotide motifs with repeat length ≥ 10 were obtained. To assure some level of independence, all microsatellites within 200 bp of another were discarded.

Each selected chimp microsatellite, with 200 bp of flanking sequence upstream and downstream, was used to perform an extremely stringent (*E*-value $\leq 1 \times 10^{-100}$) unfiltered BLAST search against the human contigs downloaded from the August 23, 2002, NCBI release at ftp://ncbi.nlm.nih.gov/genomes/H_sapiens/, using formatdb and blastall (2.2.3 release) of the NCBI Toolkit in ftp://ftp.ncbi.nih.gov/toolbox/ncbi_tools/. Thus we obtained 644 candidate microsatellite loci homologous between the two primates.

Each such microsatellite locus was retained if it had a flanking sequence of length ≥ 200 bp on at least one side of the dinucleotide repeat in both species and a flanking sequence of length ≥ 50 bp on the other side in both species. A compound repeat is defined to have more than one motif, each of repeat length ≥ 10, within a 50-bp radius. Thirty percent of loci contained compound repeats in at least one of the homologs and were excluded from further analysis. Finally, those loci whose simple repeats in at least one species were interrupted by two or more point mutations were omitted. Thus 383 candidate loci were obtained. About 70% of these loci occurred in human chromosome 7. Fifteen percent of these 383 loci were omitted as their human homologs were ≤ 9 units in repeat length. Among the remaining 321 loci 78% were AC repeats (namely, AC, CA, TG, and GT repeats), 13% were AT repeats (namely, AT and TA repeats), and 9% were AG repeats (namely, AG, GA, TC, and CT repeats). There were no CG repeats (namely, CG and GC repeats).

Among these 321 loci, 18% contained homologous pairs of once-interrupted dinucleotide repeats, which have exactly one point mutation interrupting an otherwise pure stretch of the repeat in either or both species. We count the repeat length of a once-interrupted AC repeat (iAC repeat), ignoring the interruption. For instance, the iAC repeat "ACACATACAC" is taken to be of length 5. The common practice in the literature of directly extrapolating the repeat length of a microsatellite from its PCR fragment length is the motivation behind such a characterization of repeat length for an interrupted microsatellite.

Thus we found 321 homologous pairs of simple dinucleotide repeats with at most one interruption, of which 264 were uninterrupted or pure dinucleotide repeats

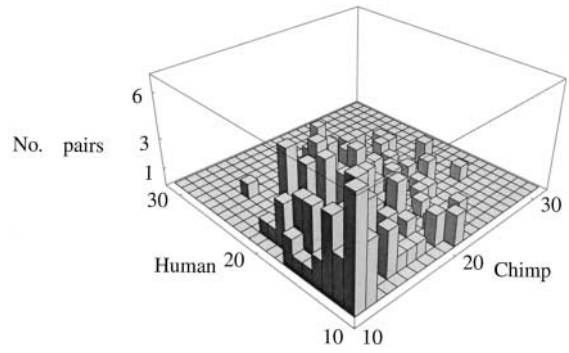

FIGURE 3.—Counts of the pairs of homologous microsatellites (pure AC repeats) between human and chimp (D_{AC}).

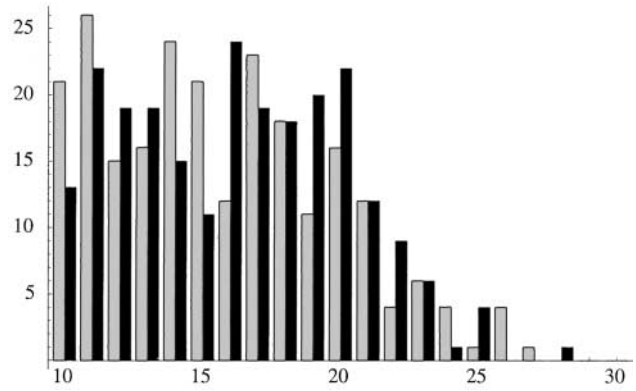

FIGURE 4.—Counts of the chimp (shaded bars) and human (solid bars) microsatellites of pure AC repeats (D_{AC}).

and 235 were pure AC repeats. This constitutes our basic human-chimp data set. The empirical joint and marginal distributions of homologous pure AC-repeat data are shown in Figures 3 and 4, respectively. We also obtained the human genomic data of perfect (devoid of interruptions) and isolated (at least 50 bp from the nearest dinucleotide microsatellite of length >4 repeat units) AC repeats as described by CALABRESE and DURRETT (2003) for comparative purposes.

To maximize the likelihood L, we transformed the constrained parameter space to an unconstrained one and performed an unconstrained optimization using the function *Findminimum* of *Mathematica* (WOLFRAM 1999). We explored most of the support of the parameter space by partitioning it into small hypercubes and using their midpoints as initial conditions to find local optima.

Stationary distribution of one-phase models: Observe that all the one-phase models including PL1 are special cases of the general birth-death chain with birth and death rates b_i and d_i representing the rate of expansion and contraction, respectively, of allele i by 1 repeat unit. Using the convention $\prod_{j=\kappa}^{\kappa-1}(\cdot) = 1$, the stationary distribution π_i, up to a normalizing factor, is given by

$$\pi_i \propto \prod_{j=\kappa}^{i-1} \frac{b_j}{d_{j+1}}.$$

For example, for the PL1 model with birth rate $\alpha(u, v, i)\,\beta(i, s)$ and death rate $(1 - \alpha(u, v, i))\,\beta(i, s)$,

$$\begin{aligned}\pi_i &\propto \prod_{j=\kappa}^{i-1} \frac{\alpha(u, v, j)\beta(j, s)}{(1 - \alpha(u, v, j+1))\beta(j+1, s)} \\ &\propto \prod_{j=\kappa}^{i-1} \frac{\alpha(u, v, j)}{(1 - \alpha(u, v, j+1))} \prod_{j=\kappa}^{i-1} \frac{\beta(j, s)}{\beta(j+1, s)} \\ &\propto \frac{1}{1 + (i - \kappa)s} \prod_{j=\kappa}^{i-1} \frac{\alpha(u, v, j)}{(1 - \alpha(u, v, j+1))}. \end{aligned} \quad (4)$$

Repeat-specific models: The presence or absence of any significant difference between the mutational mechanisms of two distinct types of dinucleotide repeats, for example, pure *vs.* interrupted repeats, or different motifs, can be investigated. The distribution of D^{I}, the data of type I, is modeled by superimposing a Markov chain model $\mathbf{X}^{I}$ with parameters Θ^{I} on the three branches of τ with terminal branches of equal length λ^{I}. D^{II}, the data of type II, is modeled in a similar manner, by $\mathbf{X}^{II}$ with its respective parameters and branch length. Thus, akin to Equation 2, our likelihood function for the data (D^{I}, D^{II}), where D^{I} is a $2 \times N^{I}$ matrix, and D^{II} is a $2 \times N^{II}$ matrix, is

$$L(\Theta^{I}, \Theta^{II}, \lambda^{I}, \lambda^{II}|(D^{I}, D^{II})) := \prod_{i=1}^{N^{I}} L_i(\Theta^{I}, \lambda^{I}|D_i^{I}) \prod_{i=1}^{N^{II}} L_i(\Theta^{II}, \lambda^{II}|D_i^{II}). \quad (5)$$

Likelihood-ratio test: Suppose $\theta = (\theta_r, \theta_s) \in (\Theta_r, \Theta_s)$ is a vector of $r + s$ parameters, where $r \geq 1$ and $s \geq 0$, and we are interested in testing the null hypothesis, H_0: $\theta_r = \theta_{r0}$, against the alternative hypothesis, H_1: $\theta_r \neq \theta_{r0}$. The likelihood-ratio test statistic (LRTS) given by

$$-2 \log \frac{\sup_{\theta_s \in \Theta_s} L(\theta_{r0}, \theta_s|D)}{\sup_{\theta \in (\theta_r, \theta_s)} L(\theta_r, \theta_s|D)} \quad (6)$$

is asymptotically χ_r^2 distributed under the null hypothesis, where the degrees of freedom r is the difference in the number of free parameters between the two hypotheses, under standard conditions (WILKS 1938). The asymptotic distribution of the LRTS when the parameter has a boundary value is obtained from SELF and LIANG (1987).

Model selection: Given an *a priori* set of candidate models, they can be ranked from the best to the worst, in an information-theoretic paradigm through a second-order Akaike information criterion (AIC_c). This ranking can help distinguish models that are nearly equally good fits *vs.* those that are poor explanations for the given data D of sample size N. The best candidate model with a total of K parameters in $(\boldsymbol{\Theta}, \boldsymbol{\lambda})$ is the one that minimizes the quantity

$$AIC_c := -2 \log L(\boldsymbol{\Theta}, \boldsymbol{\lambda}|D) + 2\left(K + \frac{K(K+1)}{N - K - 1}\right). \quad (7)$$

We use AIC_c (SUGIURA 1978; HURVICH and TSAI 1989), the second-order estimator of the Kullback-Liebeler information, instead of the first-order estimator AIC, be-

TABLE 1

Parameter estimation, maximum likelihood, and model ranking using species-pair data from 235 loci of AC repeats

		MLEs of parameters				log *L*:	AIC_c + 2491		
Models	*u*	*v*	*m*	*s*	λ	Ω = 40	Ω = 40	Ω = 60	Ω = 100
PL2	0.82	0.039	0.55	0.76	0.56	−1240.21	0	0	0
PL1	0.62	0.015	*1.00*	0.88	2.14	−1241.48	0.5	0.4	0.4
EL2	0.68	0.037	0.43	*0.00*	1.72	−1247.94	13	13	13
EL1	0.54	0.0095	*1.00*	*0.00*	12.26	−1250.40	16	16	16
EC1	0.47	*0.00*	*1.00*	*0.00*	11.09	−1294.54	102	108	108
PC1	0.47	*0.00*	*1.00*	−0.0048	10.93	−1294.52	104	110	110
EC2	0.47	*0.00*	0.99	*0.00*	11.09	−1294.54	105	110	110
PC2	0.47	*0.00*	0.99	−0.0048	10.93	−1294.52	107	112	112
PU1	*0.50*	*0.00*	*1.00*	0.28	3.68	−1342.48	198	276	347
PU2	*0.50*	*0.00*	0.99	0.28	3.68	−1342.48	200	278	349
EU1	*0.50*	*0.00*	*1.00*	*0.00*	10.33	−1432.35	376	610	882
EU2	*0.50*	*0.00*	0.94	*0.00*	8.63	−1432.31	378	609	877

The parameters that are fixed for a given submodel are shown in italics. Free parameters take their maximum-likelihood estimates (MLEs) when Ω = 40.

cause $N/\max\{K\}$ is small in our study (*e.g.*, Burnham and Anderson 1998). When the expression in the parentheses of the above equation is replaced by *K* we get AIC.

RESULTS

We initially assume a lineage-homogeneous mutational process to model the distribution of the 235 homologous pairs of pure AC repeats. Thus the same Markov chain model (*i.e.*, $\Theta_a = \Theta_c = \Theta_h = \Theta$) is superimposed on the three branches of $\boldsymbol{\tau}$ whose terminal branches are of equal length (*i.e.*, $\lambda_c = \lambda_h = \lambda$). Observe that for time-reversible Markov chains, such as PL1, we can estimate only the sum of the terminal branch lengths (2λ) along with Θ. This is because the per-locus likelihood given by Equation 1 becomes $\Sigma_{j\in\mathbb{S}}\pi_j P_{j,C_i}(\lambda)P_{j,H_i}(\lambda)$ due to lineage homogeneity and further simplifies to $\pi_{C_i}P_{C_i,H_i}(2\lambda)$ due to time reversibility. We relax, and even test, these homogeneity assumptions later when we study repeat-specific and lineage-specific processes.

Ranking the submodels: The submodels of Equation 3 define the set of candidate models to be ranked from best to worst according to their AIC_c values using Equation 7, on the basis of data D_{AC} (see Table 1). Five groupings of models are found. The best group contains the proportional-rate linear-biased models, PL1 and PL2, where longer microsatellites mutate more often than shorter ones toward an attracting focal length. The second-best group comprises EL1 and EL2. In these models, all microsatellites, irrespective of their repeat length, mutate at the same rate toward a focal length. The third-best group comprises the constant-bias models, namely, PC1, PC2, EC1, and EC2. In the presence of a constant downward bias in the mutational process none of the other features seem to matter very much. The proportional-rate, unbiased models, PU1 and PU2, constituting the fourth-best group, outperform their equal-rate, unbiased cousins, EU1 and EU2. Observe that the model ranking is unaffected by variation in the upper bound Ω except for that of the worst group. Since the AIC_c values of PL1 and PL2 are so close we resort to a LRT and attempt to reject PL1 in the next section.

Another ranking of the submodels is performed (Table 2) on the basis of the fit of their stationary distributions to the empirical distribution of pure and isolated AC-repeat lengths in the human genome as described by Calabrese and Durrett (2003). These results are largely consistent with those based on the human and chimp comparison. However, when fitting a model's stationary distribution, due to the large sample size, any increase in the degrees of freedom toward a multinomial model greatly increases its likelihood. For this reason we base our inferences on the human-chimp comparison. The AIC scores are also computed for the different models when the loci are restricted to those in the human-chimp comparison.

One phase *vs.* two phase: The null hypothesis of the simplest, one-phase model EU1 is tested against its two-phase cousin EU2*, through a LRT. The LRTS under this null hypothesis has a nontrivial mixture of χ^2_0, χ^2_1, and χ^2_2 for its asymptotic distribution, since both *p* and *m* lie on the boundary of the parameter space under the null hypothesis (Self and Liang 1987). Instead of analytically pursuing this asymptotic distribution under such nonstandard boundary conditions, we resort to parametric bootstrap to obtain an approximation to the finite sample distribution of the LRTS (see Figure 6A). On the basis of these simulations, there is not enough evidence to reject the one-phase hypothesis ($P = 0.16$). One is unable to reject EU1 in favor of the simpler equal-rate two-phase unbiased model EU2 as well, since the LRTS that is asymptotically $0.5 + 0.5\chi^2_1$ distributed

TABLE 2

Parameter estimation and model ranking using pure AC-repeat loci from the entire human genome

Models	MLEs of parameters				$\delta AIC_{33,298}$	δAIC_{235}
	u	v	m	s		
PL2	0.93	0.05	0.42	1.03	0	2
PL1	0.61	0.01	*1.00*	4.93	16	0
EL2	0.67	0.03	0.43	*0.00*	326	4
EL1	0.54	0.008	*1.00*	*0.00*	564	2
EC1	0.47	*0.00*	*1.00*	*0.00*	10,053	82
PC1	0.47	*0.00*	*1.00*	0.00	10,055	84
EC2	0.47	*0.00*	0.99	*0.00*	10,055	84
PC2	0.47	*0.00*	0.99	0.00	10,057	84
PU1	*0.50*	*1.00*	*1.00*	0.32	18,770	163
PU2	*0.50*	*0.00*	0.99	0.32	18,772	165
EU2	*0.50*	*0.00*	0.93	*0.00*	38,500	335
EU1	*0.50*	*0.00*	*1.00*	*0.00*	38,508	333

The parameters that are fixed for a given submodel are shown in italics. Free parameters take their maximum-likelihood estimates (MLEs) when $\Omega = 40$. Note that $\lambda = \infty$ at stationarity. $\delta AIC_{33,298}$ = AIC − 190,182 for the genomic data (33,298 loci), and δAIC_{235} = AIC − 1281 for the loci restricted to those used in the human-chimp comparison (235 loci).

is observed to be 0.084 ($P = 0.39$). Similarly, we are unable to reject the null hypothesis of every other one-phase model, in favor of its two-phase cousin, except in the equal-rate linear-biased case where one-phase is marginally rejected ($P = 0.013$). The EC2 model with $p = 0$, akin to a truncated version of the SMM of Fu and Chakraborty (1998), as well as PC2 and PU2 assign almost all of the probability mass to single-step jumps. Hence, in these cases, we fail to reject the one-phase hypothesis that $m = 1$ in favor of a two-phase hypothesis that $0 < m < 1$. Among the best group of models, PL1 and PL2, there is inconclusive evidence against one-phase as $P = 0.06$ (see Figure 6B). Furthermore, there is even less evidence in the data to reject PL1 in favor of the most general model ($P = 0.23$). The profile log-likelihood of m under the PL2 model is fairly flat with a wide confidence interval ([0.42, 1]) containing 1. By walking 2 log-likelihood units on either side of the maximum-likelihood value along the profile log-likelihoods, we obtain confidence intervals of the parameters u, v, s, and λ of the PL1 model (Figure 5).

Mutational bias: The absence of any mutational bias as embodied by EU1 is first rejected in favor of the constant-bias model EC1. The maximum-likelihood estimate (MLE) of the constant upward bias parameter $\hat{u} = 0.4650$. EU1 is also rejected in favor of the linear bias model EL1.

The hypothesis of constant mutational bias for all alleles, *i.e.*, EC1, is rejected in favor of the linear-bias model EL1 in the absence of rate proportionality. This LRTS is asymptotically distributed as χ^2_1 under the null hypothesis (see Figure 6C). The MLE of the focal length for the linear-bias model EL1 is 14.

To investigate the nature of mutational bias in the presence of rate proportionality we conducted similar LRTs. Once again absence of bias (PU1) is rejected in favor of its presence (PC1 and PL1) and the hypothesis of constant bias (PC1) is rejected in favor of linear bias (PL1). The MLE of the focal length for the proportional-rate linear-bias model PL1 is 18. When more general functional forms, such as piecewise linear, quadratic, or cubic, were employed to model the dependence of mutational bias on repeat length, the likelihood did not improve significantly enough to reject the linear-bias model (results not shown).

Rate equality *vs.* proportionality: We test the hypothesis of equal mutation rates for all alleles (EU1) against a hypothesis of proportional rates (PU1). This LRTS is asymptotically χ^2_1 distributed under the null hypothesis. Thus, the null hypothesis of rate constancy among alleles is rejected, in favor of a directly proportional relationship between mutation rate and repeat length ($\hat{s} = 0.2556$) in the presence of an unbiased mutation process.

To determine the relevance of rate proportionality in the presence of mutational bias two more LRTs are performed. In the presence of a constant bias, we failed to reject the null hypothesis of rate equality among alleles in favor of rate proportionality ($P = 0.022$). In the presence of linear bias, the LRTS is asymptotically distributed as χ^2_1 under the null hypothesis (see Figure 6D; Table 3). We were able to reject rate equality (EL1) in favor of rate proportionality (PL1). Thus, for pure AC repeats, the proportional-rate linear-bias model (PL1) explains the data best.

When performing multiple LRTs in a nested setting, the order in which such tests are done could affect the final conclusions drawn. We are assured, however, that this order has not influenced our conclusions, since the results of model selection are consistent with those of the hypothesis tests. All conclusions drawn above using the LRTs are robust to changes in the upper bound Ω (results not shown).

So far we have used only pure AC-repeat data (D_{AC}) for inference and assumed homogeneity in the mutational mechanisms across the loci. In doing so, we have ignored interlocus variation and could not address possible motif-specific and interruption-induced complications. Such issues are examined below using PL1, which emerged earlier as the best model.

Interlocus rate variation: The possible presence of variation in mutation rate among loci of pure AC repeats is investigated next. Since λ is estimable as the product of μ and t, variation in mutation rate (μ) translates to variation in λ, as the number of generations (t) remains identical for all loci. We model three equiproportionate classes of loci, 1, 2, and 3, with distinct mutation rates reflected by λ_1, λ_2, and λ_3, respectively. We are unable

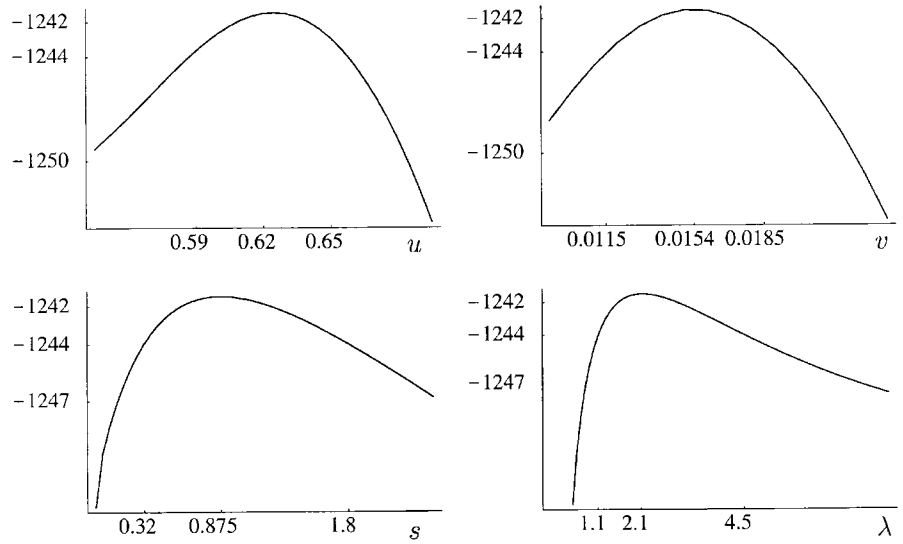

FIGURE 5.—Profile log-likelihood of the parameters u, v, s, and λ of the best model, PL1.

to reject the null hypothesis of equal rates across loci, $H_0: \lambda = \lambda_1 = \lambda_2 = \lambda_3$, in favor of interlocus rate variation, $H_1: 0 < \lambda_1 \leq \lambda_2$, $0 < \lambda_2 < \infty$, and $\lambda_2 \leq \lambda_3 < \infty$, as the asymptotically χ^2_2-distributed LRTS is observed to be 0.67 ($P = 0.73$). We could not reject an equal-rate model in favor of a model with two classes ($P = 0.46$).

Interruption-induced variation: We study possible effects of an interruption by a point mutation on the evolution of otherwise pure AC repeats. Recall that the repeat length of iAC repeat is counted ignoring the interruption. As in the previous section, the stochastic dynamics of pure AC repeats are described by a proportional-rate linear-biased one-phase model with parameters u^{AC}, v^{AC}, s^{AC}, and λ, and those of the iAC repeats are described by another such model with parameters u^{iAC}, v^{iAC}, s^{iAC}, and λ. By calculating the likelihood according to Equation 5, we test hypotheses through LRTs.

The null hypothesis of an identical mutational mechanism between pure AC repeats and iAC repeats, $H_0: u = u^{AC} = u^{iAC}$, $v = v^{AC} = v^{iAC}$, and $s = s^{AC} = s^{iAC}$, is successfully rejected in favor of the alternative, which allows distinct mutational mechanisms, $H_1: u^{AC} \neq u^{iAC}$, $v^{AC} \neq v^{iAC}$, and $s^{AC} \neq s^{iAC}$, since the asymptotically χ^2_3-distributed LRTS is observed to be 26.27. The MLE of the focal length for AC repeats is still 18 but that of the iAC repeats is longer at 21.

The scaled mutation rate $(1/\mu)\beta(i, s)$ is plotted as a function of repeat length using the MLEs of the proportional-rate parameters for pure AC repeats ($\hat{s}^{AC} = 0.83$) and iAC repeats ($\hat{s}^{iAC} = 0.37$) in Figure 7A. The ratio of the MLE of mutation rate of AC repeats over that of iAC repeats, which asymptotes to $0.83/0.37 = 2.24$, is plotted in Figure 7B. The null hypothesis H_0 of identical mutational processes in AC and iAC repeats is also rejected against a simpler alternative that assumes identical bias parameters u and v but distinct proportional-rate parameters s^{AC} and s^{iAC}. For this test the LRTS that is asymptotically distributed as χ^2_1 is observed to be 14.56.

Mutation rate estimation: Assuming 5.5 million years since human-chimp speciation and an average lifetime of 20 years for the two species leads to an estimate of 275,000 generations since speciation. Since $\mu = \lambda/t$ in

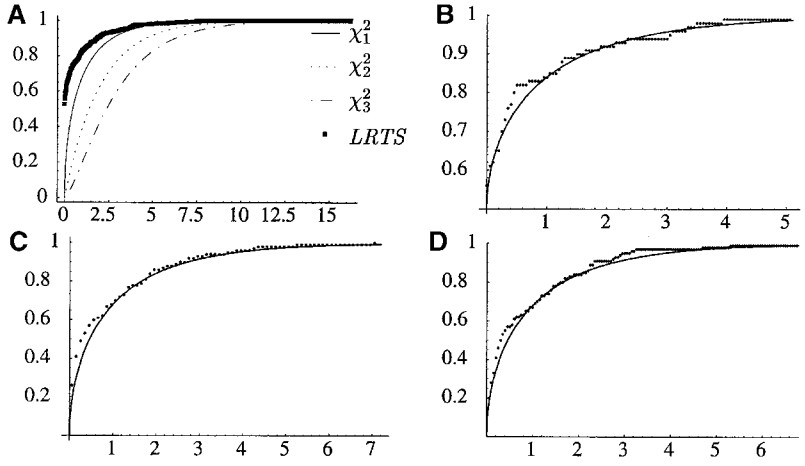

FIGURE 6.—(A) Five hundred simulations of the finite sample LRTS under the null hypothesis for EU1 *vs.* EU2* and 100 simulations each for (B) PL1 *vs.* PL2 $\sim 0.5 + 0.5\chi^2_1$, (C) EC1 *vs.* EL1 $\sim \chi^2_1$, and (D) EL1 *vs.* PL1 $\sim \chi^2_1$. The asymptotically expected distribution in each case is the solid line.

TABLE 3

Some hypothesis tests of time-homogeneous models through likelihood ratios

LRT	H_0 *vs.* H_1	Asymptotic distribution[a]	LRTS	*P*
1	EU1 *vs.* EU2	$0.5\chi^2_0 + 0.5\chi^2_1$	0.084	0.39
2	EU1 *vs.* EU2*	Simulated[b]	1.060	0.16
3	EL1 *vs.* EL2	$0.5\chi^2_0 + 0.5\chi^2_1$	4.92	0.013
4	PL1 *vs.* PL2	$0.5\chi^2_0 + 0.5\chi^2_1$	2.54	0.055
5	EU1 *vs.* EC1	χ^2_1	275.62	$\ll$0.01
6	EU1 *vs.* EL1	χ^2_2	363.91	$\ll$0.01
7	EC1 *vs.* EL1	χ^2_1	88.27	$\ll$0.01
8	EU1 *vs.* PU1	χ^2_1	179.75	$\ll$0.01
9	EC1 *vs.* PC1	χ^2_1	0.022	0.88
10	EL1 *vs.* PL1	χ^2_1	17.84	$\ll$0.01

[a] The expected asymptotic behavior of the likelihood-ratio test statistic (LRTS) under H_0, with $\Omega = 40$.

[b] Simulated finite sample distribution (Figure 6A).

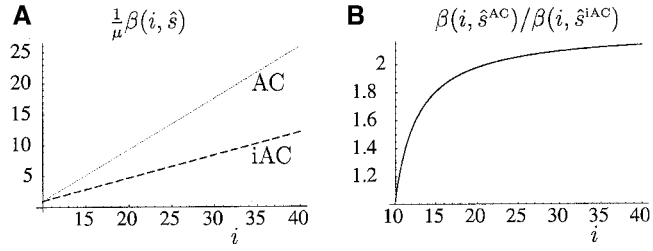

FIGURE 7.—(A) Scaled mutation rates, $(1/\mu)\beta(i, \hat{s}^{AC})$ for pure AC repeats (shaded line) and $(1/\mu)\beta(i, \hat{s}^{iAC})$ for once-interrupted iAC repeats (dashed line), as a function of repeat length under the PL1 model. (B) The rate ratio $\beta(i, \hat{s}^{AC})/\beta(i, \hat{s}^{iAC})$ as a function of repeat length *i*.

our formulation, its MLE $\hat{\mu} = \hat{\lambda}/(2.75 \times 10^5)$. Thus the MLE of the allele-specific mutation rate $\beta(i, \hat{s}) = \hat{\mu}(1 + (i - 10)\,\hat{s})$ is obtained.

To compare $\hat{\mu}$ with the estimates of mutation rates in the literature (which is done in the DISCUSSION) we obtain an average rate $\beta^* = \Sigma_i \hat{\pi}_i \beta(i, \hat{s})$, where $\hat{\pi}_i$ is the stationary probability of allele *i* under the MLEs of the model. For the best model (PL1) β^* is 4.87×10^{-5} per locus per generation and for the worst model (EU1) it is 23% less at 3.76×10^{-5}.

The confidence intervals of [1.1, 4.5] and [0.32, 1.8] for λ and *s*, respectively, translate to a confidence interval of $[1.3 \times 10^{-5}, 1.8 \times 10^{-4}]$ for the average per-locus per-generation mutation rate of pure AC repeats under the PL1 model.

Lineage-specific variation: Here, we relax the assumption of lineage homogeneity that $\Theta_a = \Theta_c = \Theta_h = \Theta$ and allow distinct Markov chain models to be superimposed on distinct branches of τ. We study lineage-specific differences in the mutational mechanism only for the PL1 model. By superimposing a proportional-rate linear-biased one-phase model with parameters u_a, v_a, and s_a upon the ancestral branch; another such model with parameters u_c, v_c, and s_c upon the chimp branch; and finally another with parameters u_h, v_h, and s_h upon the human branch, we address lineage-specific differences in the mutational mechanism of pure AC repeats.

Naturally, the lineage-homogeneous models studied thus far, in which all three branches have superimposed upon them three Markov chain models with identical parameters ($u = u_a = u_c = u_h$, $v = v_a = v_c = v_h$, and $s = s_a = s_c = s_h$), embody the essence of identical mutational mechanisms along the three lineages and constitute our null hypothesis of lineage homogeneity in the mutational process. However, there are numerous ways to model lineage-specific differences in the mutational process. The scenario of biased microsatellite expansion along the human lineage is difficult to discern from that of a biased contraction along the chimp lineage without knowledge of the stationary distribution of the ancestor or repeat length data at homologous loci in an additional outgroup species. In light of additional evidence from a human-chimpanzee-baboon study by WEBSTER *et al.* (2002), which suggests that the dinucleotide repeats in chimpanzees and baboons are similar and a change in the mutational process is more likely to have happened along the lineage leading to humans, we introduce nonhomogeneity by constraining the ancestral mutational mechanism to be identical to that of chimp. Moreover, the nonhomogeneous models that impose identical mutational mechanisms between the human and the ancestral lineages do not have better AIC_c scores (results not shown).

We marginally reject ($P = 0.018$) the null hypothesis of identical mutational mechanisms for the ancestor, chimp, and human microsatellites of the pure AC repeats (PL1 model) in favor of an alternative hypothesis of almost identical mechanisms for the three lineages with the exception of a distinct proportional-rate parameter s_h for the human lineage ($PL1^x$). Since the various alternatives are not nested we resort to AIC_c to rank the models. The better-performing nonhomogeneous models decrease the mutation rate (by decreasing s_h) for longer human microsatellites relative to that of the chimps and/or increase the focal allele of humans by 1 or 2 repeat units as evident from Table 4. Similar two-phase nonhomogeneous models did not perform better than $PL1^x$ (results not shown).

We were also able to fit nonhomogeneous models much better to the empirical distribution of isolated pure AC repeats from human genomic data. A nonhomogeneous PL1 model with seven parameters had a log-likelihood value of −95050.02 and outperformed the time-homogeneous PL2 model from Table 2 by 96 AIC units. The MLEs (not shown) suggest a scenario of ongoing repeat expansion in humans. Figure 8 shows the fits of the homogeneous and nonhomogeneous PL1 models to the empirical distribution of the AC repeats found in the human genome.

TABLE 4

Lineage-specific model ranking

		Lineage-specific parameters									
Models	K	u	u_h	v	v_h	s	s_h	m	λ	log L	δAIC_c
PL1[x]	5	*0.63*	*0.63*	*0.016*	*0.016*	1.40	0.0184	*1*	2.62	−1238.72	0.0
PL1[y]	7	0.63	0.68	0.016	0.020	1.26	0.42	*1*	2.23	−1237.83	2.5
PL1[z]	5	0.63	0.65	*0.016*	*0.016*	*0.88*	*0.88*	*1*	2.10	−1240.12	2.8
PL2	5	*0.82*	*0.82*	*0.04*	*0.04*	*0.76*	*0.76*	0.55	0.56	−1240.21	3.0
PL1	4	*0.62*	*0.62*	*0.015*	*0.015*	*0.88*	*0.88*	*1*	2.14	−1241.48	3.4

K denotes the number of parameters in a model and $\delta AIC_c = AIC_c - 2487.70$. Parameters u_h, v_h, and s_h are for the human lineage, and u, v, and s are for the chimp and ancestral lineages. Estimates shown in italics are fixed to be identical along all three lineages. The MLEs, log L, and AIC_c are computed when $\Omega = 40$.

DISCUSSION

Species-pair data from humans and chimps provide opportunities for analyzing microsatellite evolution not found in population genetic data or genomic data from a single species. A population's demography determines the distribution of its genealogy, which in turn accounts for the correlation among homologous alleles in a population sample. Thus strong demographic assumptions have to be made (Nielsen 1997) to reject one model of microsatellite evolution in favor of another. Our inferences are based on a sample of size 1 from each species and thus do not rely on assumptions regarding the demographic history of the analyzed populations. If the microsatellites themselves are undergoing neutral evolution, then the species-pair data are unaffected by selective sweeps due to samples of size 1 per locus from each species. Thus conditional on the divergence time, we may safely assume independence across loci. We have also assumed that microsatellites of a particular motif share a common mutational mechanism. Due to our small sample size we are unable to allow locus-specific heterogeneity in all the parameters of the mutational model. Since the simplest mixture models that allow interlocus variation in mutation rate for pure AC repeats do not perform significantly better than homogeneous models, it is reasonable to assume that identical mutational mechanisms for microsatellites of the same motif can model our data.

Different models can give rise to similar equilibrium distributions despite distinct finite-time transition probabilities. Thus any inference based on genomic data from one species is limited to parametric families of models whose members have distinct equilibrium distributions (Menendez *et al.* 1999). However, this approach currently has the advantage of larger data sets over our species-pair approach, as the chimp genome is not yet fully sequenced. We provide a framework for hypothesis tests directed at a mechanistic understanding of the mutational process of microsatellites using information about their divergence.

Our analysis indicates that bias in the mutational process and proportionality in mutation rate are vital for realistic stochastic models of evolution of pure dinucleotide repeats. The models with a linear bias toward a focal length, in the spirit of Garza *et al.* (1995), constitute the top four models. This suggests an important role for bias toward a target length in microsatellite evolution using interspecies divergence data, consistent with the study

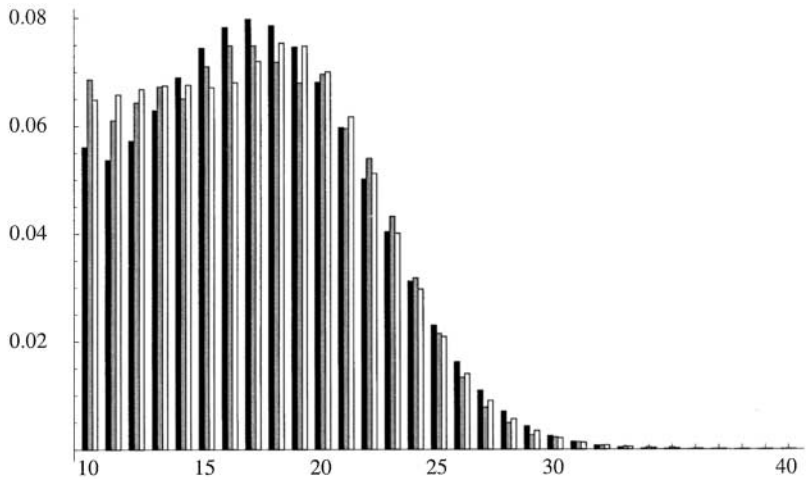

Figure 8.—Stationary distribution of the homogeneous (solid bars) PL1 model, transient distribution of the nonhomogeneous (open bars) PL1 model, and empirical distribution of the isolated AC repeats (shaded bars) in the human genome.

of Garza *et al.* (1995) and Zhivotovsky *et al.* (1997) based on population data, and further affirms the findings of Calabrese and Durrett (2003) that proportional slippage is not sufficient in the absence of mutational bias to explain the human genomic microsatellite length distribution. Finally, using data from parent-offspring transmissions of AC repeats, Whittaker *et al.* (2003) also showed that contractions become more likely for microsatellites >20 repeats while expansions become frequent for shorter repeats.

The linear bias may be construed as a signature of underlying counteracting forces in the mutational mechanism, *i.e.*, upward mutation bias of primary slippage mutations countered by the downward mutation bias at longer alleles due to the mismatch repair system (Harr *et al.* 2002). Since the effects of mutational and substitutional processes are confounded in our human-chimp data, natural selection could also be contributing to the downward bias by acting directly against longer microsatellites if they confer some selective disadvantage or by acting indirectly upon the mismatch-repair machinery itself. However, similar findings in the pedigree studies of Whittaker *et al.* (2003) suggest that linear bias may truly be a signature of the microsatellite mutational process alone.

The biased models are robust to variation in the upper bound Ω, as is evident from their stable AIC_c values for larger Ω, due to the presence of a downward or focal bias. The unbiased models, on the other hand, do considerably worse for larger values of Ω, because as microsatellites mutate without preferring contractions over expansions, they distribute themselves uniformly over the entire state space as time progresses. Thus, when Ω is large, microsatellites can attain unrealistically large repeat lengths under the unbiased models. The lower boundary κ was chosen to be 10 because we wanted it to be higher than the threshold at which slippage is empirically observed to occur (typically 8 repeat units). For one-phase models that allow jumps of only a single repeat unit at a time, the choice of the lower boundary poses little problem. One may view this, in the queuing theory terminology, as a harmless effect on the total number of customers entering the system from the lower boundary but not on their relative numbers (Calabrese *et al.* 2001). However, in more complicated models that allow microsatellites to enter the truncated state space at several distinct repeat lengths, inference can be sensitive to the choice of the lower boundary.

Among the one-phase models, rate proportionality gives a better fit to the data than rate equality among alleles in the presence of an unbiased or a linear-biased mutational process. However, it does not do so in the presence of a strong constant downward bias ($\hat{u} = 0.46$). Under a constant downward bias, most of the probability mass under stationarity is already piled over shorter alleles, and thus any increase in rate proportionality will only exacerbate this trend by reducing the mean holding time of longer alleles and thereby further reducing their stationary probability. In fact, the small negative value taken by the proportional-rate parameter ($\hat{s} = -0.0048$) reflects some level of restoration of probability mass to longer alleles, countering the effects of geometric decay caused by constant bias. In the absence of any mutational bias, on the other hand, the ratio term $\alpha(u, v, j)/(1 - \alpha(u, v, j + 1))$ in the finite product of Equation 4 simplifies to 1 for all alleles and thus makes the effects of proportionality pronounced. Any increase from 0 in the proportional-rate parameter s shifts the probability mass away from being uniformly distributed among all alleles toward shorter alleles, reflecting their increased mean holding times relative to longer alleles. Similarly, under linear bias, the effects of proportionality are pronounced as this finite product has terms both ≥ 1 and < 1 for longer alleles. Thus, rate proportionality cannot be ignored even in the presence of linear bias.

The truncated TPM of DiRienzo *et al.* (1994) fits the pure AC-repeat data by essentially mimicking the truncated SMM of Ohta and Kimura (1973). The two-phase models generally mimic their one-phase cousins, as reflected by their values of $\hat{m}$ being close to 1. Our inability to reject one-phase in favor of two-phase using human-chimp data is in contrast with experimental observations of multistep mutations. There are several explanations for this. First, noise in repeat length estimates due to indel activity in the flanking region may be at least partly responsible for elevating the experimentally observed proportion of multistep mutations. Empirical studies usually keep track of the length of a microsatellite repeat along with its flanking sequence (PCR fragment length), rather than the actual repeat length. Studies have found both interspecific and intraspecific fragment length polymorphism to be caused by indels in the flanking regions (Angers and Bernatchez 1997; Matsuoka *et al.* 2002). Thus, on a cautionary note, indels in the flanking sequence could be construed as multiunit microsatellite mutations if repeat lengths are directly extrapolated from the PCR fragment length. Most studies that found two-phase models to produce better fits than their one-phase cousins used some transformation of the PCR fragment length for their data. Since the evolution of such microsatellite-containing PCR fragment lengths is influenced by the local indel activity as well as the true microsatellite mutations, it becomes difficult to make inference on the nature of two-phase mutations inherent to microsatellites alone with such PCR fragment length data. On the other hand, two-phase models may be more appropriate than one-phase models for such PCR-extrapolated data as shown in a recent study by Whittaker *et al.* (2003).

Second, the lack of evidence for our two-phase models should really be seen as the rejection of a homogeneous two-phase mechanism that is insensitive to repeat length in favor of a homogeneous one-phase mechanism. We forged our two-phase models in the image of TPM of

DiRienzo *et al.* (1994) and an SMM of Fu and Chakraborty (1998). However, other formulations of a two-phase mutational mechanism, particularly those that allow the probability p of single-step mutations and/or the success probability m of the conditional geometric distribution specifying the lengths of multistep jumps to decrease with repeat length, may be more realistic, especially in light of empirical evidence for large contractions being more common among long alleles in yeast (Wierdl *et al.* 1997) and fruit fly (Harr *et al.* 2002). As more of the chimp genome gets sequenced such varying two-phase models should be tested to further evaluate the importance of multistep mutations.

There is a twofold decrease in the slippage rate and a 6-bp increase in the focal length of an AC repeat interrupted by just one point mutation relative to a pure repeat. This is not surprising as a point mutation is expected to create fewer opportunities for polymerase slippage and thereby decreases mutation rate as demonstrated in yeast (Petes *et al.* 1997). Moreover, longer repeats are more prone to getting interrupted by point mutations. Upon interruption, they are less likely to mutate and thereby contract, due to linear bias toward the focal length, as much as the pure repeats.

Our mutation rate estimates are not significantly different from the empirical estimates of per-locus rates of 6×10^{-4} for autosomal dinucleotide repeats (Ellegren 2000b) and 2.3×10^{-4} for CA repeats (Petrukhin *et al.* 1993). Empirical overestimates of the true mutation rate may result from sampling bias toward highly polymorphic loci, which are typically also the fastest mutating. If the loci chosen to estimate mutation rate empirically have longer alleles on average, then an overestimation of the true average may result. The sample in our study is small for reliable mutation rate estimates as reflected in the large confidence interval of $[1.3 \times 10^{-5}, 1.8 \times 10^{-4}]$.

There is evidence in the human-chimp data as well as in human genomic data to reject lineage homogeneity in favor of lineage-specific variation in the evolution of pure AC repeats. The effects of increase in focal allele length and decrease in mutation rate for longer alleles along the human lineage relative to those along the chimp lineage are consistent with other studies (Cooper *et al.* 1998; Webster *et al.* 2002). One possible explanation is that the human mismatch-repair system is not as efficient as that of the chimp. As has been pointed out by Harr *et al.* (2002), subtle differences in the mismatch-repair system between two species could give rise to distinct mutational biases. An equally compatible explanation is selection against longer repeats and the differences in the effective population size (N_e) between the two species. A lower value of N_e in humans could decrease the relative effectiveness of selection. Additional data are required to distinguish between hypotheses purporting differences in mismatch-repair machinery and those invoking selection to explain lineage-specific differences in the evolution of pure dinucleotide repeats.

These methods can be extended to more species as more primate sequences become available. One can test hypotheses and estimate parameters in a locus-specific as well as lineage-specific manner simultaneously. In particular, as data for primates accrue, it would be biologically relevant to use more general functional forms to model mutational bias as well as the nature of two-phase mutations. One may further use such species-specific and motif-specific parameter estimates in various population genetic inferences. The impact of model misspecification on signals of selective sweeps from microsatellite variation also needs to be investigated.

R.S. thanks William Amos, Dave Capella, Lounes Chikhi, Floyd Reed, Guy Reeves, Gennady Samorodnitsky, James Signorovitch, and Robert Strawderman for insightful discussions. R.S. is supported by the Integrative Graduate Education, Research and Traineeship from National Science Foundation (NSF) grant DGE-9870631; NSF grant DEB-0089487 to R. Nielsen; and NSF/National Institutes of Health grant DMS/NIGMS 0201037 to R. T. Durrett, C. F. Aquadro, and R. Nielsen.

LITERATURE CITED

Amos, W., S. J. Sawcer, R. W. Feakes and D. C. Rubinsztein, 1996 Microsatellites show mutational bias and heterozygote instability. Nat. Genet. **13:** 390–391.

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

Brémaud, P., 1999 *Markov Chains, Gibbs Fields, Montecarlo Simulations and Queues.* Springer-Verlag, New York.

Burnham, K. P., and D. R. Anderson, 1998 *Model Selection and Inference: A Practical Information-Theoretic Approach.* Springer-Verlag, New York.

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

Calabrese, P. P., R. T. Durrett and C. F. Aquadro, 2001 Dynamics of microsatellite divergence. Genetics **159:** 839–852.

Cooper, G., D. C. Rubinsztein and W. Amos, 1998 Ascertainment bias cannot entirely account for human microsatellites being longer than their chimpanzee homologues. Hum. Mol. Genet. **7:** 1425–1429.

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

Ellegren, H., 2000a Heterogeneous mutation processes in human microsatellite DNA sequences. Nat. Genet. **24:** 400–402.

Ellegren, H., 2000b Microsatellite mutations in the germline: implications for evolutionary inference. Trends Genet. **16:** 551–558.

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.

Fu, Y., and R. Chakraborty, 1998 Simultaneous estimation of all the parameters of a step-wise mutation model. Genetics **150:** 487–497.

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

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 under-representation. Genetics **155:** 1213–1220.

Harr, B., J. Todorova and C. Schlötterer, 2002 Mismatch repair-driven mutational bias in *D. melanogaster.* Mol. Cell **10:** 199–205.

Huang, Q., F. Xu, H. Shen, Q. Deng, Y. Liu *et al.*, 2002 Mutation

patterns at dinucleotide microsatellite loci in humans. Am. J. Hum. Genet. **70:** 625–634.

Hurvich, C., and C.-L. Tsai, 1989 Regression and time series model selection in small samples. Biometrika **76:** 297–307.

Jarne, P., and P. Lagoda, 1996 Microsatellites, from molecules to populations and back. Trends Ecol. Evol. **11:** 424–429.

Kruglyak, S., R. T. Durrett, M. D. 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.

Matsuoka, Y., S. E. Mitchell, S. Kresovich, M. Goodman and J. Doebley, 2002 Microsatellites in *Zea*—variability, patterns of mutations, and use for evolutionary studies. Theor. Appl. Genet. **104:** 436–450.

Menendez, M. L., D. Morales, L. Pardo and I. Vajda, 1999 Inference about stationary distributions of Markov chains based on divergences with observed frequencies. Kybernetika **35:** 265–268.

Nauta, M. J., and F. J. Weissing, 1996 Constraints on allele size at microsatellite loci: implications for genetic differentiation. Genetics **143:** 1021–1032.

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

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.

Ouellette, B. F. F., and M. S. Boguski, 1997 Database divisions and homology search files: a guide for the perplexed. Genome Res. **7:** 952–955.

Petes, T. D., P. W. Greenwell and M. Dominska, 1997 Stabilization of microsatellite sequences by variant repeats in the yeast *Saccharomyces cerevisiae.* Genetics **146:** 491–498.

Petrukhin, K. E., M. C. Speer, E. Cayanis, M. F. Bonaldo, U. Tantravahi *et al.*, 1993 A microsatellite genetic linkage map of human chromosome 13. Genomics **15:** 76–85.

Primmer, C. G., H. Ellengren, N. Saino and A. P. Moller, 1996 Directional evolution in germline microsatellite mutations. Nat. Genet. **13:** 391–393.

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

Self, S. G., and K. Liang, 1987 Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J. Am. Stat. Assoc. **82:** 605–610.

Sugiura, N., 1978 Further analysis of the data by Akaike's information criterion and the finite corrections. Comm. Stat. Theor. Meth. **A7:** 13–26.

Tachida, H., and M. Iizuka, 1992 Persistence of repeated sequences that evolve by replication slippage. Genetics **131:** 471–478.

Walsh, J. B., 1987 Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. Genetics **115:** 553–567.

Webster, M. T., N. G. C. Smith and H. Ellegren, 2002 Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. Proc. Natl. Acad. Sci. USA **99:** 8748–8753.

Whittaker, J. C., R. M. Harbord, N. Boxall, I. Mackay, G. Dawson *et al.*, 2003 Likelihood-based estimation of microsatellite mutation rates. Genetics **164:** 781–787.

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

Wilks, S. S., 1938 The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat. **9:** 60.

Wolfram, S., 1999 *The Mathematica Book*, Ed. 4. Cambridge University Press, Cambridge, UK.

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

Zhivotovsky, L. A., M. W. Feldman and S. A. Grishechkin, 1997 Biased mutations and microsatellite variation. Mol. Biol. Evol. **14:** 926–933.

Communicating editor: M. Feldman