Models of Microsatellite Evolution

Peter Calabrese[1] and Raazesh Sainudiin[2]

[1] University of Southern California, Los Angeles, California petercal@usc.edu
[2] Cornell University, Ithaca, New York rs228@cornell.edu

1 Introduction

Microsatellites are simple sequence repeats in DNA, for example the motif AT repeated twenty-five times in a row. Microsatellites mutate by changing the number of their repeats, for example the $(AT)_{25}$ mentioned in the previous sentence might become an $(AT)_{24}$ or $(AT)_{26}$ in that individual's offspring. These length-changing mutations occur at rates several orders of magnitude higher than point mutations. The reason for microsatellite's popularity as genetic markers is that their high mutation rates make them highly polymorphic, and they are relatively dense in the genomes of many organisms. For a review see the article by Ellegren [16], and for a collection of articles see the book edited by Goldstein and Schlötterer [23]. Ellegren [16] succinctly wrote, "simple repeats do not evolve simply." In this chapter, then, we will discuss many different models for microsatellite evolution.

Researchers have exploited microsatellites for many purposes. They are commonly used in the construction of genome-wide maps, in the search for disease-causing genes, and in identification for both forensics applications and paternity tests. As an example, the controversy over whether Thomas Jefferson fathered a child with his slave Sally Hemings was rekindled by a microsatellite study [19]. In these applications, two individuals are considered closely related if a large percentage of the microsatellite markers studied have the same number of repeats. However, microsatellites in different individuals are not just the same or different, they can differ by just a few repeat units or by many repeat units. Because pedigree experiments have shown that most mutations are a change in one repeat unit (85% in [54], 78% in [50]), some researchers have used microsatellites as molecular clocks. By studying the average number of repeat differences in many microsatellite loci, one can infer the time to the most recent common ancestor of two individuals. Microsatellites have been used to estimate the age of non-microsatellite mutations, for example, the $CCR5-\Delta 32$ AIDS-resistance allele [46]. In cancer research, hyper-mutable microsatellites with deficient DNA mismatch repair systems have been used

to reconstruct tumor progression [48]. Microsatellites have been used to infer selective sweeps (for a recent review see *e.g.* [41]), demographic history ([22], [12], [37], [2]), and population structure ([38], [17]).

The vast majority of microsatellites in higher organisms are believed to evolve neutrally, *i.e.*, there is no selection pressure on the number of repeats. Nonetheless, some microsatellites exist in promoter regions and may be sites for protein binding, or near such sites. In this case, the number of repeats in these microsatellites has an effect on transcription and the degree of protein binding [29]. Further, other microsatellites play a direct role in such human diseases as Fragile X syndrome, myotonic dystrophy, and Huntington's disease; these diseases are caused by trinucleotide microsatellites at specific locations expanding beyond a disease-specific threshold [39].

The predominant mechanism by which microsatellites mutate is believed to be replication slippage [15]. When DNA replicates, the two strands sometimes disassociate. In non-repetitive DNA, the strands reassociate the same way they were before the slippage event, with matching base pairs on the opposing strands. But in repetitive DNA, since there are so many possible matching base pair alignments, sometimes the strands realign differently, forming an unmatched loop on one of the strands. Then when the two strands completely disassociate and begin replication anew the strand which had the loop will contain a longer microsatellite than the opposing strand. The microsatellite on the template strand will always have the same length before and after the slippage event. If the loop is on the template strand, then the microsatellite on the replicating strand will be shorter; and if the loop is on the replicating strand, then the microsatellite on its side will be longer. For a figure of this process see [15], p. 38. These primary mutations, which depend exclusively on the DNA replication machinery, occur at rates several orders of magnitude higher than the observed mutation rate and are countered by the DNA mismatch repair machinery (for a recent review see [42]). Thus the observed mutations are those replication slippage events that have escaped repair.

Since longer microsatellites present more opportunity for slippage, we would expect mutation rates to increase as a function of microsatellite length; this prediction is experimentally supported [53]. Some microsatellites are interrupted, for example $(AT)_{12}GT(AT)_7$. Since these interruptions allow fewer realignments after a possible slippage event, we would expect interrupted microsatellites to have lower mutation rates; and this is also experimentally supported [36].

Several other factors are also known to be associated with the heterogeneity in mutation rates across microsatellite loci. Dinucleotides have a lower mutation rate than tetranucleotides (see Table 1 of [16]). Moreover, different dinucleotide motifs have a strikingly different length distribution in the human genome [8], possibly due to motif-specific differences in the efficacy of mismatch-repair [25]. A significant number of uninterrupted compound repeats ($> 30,000$) such as $(\mathrm{TG})_m$-$(\mathrm{TA})_n$, with both m and $n \geq 5$ repeat units, occur in the human genome (Sainudiin and Durrett unpublished re-

sults); their evolutionary dynamics are complex [5] and not well understood. A further complication to measuring microsatellite variability is that insertions/deletions in the flanking regions can also affect the PCR fragment length (see *e.g.* [24]).

2 Models

The oldest model for microsatellite evolution is the stepwise mutation model originally proposed by Ohta and Kimura [35] for electrophoretic alleles. In this model the number of repeat units is equally likely to increase or decrease by one at a rate independent of the microsatellite's length, subject to the constraint that the number of repeat units cannot become smaller than one. Let X be the length of the microsatellite, then

$$\begin{aligned} &X \to X+1 \text{ at rate } \gamma, \text{ and} \\ &X \to X-1 \text{ at rate } \gamma \end{aligned} \tag{1}$$

This is a symmetric random walk with a lower boundary condition. Numerous complications to this basic model have been introduced.

The first complication we will discuss is allowing the mutation rate to depend on the microsatellite's length. For example, Kruglyak *et al.* [30] proposed a proportional slippage model where the mutation rate increases linearly with the microsatellite's length

$$\begin{aligned} &X \to X+1 \text{ at rate } b(X-1), \text{ and} \\ &X \to X-1 \text{ at rate } b(X-1) \end{aligned} \tag{2}$$

Sibly, Whittaker, and Talbot [44] proposed a model with an additional constant term

$$\begin{aligned} &X \to X+1 \text{ at rate } b_0 + b_1(X-1), \text{ and} \\ &X \to X-1 \text{ at rate } b_0 + b_1(X-1) \end{aligned} \tag{3}$$

This constant term is analogous to the "indel slippage" term in [10]. Calabrese, Durrett, and Aquadro [7] further extended this model to prevent microsatellites shorter than a threshold κ from mutating

$$\begin{aligned} &X \to X+1 \text{ at rate } b(X-\kappa)^+, \text{ and} \\ &X \to X-1 \text{ at rate } b(X-\kappa)^+ \end{aligned} \tag{4}$$

(where $(X-\kappa)^+ = \max(X-\kappa, 0)$). The symmetric random walk models do not have a stationary distribution on their countable state space [32]. Nauta and Weissing [34] proposed a finite alleles version of the stepwise mutation model by imposing range constraints with reflecting boundaries to assure a uniform stationary distribution (also see [18]).

Most, but not all, observed microsatellite mutations are by one repeat unit. Therefore Di Rienzo *et al.* [11] proposed a model which allows for larger

mutations. With probability p a mutation is one repeat unit, and with probability $1-p$ the mutation could be larger. In their formulation, the one-step mutations followed the stepwise mutation model and the larger mutations were equally likely to be contractions or expansions, with the magnitude of these mutations following a truncated geometric distribution.

Another complication is to allow the mutation rates to be asymmetric. Walsh [49] proposed a linear birth death chain, *i.e.*, a proportional slippage model where the mutation rate increases linearly with the microsatellite's length in the presence of biased contractions

$$\begin{array}{l} X \to X+1 \text{ at rate } bX \\ X \to X-1 \text{ at rate } dX \end{array} \tag{5}$$

for $X \in \{2, 3, \cdots, \infty\}$ and $1 \to 2$ at a much smaller birth rate ν. He showed that a stationary distribution exists for this model when $d/b > 1$ (see also [47]). Fu and Chakraborty [20] proposed a model which allows larger mutations according to a geometric distribution in the presence of a constant mutational bias. Calabrese and Durrett [8] generalized the models described earlier to asymmetric linear and quadratic models: for the linear model

$$\begin{array}{l} X \to X+1 \text{ at rate } u_0 + u_1(X-\kappa)^+, \text{ and} \\ X \to X-1 \text{ at rate } d_0 + d_1(X-\kappa)^+ \end{array} \tag{6}$$

and for the quadratic model

$$\begin{array}{l} X \to X+1 \text{ at rate } u_0 + u_1(X-\kappa)^+ + [u_2(X-\kappa)^+]^2, \text{ and} \\ X \to X-1 \text{ at rate } d_0 + d_1(X-\kappa)^+ + [d_2(X-\kappa)^+]^2 \end{array} \tag{7}$$

The expansion and contraction rates can take the same parametric form with distinct parameters as above, or take different parametric forms as well. Xu *et al.* [54] suggested that the expansion rate be independent of microsatellite length, while the contraction rate increase exponentially with microsatellite length. Using an approximation to the Ornstein-Uhlenbeck process, Garza, Slatkin, and Freimer [21] proposed that microsatellites have a focal length in the sense that microsatellites shorter than this length have a bias upwards whereas longer microsatellites have a bias downwards (also see [57]). These models can also allow larger mutations by specifying the expectation and variance of the size of mutations and thus nest the stepwise mutation model and the model due to Di Rienzo *et al.* [11] within them. While most of the previously described asymmetric models do not stipulate this focal property *a priori*, when these models are fit to data, generally their parameter estimates do satisfy this property.

Two recent studies attempt to capture several features of microsatellite evolution just described. Whittaker *et al.* [52] proposed a class of models with the following transition rates from microsatellite length $X = i$ to length j

$$q_{ij} = \begin{cases} \gamma_u e^{\alpha_u i} e^{-\lambda_u (j-i)}, i < j \\ \gamma_d e^{\alpha_d i} e^{-\lambda_d (i-j)}, i > j \end{cases} \tag{8}$$

Sainudiin [40] proposed another class of models

$$q_{ij} = \begin{cases} \mu(1+(i-\kappa)s)[u-v(i-\kappa)]_0^1 m(1-m)^{|i-j|-1}, i<j \\ \mu(1+(i-\kappa)s)\{1-[u-v(i-\kappa)]_0^1\}\frac{m(1-m)^{|i-j|-1}}{1-(1-m)^{i-\kappa}}, i>j \end{cases} \quad (9)$$

In equation (9), the notation means

$$[\alpha]_0^1 = \begin{cases} 1, \text{ if } \alpha > 1 \\ 0, \text{ if } \alpha < 0, \text{ and} \\ \alpha, \text{ otherwise} \end{cases} \quad (10)$$

Both classes of models allow the mutation rates to increase with microsatellite length, the bias to change as a function of microsatellite length, and larger mutations to have a geometrical distribution. However, the parametric forms of these models differ.

The final model complication we will consider is point mutations. Point mutations can interrupt a microsatellite, for example transforming $(AT)_{20}$ into $(AT)_{12}GT(AT)_7$. Since most researchers measure the length of microsatellites without sequencing they would not detect this transformation. Bell and Jurka [3] proposed that such point mutations constrain the growth of microsatellites. Kruglyak *et al.* [30] proposed a model with two processes

1. single-step proportional slippage (described above): $X \to X+1$ at rate $b(X-1)$ and $X \to X-1$ at rate $b(X-1)$, and
2. point mutations: $X \to j$ where $j < X$ at rate a.

Thus a point mutation is uniformly likely to affect any of the repeat units. Durrett and Kruglyak [14] showed that this model has a stationary distribution. Sibly, Whittaker, and Talbot [44] and Calabrese and Durrett [8] followed this paradigm of considering a slippage process and a point mutation process, but they considered more general slippage processes. (Now that we are considering interrupted microsatellites the state space is more complicated. The studies referenced above chose counting schemes to limit the state space to one dimension, but they all chose to do this in different ways and this has been the source of some confusion. For every point in the genome, [30] and [14] counted all uninterrupted repeats to the left. The other studies did not consider every point in the genome. [44] counted only the left half of an interrupted repeat, and [8] counted only uninterrupted repeats.)

One final caveat is in order, many microsatellite models have been proposed. We believe this summary captures most of the important concepts, but we do not claim to be exhaustive.

3 Experiments and Analysis

One of the statistical tools used in this section is the Akaike Information Criterion (AIC) [1]. The formula for computing the AIC score for a model is simple

$$\text{AIC} = -2\log(\text{maximum likelihood}) + 2(\text{number parameters}) \qquad (11)$$

Given a list of models, we compute the AIC score for each model and choose the models with the lowest scores. This scheme has the advantage over the likelihood ratio test and parametric bootstrap (see, *e.g.*, [4] and [28]) of being able to select from a large set of models without considering all pairwise comparisons. The AIC score is intuitive because the best models should have high likelihoods and models are penalized for having a large number of parameters. But this scoring system also has a firm statistical foundation. The book by Burnham and Anderson [6] discusses the model selection problem in general and also presents a heuristic justification for the AIC scoring system, p. 239-247.

There are several types of data sets to consider when studying microsatellites. The first is pedigree data. Two of the largest such data sets (both in humans) were by Xu *et al.* [54] and Huang *et al.* [26]. Xu *et al.* [54] observed that the rate of expansions is independent of microsatellite "length" but that the rate of contractions increases exponentially as a function of microsatellite "length". Huang *et al.* [26] found a statistically significant negative relationship between the magnitude and direction of mutation and "length". In the two preceding sentences we have put the word length in quotations, because both groups of researchers did not measure the actual length of a microsatellite but rather the total length of the polymerase chain reaction (PCR) product that consists of the microsatellite and a variable amount of flanking sequence. They then applied the inverse of the distribution function of the observed lengths to obtain a number in $[0, 1]$ which they called the "length". This near universal practice of measuring the PCR product length rather than the actual number of repeat units has complicated modeling efforts.

In another large human pedigree study, Whittaker *et al.* [52] has taken the further step of using the human genome sequence and the primer sequence to infer the number of repeat units from the PCR product length. While this method cannot tell whether an individual microsatellite has been interrupted by point mutations, it is an important advance over simply using PCR product length. They measured 118,866 parent-offspring transmissions of AC repeats and observed 53 mutations, for a mutation rate of 4.5×10^{-4} per generation. Approximately 72% of the mutations were of one repeat unit. The mutation rate clearly increased super-linearly with the repeat length (see Figure 2 in [52]). They used the AIC scoring system to compare models of the class in equation (8). The cases where mutation rate increases with microsatellite length ($\alpha > 0$) were significant improvements over the cases where mutation rate was independent of microsatellite length ($\alpha = 0$). The best model in this class had asymmetric γ and α terms ($\gamma_u \neq \gamma_d$ and $\alpha_u \neq \alpha_d$) implying a mutation rate bias, but a symmetric λ term implying the distribution of the size of the larger mutations was symmetric. The estimated parameters implied that microsatellites shorter than twenty repeat units had a bias towards ex-

pansions and longer microsatellites had a bias towards contractions. However, there were large confidence intervals for all of the parameter estimates.

Another type of data set is *in vitro* experiments. During PCR, microsatellites are duplicated and there are opportunities for slippage just as in *in vivo* cell division. In single-molecule PCR experiments, Shinde *et al.* [43] found that slippage rates increase linearly with microsatellite length and there is a threshold of eight basepairs below which microsatellites do not appear to slip. For all microsatellite lengths they found a higher rate for contractions than expansions (fourteen times greater for AC microsatellites and five times greater for poly-A microsatellites). There are thermodynamic reasons to expect this asymmetry *in vitro* (see references in [43]). Clearly there are differences, however, between these *in vitro* experiments with *Taq* DNA polymerase and no mismatch repair system and *in vivo* cell division. Another set of related experiments studies microsatellite mutations *in vivo* but in organisms whose mismatch repair system has been knocked out. For an example in mice see [55] and in yeast see [36]. There are many more microsatellite mutations in individuals with deficient mismatch repair systems, and this is informative for studying microsatellite models, but in addition to the rate the pattern of mutations also appears to be different in these individuals.

In population data, many unrelated individuals are typed at numerous microsatellite loci. Nielsen [33] suggested using such data sets and likelihood ideas for model selection. The problem with this approach is that assumptions must be made about the genealogies of individuals. These assumptions will in turn affect the evaluation of the models.

Another type of data set is genome data. There are now complete (or nearly complete) genome sequences available for many species. For each such species, we thus have the length distribution of all microsatellites in one idealized individual. If we assume this distribution is at equilibrium and we consider models that have a stationary distribution then we can fit these models to genome data. All the references we will discuss assume that microsatellites are "born" at some minimum length. Kruglyak *et al.* [30] fit the proportional slippage (equation (2)) with point mutation model to the then available genome sequence of humans, mice, fruit fly, and yeast. They later fit this model to the complete genome sequence of yeast [31]. Assuming different microsatellites evolve independently, Sibly, Whittaker, and Talbot [44] then used likelihood ideas to compare different models of microsatellite evolution. They considered symmetric slippage models of the form in equation (3) with a point mutation process and found support for the parameters $b_0 \neq 0$ and $b_1 \neq 0$.

Calabrese and Durrett [8] used genomic data and the AIC scoring system to consider many different slippage models, including most of those then in the literature. They considered general slippage processes with a uniform point mutation process as described in the previous section. The data they considered was moderately-spaced dinucleotide microsatellites uninterrupted by point mutations. They found the asymmetric models explained the genome data significantly better than the symmetric models. One of the best models

had asymmetric quadratic slippage (equation (7)), where the parameters were such that dinucleotide microsatellites with length longer than twenty-five repeat units had a bias toward contractions. Moreover, for humans (but not Drosophila), they found that the different dinucleotide motifs had strikingly different distributions, as shown in Figure 1.

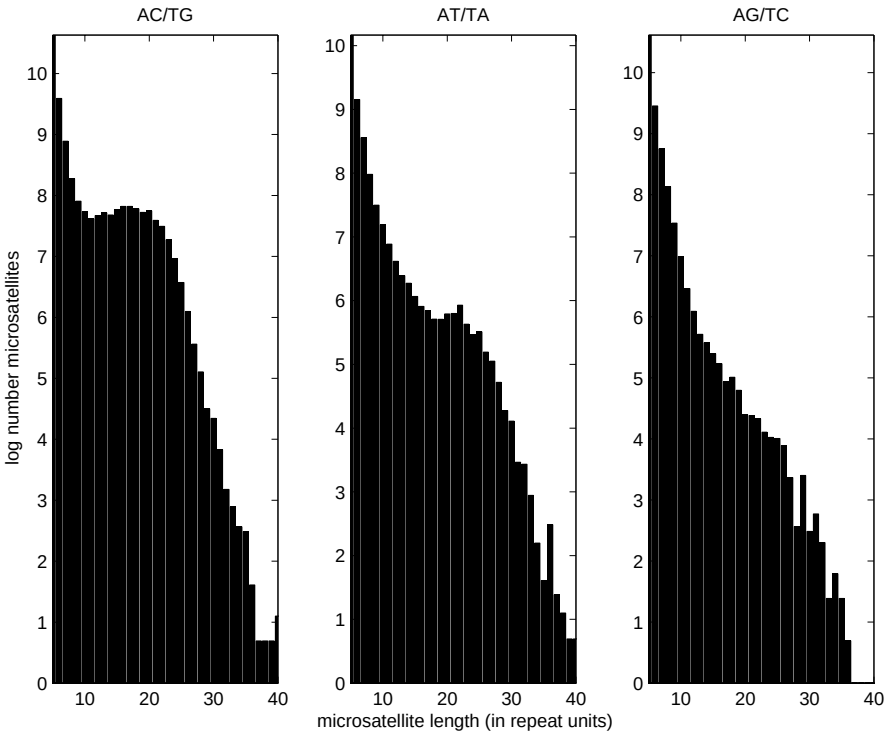

Fig. 1. Separated by motif, the natural logarithm of one plus the number of dinucleotide microsatellites of different lengths in the human genome.

Calabrese and Durrett [8] exploited a connection with queueing theory in order to calculate the stationary distribution. In the language of continuous time Markov chains, each model was specified by a set of exponential holding times $\mu(j)$ for microsatellites of length j, and the probabilities $p(j,i)$ a microsatellite of length j will next mutate to length i. The total number of microsatellites in the genome was modeled as a network of queues (specifically $M/M/\infty$ queues, where in the usual queueing theory terminology microsatellites correspond to customers, microsatellite lengths correspond to stations, and at each length or station there are an infinite number of servers); when a microsatellite is interrupted by a point mutation it exits the network. Since all microsatellites have a positive probability of leaving the network, there exists

a stationary distribution. Define arrival rates

$$r(\kappa) = \lambda + \sum_i r(i)p(i,\kappa) \tag{12}$$

$$r(j) = \sum_i r(i)p(i,j), j > \kappa \tag{13}$$

where λ is a scaling parameter that is the rate microsatellites are born at the minimum considered length κ. Then the stationary distribution is for all j the number of microsatellites with length j are independent and Poisson distributed with mean $r(j)/\mu(j)$ (see *e.g.* [13], p.192). Let $l(j)$ be the number of microsatellites of length j in the genome, and define the normalizing constant $Z = \sum_j r(j)/\mu(j)$. Since they assumed moderately-spaced microsatellites evolve independently, conditioning on the number of microsatellites, the likelihood of the data is

$$\prod_j \left(\frac{r(j)}{Z\mu(j)} \right)^{l(j)} \tag{14}$$

For each model, they numerically solved the linear system of equations (12), (13) to determine the arrival rates $r(j)$, numerically maximized the likelihood of the genome (14) over the space of model parameters, and computed the AIC score.

The final type of data set we will consider collects microsatellites from two closely-related populations or species. Since longer microsatellites are more mutable and hence more useful to experimentalists, when microsatellites are selected in one species and then compared in another species there is an ascertainment bias. Two studies that avoid this problem are Cooper, Rubinsztein, and Amos [9] and Webster, Smith, and Ellegren [51]. Despite accounting for this ascertainment bias, Cooper, Rubinsztein, and Amos [9] found that human dinucleotide microsatellites are significantly longer than their chimpanzee orthologues. Webster, Smith, and Ellegren [51] considered many more microsatellites and concurred with [9]'s findings. They also found that human mononucleotide microsatellites are more likely to be *shorter* than their chimpanzee counterparts. Webster, Smith, and Ellegren [51] selected an unbiased sample of AC dinucleotide microsatellites from a region of genomic DNA and compared the orthologues in humans and chimpanzees.

Sainudiin [40] followed this strategy and used the AIC scoring system to compare slippage models of the form in equation (9). They initially assumed that the same microsatellite model and parameters applied both to the human and chimpanzee lineages. They concluded $s \neq 0$, so longer microsatellites are more mutable, and $v \neq 0$, so there is a bias term that depends linearly on the microsatellite's length. The estimated parameters imply microsatellites shorter than eighteen repeat units have a bias towards expansions, while longer microsatellites have a bias towards contractions. When they relaxed the assumption that the same model parameters applied in both the human and

chimpanzee lineages, they found that this focal length increased to twenty-one repeats in humans while remaining at eighteen repeats in chimpanzees, further confirming the findings in [9] and [51].

Sainudiin [40] considered three Markov chains, one each on the ancestral, human, and chimpanzee lineages. These three Markov chains had rate matrices $\mathbf{Q}^{(a)}$, $\mathbf{Q}^{(h)}$, and $\mathbf{Q}^{(c)}$, specified by equation (9), possibly with different parameters. Let λ_h and λ_c be the branch lengths of the human and chimpanzee lineages respectively. Then the transition probability matrices $\mathbf{P}^{(h)}(\lambda_h)$ and $\mathbf{P}^{(c)}(\lambda_c)$, were obtained by matrix exponentiating the product of the corresponding rate matrix and branch length, *e.g.*, $\mathbf{P}^{(h)}(\lambda_h) = \exp\{\mathbf{Q}^{(h)}\lambda_h\}$. The stationary distribution of the ancestral Markov chain $\mathbf{X}^{(a)}$ was denoted by $\boldsymbol{\pi}^{(\boldsymbol{a})}$. The data considered was N homologous microsatellite lengths (H_i, C_i) in the human and chimpanzee genomes. Then the likelihood of the data is

$$\prod_{i=1}^{N} \sum_{j=\kappa}^{\Omega} \pi_j^{(a)} \; P_{j,C_i}^{(c)}(\lambda_c) \; P_{j,H_i}^{(h)}(\lambda_h). \tag{15}$$

Since the ancestral state is unknown, the likelihood can be thought of as a weighted sum over all possible ancestral states, where the weights come from the stationary distribution of the ancestral chain. The product term comes from the assumption of independence among the N loci. For each model, Sainudiin [40] numerically optimized the likelihood (15) over the space of model parameters, and computed the AIC score.

4 Discussion

We have discussed numerous microsatellite models. There is evidence from a number of different sources that the best models have the following properties

1. long microsatellites are more likely to mutate, and
2. long microsatellites have a bias towards contractions, while
3. short microsatellites have a bias towards expansions.

For dinucleotide repeats in humans this focal length apppears to be around twenty repeat units. Moreover all the model parameters depend on both the length and composition of the repeat motif. In our opinion, it is still unclear what the parametric form of the "best" model is.

We believe that the best type of data set to determine this model is pedigree data where the actual number of repeat units has been inferred (rather than just using the PCR fragment length) as in Whittaker *et al.* [52]. It would be interesting to infer the number of repeat units and re-analyze the data sets in Xu *et al.* [54] and Huang *et al.* [26]. The main advantage of using genome data is that it is plentiful. The disadvantage is that we do not observe mutations, but rather the stationary distibution; and distinct models may have

very similar stationary distributions. For example, this makes it difficult to determine the percentage of large mutations using genome data. Further, Sibly *et al.* [45] specifically investigated the distribution of interrupted repeats and found the existing slippage/point mutation models inadequate to explain the data.

One question is whether the choice of model matters. For example, let us consider using the statistic $(\delta\mu)^2$ to measure the time to the most recent common ancestor of two individuals. Let X_i and Y_i be the microsatellite lengths at the ith locus in the two individuals; define the statistic

$$(\delta\mu)^2 = \sum_{i=1}^{I} (X_i - Y_i)^2 / I$$

as the average over I loci. Goldstein *et al.* [22] showed that for the stepwise mutation model $E(\delta\mu)^2(t) = 2\gamma t$, where t is the time to the most recent common ancestor and γ is the mutation rate. For the more complicated models that we have discussed it is unlikely that there is such a simple formula, but we can simulate these models.

Let us compare the mutation model in equation (8) and the stepwise mutation model. One difficulty with length dependent mutation models, that we do not encounter with the stepwise mutation model, is that we have to make additional assumptions about the length of the ancestor. Since the model in equation (8) has a stationary distribution, let us assume that the common ancestor has a length chosen randomly from this stationary distribution. Further we can use this stationary distribution to find the average mutation rate for a microsatellite chosen randomly from this distribution. In order to fairly compare models, let us set the mutation parameter in the stepwise mutation model equal to this average. For the model parameters estimated in [52], this average mutation rate is $\gamma_1 = 1 \times 10^{-4}$. This rate is smaller than both the observed rate in [52] (4.5×10^{-4}) and in other dinucleotide studies (*e.g.* 5.6×10^{-4} in [22]). Consequently we also consider the model in equation (8) where the γ parameters have been increased to match the average mutation rate $\gamma_2 = 5 \times 10^{-4}$. When we increase the γ parameters, we preserve the estimated ratio γ_u/γ_d, and the α and λ parameters; these new γ parameters are still well within the confidence intervals estimated from the data. Likewise we consider the stepwise mutation model with elevated rate $\gamma_2 = 5 \times 10^{-4}$.

Figure 2 shows the mean of $(\delta\mu)^2$ for these two models as a function of time. The left-hand plot has average mutation rate $\gamma_1 = 1 \times 10^{-4}$, and the right-hand plot is a re-scaling with average mutation rate $\gamma_2 = 5 \times 10^{-4}$. For the left-hand plot at times less than twenty-five thousand generations, the two models are in good agreement. For greater times the two models diverge; this is because under the stepwise mutation model the $(\delta\mu)^2$ statistic continues to grow linearly while for the models with a stationary distribution this statistic eventually plateaus. For the right-hand plot since the mutation rate is five times greater, the two models start to diverge about five times earlier

at around five thousand generations. If we are interested in short divergence times then the stepwise mutation model seems a reasonable approximation. If we are interested in divergence times that are too long and we believe the true microsatellite mutation model has a stationary distribution, then microsatellites will not be useful because any divergence statistic will eventually plateau due to this stationarity.

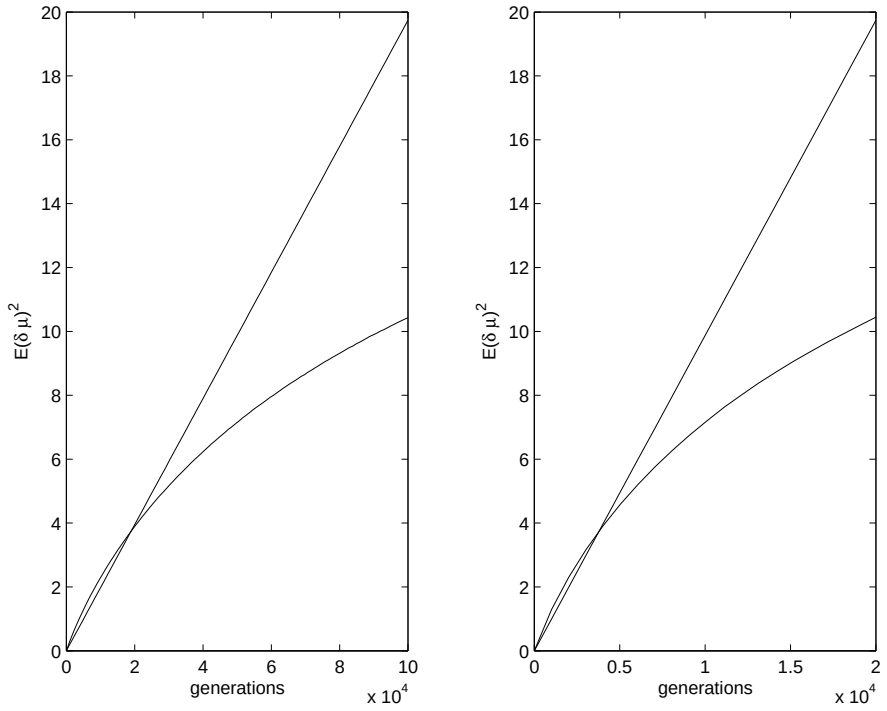

Fig. 2. The mean of the $(\delta\mu)^2$ statistic as a function of time in generations. The linear curve is for the stepwise mutation model, the nonlinear curve is for the slippage model in equation (8). The left-hand plot has average mutation rate $\gamma_1 = 1 \times 10^{-4}$, and the right-hand plot is a re-scaling with average mutation rate $\gamma_2 = 5 \times 10^{-4}$. Since the right-hand plot has mutation rate five times greater, the two models start to diverge five times earlier.

The times for which the models diverge are germane to the study of human populations. If we model the genealogy of unrelated individuals with the neutral coalescent (see *e.g.* [27]), and assume the commonly used estimate of ten thousand for the effective population size of humans, then the average time to the most recent common ancestor of two individuals is twenty thousand generations, and the average time to the most recent common ancestor of a large sample is forty thousand generations. Assuming the stepwise muta-

tion model, various statistics of microsatellite lengths have been used to infer aspects of demographic history (for *e.g.* [56], [12], [37], [58]). We have simulated the coalescent process with effective population size ten thousand and sample size fifty, and alternately used the stepwise mutation model and the model in equation (8) with the two parameter sets discussed in the previous paragraphs. In Table 1, we show the median (5% quantile, 95% quantile) for several summary statistics. We can see that such statistics and related tests will be dependent on the mutation model used.

Table 1. The median (5% quantile, 95% quantile) of the sample variance, homozygosity, and number alleles for the stepwise mutation model (SMM) and the model in equation (8) (WHB) at two average mutation rates. These values were simulated using the coalescent with effective population size ten thousand, and sample size fifty.

	$\gamma_1 = 1 \times 10^{-4}$			$\gamma_2 = 5 \times 10^{-4}$		
Model	sam.var.	homo.	num.all.	sam.var.	homo.	num.all.
SMM	1.2 (0.3, 5.9)	0.3 (0.2, 0.6)	5 (3, 7)	6.2 (1.8, 29.)	0.2 (0.1, 0.3)	9 (6,13)
WHB	1.9 (.02, 11.)	0.4 (0.2, 1.0)	5 (2, 10)	7.0 (1.0, 20.)	0.2 (0.1, 0.5)	10 (4, 14)

References

1. Akaike, H.: A new look at the statistical model identification. IEEE Trans. on Automatic Control, **19:** 716–723 (1974)
2. Beaumont, M. : Detecting population expansion and decline using microsatellites. Genetics **153:**2013–2029 (1999)
3. Bell, G.I., Jurka, J.: 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 (1997)
4. Bickel P.J., Doksum, K.A.: Mathematical Statistics. Prentice Hall, New Jersey (1977)
5. Bull, L. N. , Pabón-Peña C. R. , Nelson B. Freimer: Compound microsatellite repeats: practical and theoretical features. Genome Research **9:**830–838 (1999)
6. Burnham, K.P., Anderson, D.R.: Model Selection and Inference. Springer, New York (1998)
7. Calabrese, P.P., Durrett, R.T., Aquadro, C.F.: Dynamics of microsatellite divergence and proportional slippage/ point mutation models. Genetics **159:** 839–852 (2001)
8. Calabrese, P., Durrett, R.: Dinucleotide repeats in the Drosophila and human genomes have complex, length-dependent mutation processes. Mol. Biol. Evol. **20:** 715–725 (2003)
9. Cooper, G., Rubinsztein, D.C., Amos, W.: Ascertainment bias cannot entirely account for human microsatellites being longer than their chimpanzee homologues. Human Molecular Genetics **7:** 1425–1429 (1998)

10. Dieringer, D., Schlötterer, C.: Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. Genome Research **13:** 2242–2250 (2003)
11. Di Rienzo, A., Peterson, A.C., Garza, J.C., Valdes, A.M., Slatkin, M., *et al.:* Mutational processes of simple-sequence repeat loci in human populations. Proc. Natl. Acad. Sci. USA **91:** 3166–3170 (1994)
12. Di Rienzo, A. , Donnelly, P. , Toomajian, C. , Sisk, B. , Hill, A. , Petzl-Erler, M. L. , Haines, G. K. , Barch, D. H. : Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. Genetics **148:**1269–1284 (1998)
13. Durrett, R.: Essentials of stochastic processes. Springer, New York (1999)
14. Durrett, R., Kruglyak, S.: A new stochastic model of microsatellite evolution. J. Apl. Prob. **36:** 621–631 (1999)
15. Eisen, J.A.: Mechanistic basis for microsatellite instability. In: Goldstein, D.B., Schlötterer, C. (ed) Microsatellites. Evolution and Applications. Oxford University Press, Great Britain (1999)
16. Ellegren, H.: Microsattelite mutations in the germ line: Implications for evolutionary inference. TIG **16:** 551–558 (2000)
17. Falush, D. , Stephens, M., Pritchard, J. K. : Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics **164:**1567–1587 (2003)
18. Feldman, M.W., Bergman, A., Pollock, D.D., Goldstein, D.B.: Microsatellite genetic distances with range constraints: Analytic description and problems of estimation. Genetics **145:** 207–216 (1997)
19. Foster, E.A., Jobling, M.A., Taylor, P.G., Donnelly, P., de Knijff, P., Mieremet, R., Zerjal, T., Tyler-Smith, C.: Jefferson fathered slave's last child. Nature **396:** 27–28 (1998)
20. Fu, Y. Chakraborty, R. : Simultaneous estimation of all the parameters of a step-wise mutation model. Genetics **150:**487–497 (1998)
21. Garza, J.C., Slatkin, M., Freimer, N.B.: Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. Mol. Biol. Evol. **12:** 594–603 (1995)
22. Goldstein, D.B., Ruiz-Linares, A., Cavalli-Sforza, L.L., Feldman, M.W. Genetic absolute dating based on microsatellites and modern human origins. Proc. Natl. Acad. Sci. USA **92:** 6723–6727 (1995)
23. Goldstein, D.B., Schlötterer, C. (editors): Microsatellites. Evolution and Applications. Oxford University Press, Great Britain (1999)
24. Grimaldi, M. C. , Crouau-Roy, B. : Microsatellite allelic homoplasy due to variable flanking sequences. Jnl. Mol. Evol. **44(3):**336–334 (1997)
25. Harr, B. Todorova, J. and Schlötterer, C. : Mismatch repair-driven mutational bias in *D. melanogaster*. Mol. Cell **10:**199–205 (2002)
26. Huang, Q-Y., Xu, F-H., Shen, H., Deng, H-Y., Liu, Y-J., Liu, Y-Z., Li, J-L., Recker, R.R., Deng,H-W.: Mutational patterns at dinucleotide microsatellite loci in humans. Am. J. Hum. Genet. **70:** 625-634 (2002)
27. Hudson, R.R.: Gene genealogies and the coalescent process. Oxford Surveys Evol. Biol. **7:**1–44 (1990)
28. Huelsenbeck, J.P., Rannala, B.: Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. Science **276:** 227 – 232 (1997)

29. Kashi, Y., Soller, M.: Functional roles of microsatellites and minisatellites. In: Goldstein, D.B., Schlötterer, C. (ed) Microsatellites. Evolution and Applications. Oxford University Press, Great Britain (1999)
30. Kruglyak, S., Durrett, R., Schug, M.D., Aquadro, C.F.: Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc. Natl. Acad. Sci. USA **95:** 10774–10778 (1998)
31. Kruglyak, S., Durrett, R., Schug, M.D., Aquadro, C.F.: 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 (2000)
32. Moran, P.A.P.: Wandering distributions and the electrophoretic profile. Theoretical Population Biology **8:** 318–330 (1975)
33. Nielsen, R.: A likelihood approach to populations samples of microsatellite alleles. Genetics **146:** 711–716 (1997)
34. Nauta, M.J., Weissing, F.J.: Constraints on allele size at microsatellite loci: implications for genetic differentiation. Genetics **143:**1021–1032 (1996)
35. Ohta, T., Kimura, M.: A model of mutation appropriate to estimate the number of electrophoretic detectable alleles in a finite population. Genet. Res. **22:** 201–204 (1973)
36. Petes, T.D., Greenwell, P.W., Dominska, M.: Stabilization of microsatellite sequences by variant repeats in the yeast *Saccharomyces cerevisiae*. Genetics **146:** 491–498 (1997)
37. Reich, D. E. Goldstein, D. B. : Genetic evidence for a paleolithic human population expansion in Africa. Proc. Natl. Acad. Sci. USA **95:**8119–8123 (1998)
38. Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.W., Kidd, K.K., Zhivotovsky, L.A., Feldman, M.W.: Genetic structure of human populations. Science **298:** 2381–2385 (2002)
39. Rubinsztein, D.C.: Trinucleotide expansion mutations cause disease which do not conform to classical Mendelian expectations. In: Goldstein, D.B., Schlötterer, C. (ed) Microsatellites. Evolution and Applications. Oxford University Press, Great Britain (1999)
40. Sainudiin, R.: Statistical inference of microsatellite models: an application to humans and chimpanzees. MS Thesis, Cornell University, New York (2003)
41. Schlötterer, C. : Hitchhiking mapping - functional genomics from the population genetics perspectve. Trends in Genetics **19:**32–38 (2003)
42. Schofield, M. J. Hsieh, P. : DNA mismatch repair: molecular mechanisms and biological function. Annu Rev Microbiol **57:**579–608 (2003)
43. Shinde, D., Lai, Y., Sun, F., Arnheim, N.: *Taq* DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: $(CA/GT)_n$ and $(A/T)_n$ microsatellites. Nucleic Acids Research **31:** 974–980 (2003)
44. Sibly, R.M., Whittaker, J.C., Talbot, M.: A maximum-likelihood approach to fitting equilibrium models of microsatellite evolution. Mol. Biol. Evol. **18:** 413–417 (2001)
45. Sibly, R.M., Meade, A., Boxall, N., Wilkinson, M.J., Corne, D.W., Whittaker, J.C.: The structure of interrupted human AC microsatellites. Mol. Biol. Evol. **20:** 453–459 (2003)
46. Stephens, J.C., Reich, D.E., Goldstein, D.B., Shin, H.D., Smith, M.W., *et al.:* Dating the origin of the $CCR5-\Delta 32$ AIDS-resistance allele by the coalescence of haplotypes. Am. J. Hum. Genet. **62:** 1507–1515 (1998)
47. Tachida, H. Iizuka, M. : Persistence of repeated sequences that evolve by replication slippage. Genetics **131:**471–478 (1992)

48. Tsao, J-L., Yatabe, Y., Salovaara, R., Järvinen, H.J., Mecklin, J-P., Aaltonen, L.A., Tavaré, S., Shibata, D.: Genetic reconstruction of individual colorectal tumor histories. Proc. Natl. Acad. Sci. USA **97:** 1236–1241 (2000)
49. Walsh, J. B. : Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. Genetics **115:**553–567 (1987)
50. Weber, J.L., Wong, C.: Mutation of human short tandem repeats. Hum. Mol. Genet. **2:** 1123–1128 (1993)
51. Webster, M.T., Smith, N.G.C., Ellegren, H.: Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. Proc. Natl. Acad. Sci. USA **99:** 8748–8753 (2002)
52. Whittaker, J.C., Harbord, R.M., Boxall, N., Mackay, I., Dawson, G., Sibly, R.M.: Likelihood-based estimation of microsatellite mutation rates. Genetics **164:** 781–787 (2003)
53. Wierdl, M., Dominska, M., Petes, T.D.: Microsatellite instability in yeast: dependence on the length of the microsatellite. Genetics **146:** 769–779 (1997)
54. Xu, X., Peng, M., Fang, Z., Xu, X.: The direction of microsatellite mutation is dependent upon allele length. Nat. Genet. **24:** 396–399 (2000)
55. Yao, X., Buermeyer, A.B., Narayanan, L., Tran, D., Baker, S.M., Prolla, T.A., Glazer, P.M., Liskay, R.M., Arnheim, N.: Different mutator phenotypes in *Mlh1*- versus *Pms2*- deficient mice. PNAS, USA **96:** 6850–6855 (1999)
56. Zhivotovsky, L. A. Feldman, M. W. : Microsatellite variability and genetic distances. PNAS, USA **92:** 11549–11552 (1995)
57. Zhivotovsky, L. A. Feldman, M. W. , Grishechkin, S. A. : Biased mutations and microsatellite variation, MBE **14:** 926–933 (1997)
58. Zhivotovsky, L. A. , Bennett, L. , Bowcock, A. M. , and Feldman, M. W. : Human population expansion and microsatellite variation. Mol. Biol. Evol. **17:** 757–767 (2000)