Likelihood-Based Estimation of Microsatellite Mutation Rates

John C. Whittaker,[*,†,1] **Roger M. Harbord,**[*,‡] **Nicola Boxall,**[§] **Ian Mackay,**[**] **Gary Dawson**[**] **and Richard M. Sibly**[§]

[*]*School of Applied Statistics, University of Reading, Reading RG6 6FN, United Kingdom,* [†]*Department of Epidemiology and Public Health, Imperial College London, London W2 1PG, United Kingdom,* [‡]*Department of Social Medicine, University of Bristol, Bristol BS8 2PR, United Kingdom,* [§]*School of Animal and Microbial Sciences, University of Reading, Reading RG6 6AJ, United Kingdom and* [**]*Oxagen Ltd., Abingdon OX14 4RY, United Kingdom*

Manuscript received November 5, 2002
Accepted for publication February 25, 2003

ABSTRACT

Microsatellites are widely used in genetic analyses, many of which require reliable estimates of microsatellite mutation rates, yet the factors determining mutation rates are uncertain. The most straightforward and conclusive method by which to study mutation is direct observation of allele transmissions in parent-child pairs, and studies of this type suggest a positive, possibly exponential, relationship between mutation rate and allele size, together with a bias toward length increase. Except for microsatellites on the Y chromosome, however, previous analyses have not made full use of available data and may have introduced bias: mutations have been identified only where child genotypes could not be generated by transmission from parents' genotypes, so that the probability that a mutation is detected depends on the distribution of allele lengths and varies with allele length. We introduce a likelihood-based approach that has two key advantages over existing methods. First, we can make formal comparisons between competing models of microsatellite evolution; second, we obtain asymptotically unbiased and efficient parameter estimates. Application to data composed of 118,866 parent-offspring transmissions of AC microsatellites supports the hypothesis that mutation rate increases exponentially with microsatellite length, with a suggestion that contractions become more likely than expansions as length increases. This would lead to a stationary distribution for allele length maintained by mutational balance. There is no evidence that contractions and expansions differ in their step size distributions.

MICROSATELLITES consist of repeats of short sequences (1–6 bp) of DNA and are very common in eukaryotic genomes. They are highly mutable, with the primary mutational mechanism believed to be replication slippage (Ellegren 2000a): during replication of the microsatellite region the strands may be displaced and then realign incorrectly, leading to insertion or deletion of a number of repeat units. Their abundance and high variability has led to wide use in a variety of genetic analyses, many of which require reliable estimates of microsatellite mutation rates, yet the factors determining mutation rates are as yet poorly understood (Ellegren 2000a,b; Kayser *et al.* 2000; Sibly *et al.* 2001, 2003).

The simplest model of microsatellite mutation, often known as the stepwise mutation model (SMM; Ohta and Kimura 1973), assumes that length, measured as number of repeat units, changes by only 1 unit per mutation, with expansions and contractions equally likely. A number of developments of this model have also been proposed: for example, the failure of the basic SMM to explain why some alleles do not expand indefinitely led to the suggestion (Bell and Jurka 1997) that point mutation might prevent expansion by breaking long microsatellites into smaller ones, an idea further developed in Kruglyak *et al.* (1998, 2000).

However, fitting these mutational models to data is not straightforward, because of the difficulty of obtaining sufficient mutational events. Two broad strategies have been pursued. First, the mutational model may be used to provide an equilibrium distribution of microsatellite length, which may then be compared with that observed in DNA sequences, either by genotyping a number of individuals at one or more markers (Ewen *et al.* 2000) or by cumulating many microsatellites from the published genome sequence (Kruglyak *et al.* 1998; Sibly *et al.* 2001). These analyses are attractive because of the easy availability of the appropriate data, but have the disadvantage that they do not allow direct observation of mutational events and so rely on evolutionary assumptions, notably that the mutational process has reached equilibrium. Furthermore, we often find that many models of the mutational process are compatible with the observed length distributions.

Direct observation of the mutational events is preferable if possible, but involves much more genotyping.

[1] *Corresponding author:* Department of Epidemiology and Public Health, Imperial College School of Medicine, St. Mary's Campus, Norfolk Pl., London W2 1PG, United Kingdom.
E-mail: j.whittaker@ic.ac.uk

Typing of large numbers of sperm is one approach (Leeflang *et al.* 1995), but a number of recent studies (Xu *et al.* 2000; Huang *et al.* 2002) have exploited the large quantities of microsatellite genotype data collected on human pedigrees for studies of human disease. The disadvantage of such studies is that the microsatellites used as molecular markers in such studies are selected first to be highly polymorphic and second to be easy to genotype: they may therefore not be representative of the global population of microsatellites.

Despite these difficulties, some interesting results have emerged. It is now generally accepted that mutation rate increases with allele length, measured as number of repeats (Brinkmann *et al.* 1998; O'Connell and Weeks 1998; Ellegren 2000a,b; Kayser *et al.* 2000; Kruglyak *et al.* 2000; Xu *et al.* 2000; Sibly *et al.* 2001), and that although mutations generally consist of a length change of one repeat, two or occasionally more are known (Ellegren 2000a); indeed, Huang *et al.* (2002) suggest that multistep mutations may be very common. Several studies have suggested a mutation bias in favor of expansions, while some recent analyses have argued that long alleles tend to mutate to shorter alleles, so that infinite expansion is prevented by mutational balance (Ellegren 2000a,b; Xu *et al.* 2000; Huang *et al.* 2002). However, much of the evidence in support of this hypothesis comes from direct observation of allele transmissions. In addition to the ascertainment problem referred to above, these studies have the limitation that, with the exception of microsatellites on the Y chromosome (Kayser *et al.* 2000), previous analyses have not made full use of available data and may have introduced bias. This arises because mutations have been identified only where child genotypes could not be generated by transmission from parents' genotypes, and so the probability that a mutation is detected depends on the distribution of allele lengths and varies with allele length. Perhaps more seriously, the dependence on simple counting schemes makes it difficult to formally compare the fit of competing models of microsatellite evolution. In this article we introduce a likelihood-based approach that avoids both of these problems, properly allowing for undetected mutations and allowing comparison of models via the standard likelihood-based statistical machinery. Our new method is applied to data on almost 400 *AC* dinucleotide microsatellites from a genome scan conducted by Oxagen Ltd. on 123 extended families of two to four generations.

METHODS

Data collection: Automated genotyping of blood samples from 680 individuals was performed by an ABI PRISM 377 DNA sequencer and interpreted with associated software GeneScan and Genotyper Software v 3.6 (Applied Biosystems, Foster City, CA) to filter out stutter peaks and A+ peaks. Genotyping errors are removed using the protocols in Ewen *et al.* (2000). In brief, Mendelian-inheritance errors were identified with PedCheck v 1.00 (O'Connell and Weeks 1998) and checked manually to remove errors due to unreliable pedigrees, differential or preferential amplification, allele misassignation, low-intensity data, multiple peaks, PCR failure, noisy baseline, well-to-well leakage, bleed through, capillary failure, and contamination. Alleles were classified as nulls and removed from the data set if a parent and its offspring were apparently homozygous for the same allele, since this possibility cannot be distinguished from the possibility that one of the two alleles at the locus failed to amplify. Simulations showed that removing nulls has little effect on the results (Harbord 2001). See Ewen *et al.* (2000) for a more detailed discussion of error and mutation detection in high-throughput microsatellite genotyping.

Allele sizes were converted to number of repeats by stripping out the nonmicrosatellite bases between the two primer sequences using sequence data from the Foundation Jean Dausset-CEPH database (version 8.1; Dib *et al.* 1996). The marker set was based on the ABI PRISM linkage mapping set version 2 (http://home.appliedbiosystems.com/), although a small number of markers that had given consistently high error rates or low genotyping success rates in the Oxagen laboratory had been substituted to maintain an approximate genome-wide marker-to-marker interval of 10 cM.

Statistical analysis: The likelihood of the observed data set was calculated for each of the models described in results, as follows. Consider a triplet with parental marker genotypes (x_1, x_2, x_3, x_4) and child genotype (x_5, x_6), where alleles (x_1, x_2) are carried by one parent and (x_3, x_4) by the other. To avoid the need to model parental genotype probabilities, we worked with the likelihood conditional on parental genotypes. Writing p_{ij} for the probability that a microsatellite of length i in the parental generation mutates to length j in the child generation, this conditional likelihood is easily shown to be equal to

$$(p_{x_1x_5} + p_{x_2x_5})(p_{x_3x_6} + p_{x_4x_6}) + (p_{x_1}p_{x_6} + p_{x_2x_6})(p_{x_3x_5} + p_{x_4x_5}). \tag{1}$$

We consider a number of possible models for p_{ij} below.

Since child genotypes are independent even for sibs once we have conditioned on parental genotypes, the likelihood for the complete data set is then simply the product over all parent-offspring triplets. Maximization for the models discussed above gives estimates for the underlying parameters and allows the calculation of confidence intervals and the comparison of nested models via the usual statistical machinery.

Statistical analysis was performed using the statistical language S-PLUS. Likelihoods were maximized using the built-in function **nlminb**, calling functions for likelihood calculations coded in C for speed. Times for a

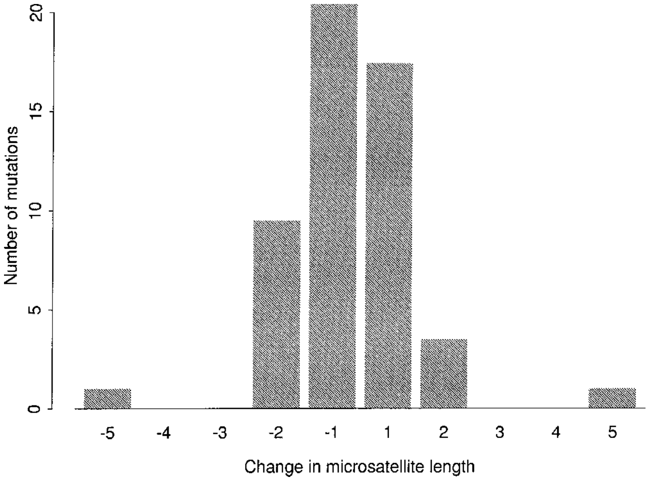

FIGURE 1.—Classification of the 53 naive mutations by step size (measured in numbers of repeats) and direction.

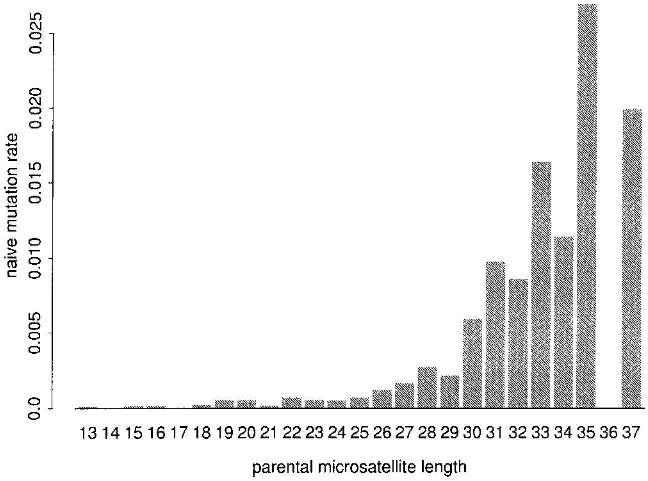

FIGURE 2.—Length dependence of naive mutation rate, per haplotype per generation. Length is measured in number of repeats.

single maximization were of the order of 10–100 sec, depending on the complexity of the model fitted, on a Sun Ultra 10 workstation running Solaris (SPECfp95 12.9). We found that working with logistic transforms of the transition probabilities p_{ij} gave much-improved numerical properties. Model comparisons are based on the usual result that for two nested models for which d_1 and $d_2 > d_1$ and L_1 and L_2 denote the number of independent parameters and values of the maximized likelihood, respectively, $2(\log L_2 - \log L_1)$ has asymptotically a $\chi^2_{d_2-d_1}$ distribution when model 1 is true (Cox and HINKLEY 1974). We also give values of Akaike's information criteria (AIC),

$$\text{AIC} = -2(\text{maximized log-likelihood}) + 2(\text{no. of parameters}), \tag{2}$$

for the models considered. Standard procedure is to choose the model minimizing AIC as optimal (ANDERSON and BURNHAM 1998). Confidence intervals were obtained using a modified version of the S-PLUS function **vcov.nlminb** (VENABLES and RIPLEY 1999) to give standard errors of parameter estimates, followed by an assumption of normality on the logit scale.

RESULTS

The data set contained 118,866 parent-offspring transmissions from 59,433 parent-offspring triplets. We identified 53 Mendelian discrepancies, giving a naive estimate of the overall mutation rate of 4.5×10^{-4} per allele transfer. It is never possible to identify with certainty the mutational event causing the discrepancy, but by taking from the set of possible mutations the one involving the smallest change in length, measured as number of *AC* repeats as described above, we were able to classify the mutations by step size and direction (Figure 1) and thus obtained naive estimates of mutation rates by length (Figure 2).

These naive estimates may be biased, as explained in the Introduction. To circumvent this, we adopted the likelihood-based approach described in METHODS and applied it to a series of mutational models chosen to be compatible with existing knowledge and with Figures 1 and 2. For convenience, the parameters of the models are listed in Table 1.

In the first model considered, we allow for mutations of any step size, but acknowledge that smaller changes in the number of repeats are more common than larger changes by assuming that the probability of a mutation of step size k declines geometrically with k. This model therefore has two parameters: μ, which represents the overall mutation rate, and an exponential decay rate parameter λ. This model (1 in Table 1) is symmetric in that the same relationship is assumed for up and down mutations, and it is length independent in that mutation rates are independent of microsatellite length. Generalizations (2–4 in Table 1) allow for either overall mutation rate or the exponential decay parameter, or both, to vary according to the direction of the mutation. Thus the most general model in this family has four parameters, μ_u and μ_d controlling overall mutation rate and λ_u and λ_d controlling the exponential decay with increasing step size, with the subscripts u and d indicating whether the mutation gives an increase or decrease in microsatellite length, respectively. With p_{ij} the probability that a microsatellite of length i in the parental generation mutates to length j in the child generation as above, we thus obtain

$$p_{ij} = \begin{cases} \mu_u e^{-\lambda_u(j-i)}, & j > i \\ \mu_d e^{-\lambda_d(i-j)}, & j < i \\ 1 - \sum_{i \neq j} p_{ij}, & j = i. \end{cases} \tag{3}$$

However, since there is strong evidence that mutation

TABLE 1

Log-likelihood and AIC values for the multistep length-independent and length-dependent models, 1–4 and 5–9, respectively

Model	Parameters	No. of parameters	Log-likelihood	AIC
1	μ, λ	2	0.00	0.00
2	μ, λ_u, λ_d	3	0.00	1.99
3	μ_u, μ_d, λ	3	0.46	1.08
4	μ_u, μ_d, λ_u, λ_d	4	0.46	3.08
5	γ, α, λ	3	40.46	−78.93
6	γ_u, γ_d, α, λ	4	41.73	−79.45
7	γ, α, λ_u, λ_d	4	40.63	−77.26
8	γ_u, γ_d, α_u, α_d, λ	5	43.16	−80.31
9	γ_u, γ_d, α_u, α_d, λ_u, λ_d	6	43.27	−78.55

Log-likelihood and AIC values are given relative to those of model 1.

TABLE 2

Parameter estimates and 95% confidence intervals for models 5 and 8

Model	Quantity	Estimate	Confidence interval
5	$\hat{\gamma}$	8.8×10^{-7}	$(2.7, 28.7) \times 10^{-7}$
	$\hat{\lambda}$	1.06	(0.69, 1.43)
	$\hat{\alpha}$	0.263	(0.215, 0.310)
8	$\hat{\gamma}_u$	3.1×10^{-6}	$(0.48, 19.2) \times 10^{-6}$
	$\hat{\gamma}_d$	4.0×10^{-7}	$(0.86, 18.7) \times 10^{-7}$
	$\hat{\lambda}$	1.06	(0.68, 1.43)
	$\hat{\alpha}_u$	0.200	(0.130, 0.269)
	$\hat{\alpha}_d$	0.302	(0.244, 0.360)

See Table 1 for model specifications.

rates increase with the length of the parental allele (Weber and Wong 1993; Leeflang *et al.* 1995; Sibly *et al.* 2001), we also consider length-dependent models (5–9 in Table 1) that allow the mutation rate to depend on the length of the parental microsatellite, *i*, by making the overall mutation rate μ a function of *i*. Guided by the results in Figure 2 we choose an exponential relationship between rate and parental length, and again the model can be further extended to allow dependence on the direction of the mutation. Our most general model therefore has six parameters, γ_u and γ_d determining the underlying mutation rate, α_u and α_d controlling the rate of change of mutation rate with parental microsatellite length, and λ_u and λ_d as in the length-independent models. Thus p_{ij} is as above but with

$$\mu_u(i) = \gamma_u e^{\alpha_u i}$$

$$\mu_d(i) = \gamma_d e^{\alpha_d i}.$$

Maximized log-likelihoods and values of the AIC are given in Table 1 for each of these models. The length-independent models (1–4) differ little in log-likelihood (Table 1), with none giving a significant improvement over model 1. However, all the length-dependent models (5–9) give hugely significant improvements in fit compared to any of the length-independent models ($P < 10^{-15}$, by the usual likelihood-ratio tests). The best-fitting model based on the AIC is model 8, in which the direction of mutation affects the dependence of mutation rate on parental allele length but not on step size. The natural model with which to compare model 8 is model 5, in which length dependence is the same for up and down mutations. Comparing these two models gives a change in log-likelihood of 2.69, which, by referring 2×2.69 to the χ^2_2 distribution, gives $P = 0.068$, and so is marginally significant. Parameter estimates and 95% confidence intervals of the parameters in these models are given in Table 2. Figure 3 shows the mutation rates predicted by model 8, plotted on a log scale against parental allele length.

Model 8 predicts that up mutations will exceed down mutations for microsatellites of <20 repeats, with the opposite true for microsatellites with >20 repeats. Though we must bear in mind the considerable uncertainty in parameter estimates when interpreting these values, this would lead to an equilibrium distribution for microsatellite length with a mode at 20, which is exactly what we see in the distribution of parental alleles plotted in Figure 4. Thus it is possible that the length distribution at the loci we studied is maintained by the length-dependent mutation bias reported in Figure 3. This cannot apply to all AC loci, however, since in the genome overall the frequency of AC alleles decreases with their length (Sibly *et al.* 2003). Thus the microsatellite loci used in linkage studies may have rather different mutation rates from those in the genome overall. This emphasizes the difficulty of reconciling results from studies such as the one reported here with genome-wide indirect studies.

We also considered models in which mutation rate increased linearly with allele length as in Sibly *et al.* (2001) but these models led to predictions of negative mutation rates over much of the range of interest. Restriction to positive mutation rates gave an AIC of 22.88 (note that the linear and exponential models are not nested and therefore cannot be compared by likelihood-ratio tests), indicating poor fit relative to the models described above. Results are therefore not presented here. Similarly, we found no evidence of sex bias in mutation rate, in contradiction to the expectation of elevated rates in males relative to females (Weber and Wong 1993); again results are not presented here.

DISCUSSION

The maximum-likelihood method introduced here avoids the biases inherent in the naive estimation methods used in previous studies and makes fuller use of the

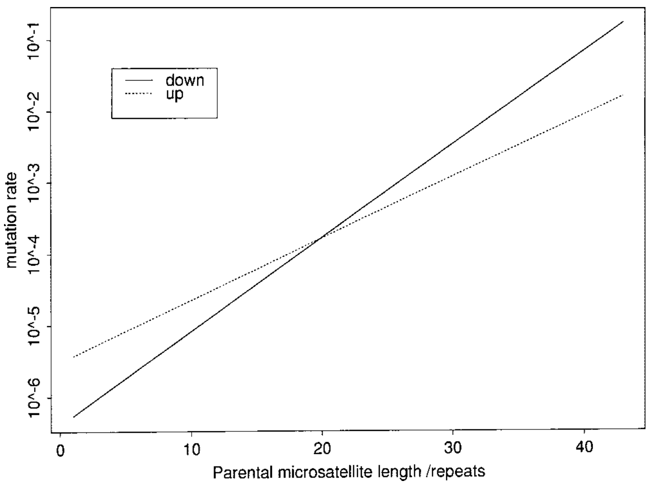

FIGURE 3.—Log-mutation rates per haplotype per generation *vs.* parental microsatellite length in number of repeats for model 8.

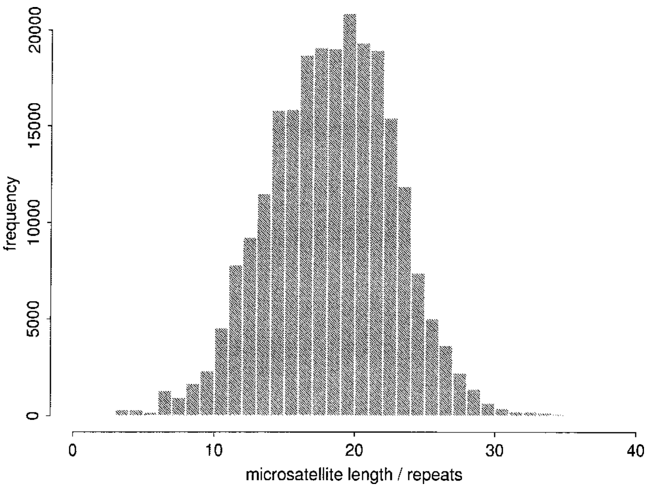

FIGURE 4.—Histogram of microsatellite lengths in number of repeats for parental alleles.

available data. Simulation results suggest that for our data set naive estimates of mutation rates are biased downward by ~12% (HARBORD 2001). Biases in other data sets will in general be different, since the extent of bias depends on the distribution of parental genotypes. Perhaps more importantly, when using a likelihood approach hierarchical models can be compared to establish whether particular model features are worth including, as illustrated above in the comparisons between the models of Table 1.

The method described here involves conditioning on parental genotypes (1), so that all information on mutation is obtained directly from parent-child transmissions. It is also possible to write down the complete likelihood, including terms for the likelihood of parental genotypes. However, these depend on the population relative frequencies of genotypes in the population, which in turn depend both on the mutational model and on population history. In principle, we could therefore write down a complete likelihood, incorporating both a mutational model and a model for population history, which unifies the "direct" and "indirect" approaches to inference on microsatelite mutation. However, here we have preferred to concentrate on direct inference of mutational mechanisms, avoiding dependence on population history.

Our method also assumes that at each locus we have no information about which parental allele is transmitted aside from allele length at that locus. If information on linked markers is available this could be incorporated, thus reducing uncertainty about the parental origin of transmitted alleles. It should be possible to modify the algorithms used for likelihood calculations in linkage analysis (*e.g.*, KRUGLYAK *et al.* 1996) to achieve this.

Size and direction of mutational steps: The naive analysis presented in Figure 1 suggests that most mutations consist of steps of one repeat, with the distribution of step sizes similar in both up and down directions. The maximum-likelihood analysis reinforces these conclusions (Table 1). Thus although in the best-fitting model (8) up and down mutation rates have separate relationships with parental allele length, the distribution of step sizes given that a mutation occurs is controlled by a single parameter, λ, and no significant increase in likelihood was obtained by separating mutations according to whether they increased or decreased microsatellite length.

The maximum-likelihood estimate of the exponential decay rate parameter λ was 1.06 (Table 1). This suggests that 65% of mutations were of step size 1, 23% of size 2, and the remaining 12% of step size greater than 2. The frequency of multistep changes is higher than that recorded in most previous studies, which give values in the range 0–14% (WEBER and WONG 1993; AMOS and RUBINSTZEIN 1996; BRINKMANN *et al.* 1998; ELLEGREN 2000b; KAYSER *et al.* 2000; XU *et al.* 2000), but lower than the 63% recorded by HUANG *et al.* (2002). Note that we obtain a 95% confidence interval for λ of (0.68, 1.43) $\times 10^{-7}$; taking the boundary points gives 23 and 50% multistep changes. If we treat the 61 multistep mutations identified by HUANG *et al.* (2002) as having a binomial distribution conditional on their total number of mutations (97), we get a 95% confidence interval of (56%, 72%), which suggest that the difference between these studies is not due to statistical noise.

HUANG *et al.* (2002) suggest that the reason they observed more multistep changes than previous studies did may be that their analysis was of a large number of loci in the same families, whereas other studies generally pooled single-locus data from a large number of families. However, this cannot explain the discrepancy with our results, since the structure of their data set is similar to ours. Specifically, HUANG *et al.* (2002) analyzed 362 loci from 630 subjects from 53 pedigrees, whereas we analyzed 400 loci from 680 subjects from 123 pedigrees. The loci analyzed were also very similar, in both cases, being based on the ABI PRISM linkage mapping set version 2. It seems unlikely therefore that the discrepancy between their results and our lies in the structure of the data sets.

Length-dependent mutational bias and evolutionary equilibrium: Several previous studies have suggested that microsatellite mutations are biased toward expansion, reporting an excess of increases over decreases in microsatellite lengths (Amos and Rubinstzein 1996; Brinkmann *et al.* 1998; Primmer *et al.* 1998; Cooper *et al.* 1999; Ellegren 2000a; Kayser *et al.* 2000). By contrast our results (model 8) suggest that up mutations will exceed down mutations for microsatellites of <20 repeats, with the opposite true for microsatellites with >20 repeats (Figure 3). Length-dependent mutational bias has also been reported in human microsatellites by Ellegren (2000a), Kayser *et al.* (2000), and Xu *et al.* (2000), although they were not able to discriminate the dependence of bias on absolute microsatellite length. Instead they used a relative measure, standardizing length within each locus relative to the longest observed allele. Length-dependent mutational bias has also been reported in Drosophila (Harr and Schlotterer 2000) and yeast (Wierdl *et al.* 1997).

Demonstration of length-dependent mutational bias is particularly important in the context of the controversy regarding the nature of the factors that constrain microsatellite lengths. Early models of the slippage mutation process carried the implication that some microsatellite lengths would increase indefinitely over evolutionary time, but in practice lengths only very rarely exceed a few tens of repeats (Tautz 1993). Two principal theories have been advanced to account for this discrepancy (Ellegren 2000b). The first supposes that the frequency distributions of microsatellites represent a balance between the expansionary tendencies of slippage mutations and the contractions caused by point mutations breaking longer microsatellites into smaller units. The second supposes that longer microsatellites experience more contractions than expansions, so that there is a length-dependent mutation bias. Our results, together with those of Wierdl *et al.* (1997), Schlotterer *et al.* (1998), Ellegren (2000a), Harr and Schlotterer (2000), Xu *et al.* (2000), and Huang *et al.* (2002) suggest that length-dependent mutation bias exists, although it is of course possible that point mutation also plays a role. Note that the data analyzed here cannot identify point mutations, since these do not change the interprimer length of DNA.

CONCLUSIONS

We have introduced a likelihood-based procedure that allows formal comparison of competing models of microsatellite mutation. Application to data composed of 118,866 parent-offspring transmissions of AC microsatellites provides very strong evidence that the mutation rate of microsatellite loci is length dependent and some support for the hypothesis that the stationary distribution of microsatellites is maintained by mutational balance (Ellegren 2000b; Xu *et al.* 2000; Huang *et al.* 2002). In contrast to Huang *et al.* (2002), but in agreement with previous studies (Weber and Wong 1993; Amos and Rubinstzein 1996; Brinkmann *et al.* 1998; Ellegren 2000b; Kayser *et al.* 2000; Xu *et al.* 2000), our results suggest that mutation rate declines rapidly with change in number of repeat units, so that multistep mutations are much rarer than single-step changes.

Key advantages of the method presented here are that it avoids bias in parameter estimation due to unobserved mutations, it makes full use of the available data, and, by allowing comparisons between models, it is readily extended to investigate other aspects of microsatellite evolution. For example, if data from microsatellites with several repeat motifs were available it would be easy to add dependence of mutation rate on repeat motif to the models discussed here.

We thank Bill Amos for helpful comments on this work and for giving us the Excel macro used to convert allele sizes to number of repeats and Mark Beaumont for his insightful reading of an earlier draft of this article. Two referees made helpful comments on the original submission.

LITERATURE CITED

Amos, W., and D. C. Rubinstzein, 1996 Microsatellites are subject to directional evolution. Nat. Genet. **12:** 13–14.

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

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.

Brinkmann, B., M. Klintschar, F. Neuhuber, J. Huhne and B. Rolf, 1998 Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am. J. Hum. Genet. **62:** 1408–1415.

Cooper, G., N. J. Burroughs, D. A. Rand, D. C. Rubinsztein and W. Amos, 1999 Markov chain Monte Carlo analysis of human Y-chromosome microsatellite provides evidence of biased mutation. Proc. Natl. Acad. Sci. USA **96:** 11916–11921.

Cox, D. R., and D. V. Hinkley, 1974 *Theoretical Statistics.* Chapman & Hall, London.

Dib, C., S. Faure, C. Fizames, D. Samson, N. Drouot *et al.*, 1996 A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature **380:** 152–154.

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.

Ewen, K. R., M. Bahlo, S. A. Treloar, D. F. Levinson, B. Mowry *et al.*, 2000 Identification and analysis of error types in high-throughput genotyping. Am. J. Hum. Genet. **67:** 727–736.

Harbord, R. M., 2001 Modelling microsatellite evolution using directly observed mutations. M.Sc. Dissertation, University of Reading, Reading, UK.

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

Huang, Q.-Y., F.-H. Xu, H. Shen, H.-Y. Deng, Y.-J. Liu *et al.*, 2002 Mutation patterns at dinucleotide microsatellite loci in humans. Am. J. Hum. Genet. **70:** 625–634.

Kayser, M., L. Roewer, M. Hedman, L. Henke, J. Henke *et al.*, 2000 Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. Am. J. Hum. Genet. **66:** 1580–1588.

Kruglyak, L., M. J. Daly, M. P. Reeve-Daly and E. S. Lander, 1996 Parametric and non-parametric linkage analysis: a unified approach. Am. J. Hum. Genet. **58:** 1347–1363.

Kruglyak, S., R. Durrett, D. Schug and C. 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.

Kruglyak, S., R. Durrett, D. Schug and C. Aquadro, 2000 Distribution and abundance of microsatellites in the yeast genome can be explained by a balance between slippage events and point mutations. Mol. Biol. Evol. **17:** 1210–1219.

Leeflang, E. P., L. Zhang, S. Tavare, R. Hubert, J. Srinidhi *et al.*, 1995 Single sperm analysis of the trinucleotide repeats in the Huntington's disease gene: quantification of the mutation frequency spectrum. Hum. Mol. Genet. **4:** 1519–1526.

O'Connell, J., and D. Weeks, 1998 PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am. J. Hum. Genet. **63:** 259–266.

Ohta, T., and M. Kimura, 1973 The model of mutation appropriate to calculate the number of electrophoretically detectable alleles in a genetic population. Genet. Res. **22:** 201–204.

Primmer, C. R., N. Saino, A. P. Moller and H. Ellegren, 1998 Unraveling the processes of microsatellite evolution through analysis of germ line mutations in barn swallow *Hirundorustica*. Mol. Biol. Evol. **15:** 1047–1054.

Schlotterer, C., R. Ritter, B. Harr and G. Brem, 1998 High mutation rate of a long microsatellite allele in Drosophila melanogaster provides evidence for allele-specific mutation rates. Mol. Biol. Evol. **15:** 1269–1274.

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

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

Tautz, D., 1993 DNA fingerprinting: state of the science, pp. 21–28 in *DNA Fingerprinting: State of the Science*, edited by S. D. J. Pena, R. Chakraborty, J. T. Epplen and A. J. Jeffreys. Birkhauser, Basel, Switzerland.

Venables, W. N., and B. D. Ripley, 1999 *Modern Applied Statistics with S-PLUS*. Springer, New York.

Weber, J. L., and C. Wong, 1993 Mutation of human short tandem repeats. Hum. Mol. Genet. **2:** 1123–1128.

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

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

Communicating editor: Y.-X. Fu