Review Article

Launching Microsatellites: A Review of Mutation Processes and Methods of Phylogenetic Inference

D. B. Goldstein and D. D. Pollock

Microsatellites are short segments of DNA in which a specific motif of 1–6 bases is repeated up to a usual maximum of 60 or so. Due to their exceptional variability and relative ease of scoring, microsatellites are now generally considered the most powerful genetic marker. It is typical to observe loci with more than 10 alleles and heterozygosities above 0.60, even in relatively small samples (Bowcock et al. 1994; Deka et al. 1995), while certain loci can be considerably more variable (Primmer et al. 1996). In addition to being highly variable, microsatellites are also densely distributed throughout eukaryotic genomes, making them the preferred marker for very-high-resolution genetic mapping (Dib et al. 1996; Dietrich et al. 1996). Microsatellites are rapidly replacing RFLPs and RAPDs in most applications in population biology, from identifying relatives to inferring demographic parameters (Blouin et al. 1996; Bowcock et al. 1994; Goldstein et al. 1996; Jame and Lagoda 1996). Part of the appeal of microsatellites over RFLPs and RAPDs is that the genetic basis of microsatellite variability is readily apparent: unique primers amplify a genomic region including a well-defined repeat structure that is responsible for the observed variation. This allows the development of inferential methods based on explicit models of microsatellite evolution (Feldman et al. 1996; Goldstein et al. 1995a,b, 1996; Pollock DD, Bergman A, Feldman MW, and Goldstein DB, submitted; Slatkin 1995a,b). These advantages suggest that microsatellites will enjoy a lengthy reign in population studies.

One perceived difficulty with microsatellites is the long lead time in identifying and characterizing microsatellites in new taxonomic groups. This problem is partially alleviated, however, by the continuing popularity of microsatellites in genetic mapping. Microsatellite maps are now available in nearly all organisms of genetic and/or economic interest including humans, mice, fruit flies, cows, sheep, chickens, pigs, tomatoes, soybeans, and rice, among others (Akkaya et al. 1995; Broun and Tanksley 1996; Causse et al. 1994; Crawford et al. 1995; Crooijmans et al. 1996; Dib et al. 1996; Dietrich et al. 1996; Goldstein and Clark 1995; Ma et al. 1996; Postlethwait et al. 1994; Rohrer et al. 1996; Su and Willems 1996; Taramino and Tingey 1996). In addition, large databases of microsatellites isolated for population work are accumulating: one maintained at the Smithsonian Laboratory of Molecular Systematics includes 25 species, and is certainly an underestimate of those available. One practical long-term difficulty with microsatellite markers is the requirement of determining fragment lengths, which would seem to complicate automation. Ultimately the future may belong to markers amenable to yes/no tests which can be set up on dense chips (e.g., single nucleotide polymorphisms).

In contrast with their importance in intraspecific studies, microsatellites have yet to make any real contribution to phylogeny reconstruction. This failure has come as a surprise to those who suspected that the huge number of microsatellites available, coupled with their very rapid rate evolution, would make them particularly useful in working out the relationships among very closely related species (e.g., Goldstein et al. 1995a). Although it is not yet entirely clear why microsatellites have not been more successful in reconstructing phylogenies, part of the difficulty certainly stems from restrictions to divergence imposed by range constraints, irregularities and asymmetries in the mutation process, and the degradation of microsatellites over time. A number of recent studies have developed theoretical methods to both estimate the relevant molecular details and to correct for them statistically, but they have yet to be tested. Our purpose here is to provide a nontechnical introduction to the concerns related to

From the Department of Zoology, South Parks Road, University of Oxford, Oxford OXI 3PS, England (Goldstein), and the Department of Mathematical Biology, National Institute for Medical Research, London, England (Pollock). We thank Dr. Stephen O'Brien and members of the Laboratory of Genomic Diversity for encouraging us to write this review. D.D.P. is a Hitchings-Elion Fellow of the Burroughs Wellcome Fund. A computer program written by Minch et al., which calculates statistics for microsatellite data, is available on L. L. Cavalli-Sforza's home page at http://lotka.stanford.edu or by writing to Dr. Eric Minch at minch@malecot.stanford.edu.

Journal of Heredity 1997;88:335–342; 0022-1503/97/$5.00

the available methods and to suggest how the methods might be applied. We begin with a review of microsatellite mutation and evolution, emphasizing those features relevant to the basic assumptions of the early stepwise distances (Goldstein et al. 1995a; Slatkin 1995b). In particular, we consider (1) the mutation rate, (2) the distribution of mutation sizes, (3) constraints on the number of repeats (repeat count or allele size), (4) the degree of asymmetry in the mutation distribution, and (5) the dependence of the mutation process on diploid genotype. Next we describe the principle analytic distances for microsatellites and a number of recent modifications that have been made, focusing in particular on their assumptions about the molecular details of microsatellites. We will also indicate how the modified distances can be used to estimate the parameters governing microsatellite mutation and evolution. Such estimation may ultimately allow the partitioning of loci into classes appropriate for particular problems.

Molecular Details

Mutation Rate

A variety of in vivo and in vitro studies indicate that microsatellite loci are highly unstable, having some of the highest mutation rates observed at molecular loci. Microsatellite mutation processes have been inferred by direct observations both on artificial constructs in yeast (Henderson and Petes 1992) and in human pedigrees (Weber and Wong 1993). The general conclusion from these studies is that there is an exceptionally high rate of mutation adding or subtracting a small number of perfect repeats. In humans, the average overall mutation rate for 28 di- and tetranucleotide microsatellites was estimated at about 0.001, with the tetranucleotide repeats significantly more mutable than the dinucleotide repeats. The most popular explanation for the high mutation rate is polymerase slippage (Levinson and Gutman 1987), a hypothesis that received considerable support from an elegant in vitro analysis showing that polymerase tends to miscopy repeated tracks of DNA (Schlötterer and Tautz 1992).

Distribution of Mutation Sizes

While the majority of observed mutations are of a single step (one repeat unit), a significant minority of mutations may be of larger size. Out of 22 observed germ line mutations, Weber and Wong (1993) confirmed no mutations of larger than two repeats. Twenty of these mutations involved a change of a single step for a ratio of 0.91 single-step to two-step mutations. A subsequent study by Amos et al. (1996) confirmed only a single mutation of larger than one repeat unit out of 15 observed mutations. Engineered repeat tracks in yeast also show a great preponderance of single- and two-step mutations (Henderson and Petes 1992). The general conclusion from these studies is that the majority of mutations are of one or two steps. It should be kept in mind, however, that in observing relatively few mutations these studies are biased toward the most common types of mutation. It remains possible that mutations of much larger sizes occur, but too infrequently to be routinely picked up in such studies. Indirect evidence for such mutations comes from the study of distances among human populations in which the calibrated mutation rate is somewhat higher than that observed in pedigrees (Goldstein et al. 1995b).

The occurrence of mutations larger than one or two steps is confirmed by studies of trinucleotide expansions in which alleles beyond a certain size threshold have asymmetric distributions of mutations, including some of very large size. In fragile chromosome sites, for example, the disease-causing allele may have over 1000 repeats of CCG. The CAG repeats associated with some neurological disorders may also mutate to alleles with over 1000 repeats when they occur outside of coding regions (Ashley and Warren 1995). Perhaps surprisingly, these very large sizes associated with diseases are rarely reported at other microsatellites, although it should be noted that a sampling bias exists in that the expanded trinucleotides are identified by their phenotypic effects. A detailed characterization of maximal allele sizes at loci not associated with disease is necessary to determine whether trinucleotide expansion behavior can be generalized to other types of microsatellites. Since the expanded trinucleotides have phenotypic effects even when they are not expressed, alleles above a certain size are probably eliminated quickly. Given that atypically large alleles are hypermutable, leading to the production of expanded, symptomatic alleles, this suggests a potential mechanism of size constraint.

Asymmetry of Mutation Distribution

A tendency to mutate to alleles of larger size (positive asymmetry) was first observed in the asymmetric mutation distribution for large alleles at trinucleotide expansion loci (Ashley and Warren 1995). Subsequently, positive asymmetry was invoked by Rubinsztein et al. (1995) as part of an explanation of observed differences in average repeat sizes between humans and other primates. They posited that microsatellites in humans have a greater tendency toward positive asymmetric mutation than those in other primates. One major problem with this inference is that since the loci were selected in humans, any real differences in microsatellite characteristics between the species are confounded with ascertainment bias (Ellegren et al. 1995; Box A). Subsequent studies, however, have demonstrated that asymmetric mutation is not restricted to trinucleotide expansion loci. Primmer et al. (1996) made a detailed study of a single highly polymorphic tetranucleotide locus in the swallow (*Hirundo rustica*). Out of 841 meioses, 26 mutations increasing size were observed, compared with 7 decreasing it, with the majority of changes involving the gain or loss of a single repeat unit. Amos et al. (1996) added 15 new germ-line mutations to those reported in Weber and Wong (1993), and showed a significant excess of mutations increasing allelic size. The generality of these results is not yet clear, however, especially given that artificially constructed repeat tracks introduced into both bacteria (Levinson and Gutman 1987) and yeast (Henderson and Petes 1992) show an asymmetry toward mutations that decrease size. These observations, together with the behavior of expanded trinucleotide alleles, suggest that the degree of asymmetry may depend on allele size. In assessing asymmetry, it will therefore be important not only to consider differences among loci but also differences among alleles within a locus.

These results raise key questions about microsatellite persistence. In particular, since loci with more than 60 or so repeats are rarely observed (but see Primmer et al. 1996), something must restrict the size of those loci showing positive asymmetry. Alternatively, microsatellites may be unstable above a certain threshold and quickly degrade through large deletions or through the introduction of imperfections. It is interesting in this regard that large GT repeat tracks engineered into plasmids tend to undergo large deletions (Levinson and Gut-

Box A. Ascertainment Bias

Rubinsztein et al. (1995), in a study of the average sizes of microsatellite loci in humans and other primates, reported a significant bias toward greater length in humans. Based on this they suggested that there is an inherent difference between human microsatellites and those in other primates, perhaps having to do with the degree of asymmetry in the mutation process. Ellegren et al. (1995), however, pointed out that the length differences could be due to ascertainment bias: microsatellites tend to be selected in a focal species (the species in which the microsatellites were first developed) to be either polymorphic or long. Since length and polymorphism are positively correlated (see above), both criteria result in loci longer than average. Amos and Rubinsztein (1996) defended their initial interpretation with a number of novel statistical approaches, but the issue of ascertainment bias itself has been largely dropped as the original participants in the debate moved on to detailed characterizations of asymmetry in the mutation distribution at particluar loci in specific taxa. This is unfortunate because whatever the level of asymmetry, and despite assertions to the contrary by Amos and Rubinsztein (1996), ascertainment bias will influence all interspecific comparisons and must be carefully taken into account.

In fact, it is straightforward to make a quantitative assessment of ascertainment bias. Imagine that in a focal species, microsatellites are selected from a pool of loci with a range of R (that is, alleles may have any number of repeats from 1 to R). For convenience, we will refer to the average length of alleles at a locus (in one taxon) as the length of that locus. Assume that in the focal species only microsatellites longer than C repeat units are accepted for subsequent analysis (the cutoff being imposed directly by a preference for clones with long alleles, indirectly by the screening process, or by a preference for polymorphic markers). On average the selection process in the focal species results in microsatellites of length $(R + C)/2$. In a related but sufficiently diverged species the average length at the same loci would be $R/2$. If the difference in length due to ascertainment bias is denoted D_a, then we have $D_a = (R + C)/2 - R/2 = C/2$. The magnitude of the difference is therefore independent of R. This argument could be refined by taking account of various complications (especially correlations in size between the focal and related species), but the point is already clear: the absolute bias is substantial, and for moderate R it is substantial as a fraction of R. It is especially interesting to note that it is customary to focus on microsatellites with 10 or more repeats, as these are often polymorphic. Then $C = 10$ and we predict that humans would have, on average, five more repeats than other primates, in striking agreement with the reported difference of four repeats between humans and chimpanzees. Thus, once ascertainment bias is taken into account, we see that in fact there is nothing to explain with respect to the difference in average length between human and other primate microsatellites reported by Rubinsztein et al. (1995).

The point here is not to further belabor the argument of whether the differences between humans and other primates reflects some inherent "directional" difference as claimed by Rubinsztein et al. (1995). That argument should (and certainly will) be settled by comparing microsatellites first selected in other primates with those first selected in humans. The point is rather to demonstrate that differences in the average length between species are expected whenever microsatellites selected in one species are carried over to another. Since length and variability are correlated, this difference imposes a bias in the variability expected in the focal and related species. Moreover, an additional contribution to such bias arises from the preference for pure stretches of repeats in the focal species. Even in closely related species these stretches will often be interrupted by imperfections (Crouau-Roy et al. 1996; Garza and Freimer 1996). Since imperfections are known to stabilize microsatellites, this difference will further the contribution that ascertainment bias will make to the differences between species in variability at microsatellite loci. For these reasons it is critical that sets of microsatellites with consistent structures be used to calculate genetic distances, and especially to compare variabilities among taxa.

man 1987). This would, however, seem to predict a shorter life span for microsatellites than is consistent with observations on at least some loci. Coote and Bruford (1996), for example, found a set of microsatellites first identified in humans that are polymorphic in the majority of apes and Old World monkeys, which includes species that last shared a common ancestor about 30 million years ago. More dramatically, FitzSimmons et al. (1995) reported conservation of orthologous microsatellite loci over 300 million years in marine turtles. It will be especially interesting to determine whether a relationship exists between microsatellite longevity and mutational asymmetry.

Range Constraints

Perhaps the most compelling evidence that the number of repeats at microsatellite loci is under some form of constraint is simply the absence of alleles of very large size. Given the high mutation rate, and the very large number of loci that have been characterized, it is clear that if the process were an unconstrained random walk we would expect to regularly observe loci with very large alleles. In fact, with the exception of trinucleotide expansion loci, alleles much greater than 60 repeats are very rarely observed (but see Primmer et al. 1996).

Other lines of argument have provided less direct evidence of a length ceiling. Bowcock et al. (1994), for example, found that the variance in repeat score among primates is not significantly larger than that among human populations. Under an unconstrained random walk, the greater evolutionary distance among primates would be expected to lead to a greater variance by increasing the between-group component of the total variance. Similarly, it has been reported that the ratio of the genetic distance between apes and humans compared with that between African and non-African populations is much less than would be expected in the absence of range constraints (Garza et al. 1995). Since these loci were first selected in humans, however, they are expected to represent a biased sample of the locus properties in humans. Some of these observations, therefore, may be due to ascertainment bias as opposed to range constraints per se (see Box A).

Dependence of the Mutation Process on Allele Size and Sequence

In trinucleotide repeat expansion loci, the rate and distribution of mutations change dramatically as allele sizes pass from the premutation (atypically large but nonsymptomatic) to the full-mutation state (Ashley and Warren 1995). A number of population studies have also tested the dependence of the mutation rate on allelic size by correlating observed levels of variation with average allele size. This approach utilizes the theoretical result of Moran (1975), who showed that at mutation-drift equilibrium the variance in size at a locus undergoing stepwise mutations is $2(N - 1)\beta$, where N is the haploid population size and 2β is the total mutation rate. The largest such study to date (Valdes et al. 1993) used PCR fragment size as a substitute for the number of repeats and reported no correlation between average fragment size and the observed variance. The relationship between PCR fragment size and number of repeats is not particularly tight, however, because the size of the nonrepeat portion of the PCR fragment varies from locus to locus. Moreover, systematic bias may have been introduced in the data by the research procedure used to select primers; algorithms often seek fragments in a specified size range. Goldstein and Clark (1995) analyzed the dependence of the allelic variance on the repeat count itself, considering both the average size and the maximum size at a locus. Both correlations were significant, but the latter more so. This suggests that the increase in mutation rate with repeat size is not linear, or that some other assumption of the stepwise mutation model is violated. Interestingly, the same pattern was observed for both di- and trinucleotide microsatellites.

It has also been inferred from population studies that imperfections in the repeat array tend to stabilize microsatellites (Goldstein and Clark 1995). This conclusion is supported by the observation that normal alleles at trinucleotide expansion loci often carry imperfections, while the pre- and full-mutation alleles do not (Ashley and Warren 1995). The sensitive dependence on the exact sequence in the repeated regions further complicates comparisons of microsatellite variability between species and suggests that microsatellite degradation may involve the introduction of imperfections (Garza and Freimer 1996). It is therefore especially important to sequence at least a single allele from each locus when extending primers from a focal species to close relatives (see Box A).

Effects of Heterozygosity

In the case of minisatellite regions, which involve repetitions of longer sequence motifs than microsatellites, it is known that mutation can result from unequal exchange during meiosis (Jeffreys et al. 1988), and it seems reasonable that this mechanism can also operate at microsatellite loci, at least for the larger alleles. The suggestion in Amos et al. (1996) that the probability of mutation increases with the difference in size between homologous alleles is consistent with a role for unequal exchange in microsatellite mutation. Mutational dependence on diploid genotype would have a dramatic impact on the dynamics of allele frequency evolution at microsatellite loci and warrants more detailed study.

Genetic Distance Measures

If a distance is used to estimate relative times of divergence, it is essential that its expectation increases linearly with time and beneficial if the coefficient of variance is low. For reconstruction of phylogenetic relationships, the combination of linearity and variance determines the performance (Goldstein and Pollock 1994; Pollock and Goldstein 1995). A useful measure that combines these features is the accuracy index of Tajima and Takezaki (1994), defined as the slope of the distance at any time divided by its standard deviation. Thus, if the variance is constant, distances will be most accurate over time if they maintain a constant rather than a decreasing slope. In general, distances are constructed to be both as linear and as precise as possible under the assumption of a particular model of evolution. It should be appreciated, however, that a trade-off between the two often exists (Goldstein and Pollock 1994; Pollock and Goldstein 1995).

The majority of mutations at microsatellite loci are stepwise in nature, changing allelic sizes by one or a very few number of repeats, and thus distances that are designed specifically to apply to microsatellites generally assume Ohta and Kimura's (1973) stepwise mutation model (SMM) or one of its generalizations. Most classical distance measures, however, are based either on Kimura and Crow's (1964) infinite alleles model (IAM), or upon multidimensional geometric considerations without reference to a particular evolutionary model. The assumptions of the SMM differ sharply from the assumptions of the IAM, and therefore distances designed to increase linearly under the IAM, such as Nei's standard distance, are both nonlinear and inaccurate for microsatellite loci (Goldstein et al. 1995a; Takezaki and Nei 1996).

Despite the fact that they are not based on the SMM or any other evolutionary model, a group of related distances performs well for reconstruction of phylogenies when taxa are closely related. Cavalli-Sforza and Edwards's (1967) chord distance (D_C), Nei et al.'s (1983) distance (D_A), and Stephens et al.'s (1992) allele sharing distance (D_{AS}) all make use of the product of allele frequencies shared between populations (see Box B), and have been shown to reconstruct closely related phylogenies better than SMM-based distances (Goldstein et al. 1995a,b; Takezaki and Nei 1996). It is clear from these studies that these distances do not increase linearly with time, however, and become extremely flat as time becomes large. Thus they do not reflect divergence time unless taxa are very closely related. Their accuracy at short distances stems from their use of the information available in the degree of overlap between the allele frequency distributions of two populations, and they are less accurate at greater distances where the amount of overlap cannot decrease below zero and the distance between distributions becomes important. The amount of overlap between distributions is also sensitive to fluctuations in the effective population size, and thus it is not surprising that these distances are much less accurate when population bottlenecks have occurred (Takezaki and Nei 1996). It should be noted that each locus may have been subjected to different apparent fluctuations in effective population size due to positive, balancing, or slightly deleterious selection (Nauta and Weissing 1996; Slatkin 1995a). The sensitivity of these distances to fluctuations in effective population thus presents a very serious complication.

Three distances have recently been developed specifically for application to microsatellite evolution assuming the SMM (see also Chakraborty and Nei 1977). Goldstein et al.'s (1995a) and Slatkin's (1995b) distance (ASD, described in Box C) increases linearly with time under the unconstrained SMM model. The main difficulty with this distance is its high variance, partly due to its dependence on the variation within populations. In addition to this, because population sizes are likely to vary among taxa in any phylogeny, the inclusion of the intrapopulation variance term $2(N - 1)\beta$ obscures the relationship between separation time and the observed value of ASD. The intrapopulation

Box B. $D_{AS,}$ D_A, and D_c

Three important distances that are not based on a model of evolution all focus on the sum of the products of frequencies of those alleles shared between two populations; that is, applied to a single locus, the three distances take the form:

$$D = c\left[1 - \sum_i \left(x_i y_i\right)^a\right]^b,$$

where x_i and y_i are the frequencies of alleles of size i in populations x and y, respectively, and a, b, and c are constants. For D_{AS}, a, b, and c are all equal to one. For D_A, a equals 0.5, while b and c are equal to 1, while for D_c, a and b equal 0.5, and c equals $2(2)^{.5}/\pi$. For multiple loci, the average distance is taken over all loci. Although the expectations of these distances are clearly different, it has not been shown that there is substantial difference between the accuracy of D_{AS} and D_A in reconstructing phylogenetic trees. D_c is slightly less efficient than D_A in phylogenetic reconstruction under the SMM model (Takezaki and Nei 1996). In the absence of range constraints these distances vary between some nonzero positive number and c as time progresses. Latter's F_{ST} distance and Nei's minimum genetic distance both separately incorporate the sum of the squared allele frequencies in each population in order to create distances which vary from zero to some positive number, but these distances are generally less accurate than D_A and/or D_c (Takezaki and Nei 1996).

While the between-locus variance of these measures is large, making it essential to bootstrap over loci to assess reliability, the degree of allele sharing may be affected by sampling, especially when the sample size is small. In such cases, bootstrapping over individuals may provide useful information, although never as a substitute for bootstrapping over loci.

variance term also makes ASD very sensitive to fluctuations in population size in a manner similar to the geometric distances described above (Nauta and Weissing 1996; Takezaki and Nei 1996). Goldstein et al.'s (1995b) distance, $(\delta\mu)^2$, was specifically designed to overcome problems associated with the variance term (see Box C). This distance increases linearly with time at the same rate as ASD, but has a lower variance and thus seems to be always preferable for both phylogenetic reconstruction and estimation of relative separation times. Although independence of this distance from population size was derived under the assumption of constant population size, computer simulations show that the standardization achieved by averaging scores within populations results in a distance that is extremely robust to fluctuations in population size, perhaps more so than any other distance defined in terms of allele frequencies (Takezaki and Nei 1996).

Shriver et al.'s (1995) stepwise distance (D_{sw}) is similar in form to ASD (with variance correction), but with an absolute value operation replacing the square function on the difference between allele sizes (Box C). D_{sw} was developed through heuristic argument, and an explicit dynamic has not been derived. Nevertheless, the exact linearity of ASD implies that D_{sw} cannot be linear, an inference borne out by computer simulations (Shriver et al. 1995). Under some circumstances, however, D_{sw} has a lower coefficient of variance and may therefore be preferred for phylogenetic reconstruction. Under the conditions analyzed by Takezaki and Nei (1996), however, this was rarely the case. In addition, this distance is extremely sensitive to variation in population size.

Although ASD and $(\delta\mu)^2$ were derived assuming a strict stepwise mutation model, they are in fact considerably more general. In particular, if mutation sizes vary, the expectation of ASD is altered only by replacing the overall mutation rate, 2β, with the product of the mutation rate and the variance of mutational step sizes (Slatkin 1995b). Kimmel et al. (1996) also note that the linearity of stepwise distances is independent of the assumptions of both single-step sizes and symmetry in the mutation rate, the latter point being of particular significance given recent demonstrations of asymmetry as described above. These results suggest that the two greatest concerns for extension of microsatellite loci to phylogenetic reconstruction of more distantly related organisms are constraints on allele sizes and the longevity of the mutational properties of microsatellite loci (Garza and Freimer 1996). The rate of degradation of microsatellite loci requires careful comparative analysis. As described above, it is essential to sequence microsatellites in all taxa to confirm that the repeated motifs have not been interrupted by imperfections. Assuming that microsatellites with sufficient longevity can be identified, the restrictions on divergence imposed by range constraints must be accounted for.

If the number of repeats attainable by microsatellite loci is restricted, the accuracy and linearity with time of all distances are strongly affected (Feldman et al. 1996; Goldstein et al. 1995a; Pollock DD, Bergman A, Feldman MW, and Goldstein DB, submitted). To statistically adjust for the effects of range constraints, our group has introduced a number of new distances (see Box D), including a log-based distance denoted D_L (Feldman et al. 1996), a least squares distance denoted D_{LS} (Pollock DD, Bergman A, Feldman MW, and Goldstein DB, submitted), and a generalized least squares distance denoted D_{GLS} (Pollock DD, Bergman A, Feldman MW, and Goldstein DB, submitted). Although the model upon which these distances are based is overly simplified (the range terminates in reflecting boundaries at the upper and lower ends; there is no asymmetry or dependence in the mutation rate with repeat number), we suspect that some of the results are significantly more general than these assumptions suggest. In particular, much of the behavior of the distances can be attributed to the fact that the length of time over which a locus will accurately reflect separation times decreases both as the mutation rate increases and as the number of attainable states decreases. This basic interaction between the range and the mutation rate is likely to come into effect for any reasonable mutation model.

Since the usefulness of a locus for assessment of deep divergence times depends upon the locus range and mutation rate, it is critical to accurately assess these parameters. Pollock DD, Bergman A, Feldman MW, and Goldstein DB (submitted) developed methods to assess range constraints under Feldman et al.'s (1996) and Nauta and Weissing's (1996) reflecting boundary SMM model. The obvious esti-

Box C. ASD, $(\delta\mu)^2$, and D_{SW}

ASD, $(\delta\mu)^2$, and D_{sw} make use of the difference in size between alleles, which is expected to carry useful information under the SMM model. ASD takes the form

$$\sum_i \sum_j (i - j)^2 x_i y_j \tag{C1}$$

where x_i and y_j are the frequencies in population x and y of alleles with i and j repeats, respectively. $(\delta\mu)^2$ and D_{sw} take the form

$$\sum_i \sum_j f(i - j)x_i y_j - \left(\sum_i \sum_j f(i - j)x_i x_j + \sum_i \sum_j f(i - j)y_i y_j\right)\Big/2, \tag{C2}$$

where for D_{sw}, f is the absolute value function, and for $(\delta\mu)^2$, f is the square function. $(\delta\mu)^2$, of course, conveniently simplifies to its familiar form, $(\delta\mu)^2 = (m_x - m_y)^2$, where m_x and m_y are the means of allele sizes in populations x and y, respectively. While ASD and $(\delta\mu)^2$ were derived analytically under the SMM model, and both are expected to increase linearly with time, D_{sw} was derived heuristically and does not. Both $(\delta\mu)^2$ and D_{sw} equal zero when the time separating two populations is zero, and increase with time. The expectation of ASD, however, is $2\beta\tau + V_x + V_y$, where τ is the time since separation, β is the stepwise mutation rate in each direction, and V_x and V_y are the variances in allelic score of populations x and y, respectively. Thus ASD does not begin at zero, and the inclusion of the variance terms add significantly to the variance of the distance measure.

In the preceding discussion it is assumed that sampling is complete. In practice, however, we must estimate $(\delta\mu)^2$, which describes populations x and y, based on samples from these populations. Equation (C2) for $(\delta\mu)^2$ can be rewritten as $(\delta\mu)^2 = \text{ASD} - V_x - V_y$. Goldstein et al. (1995a) showed that ASD calculated for the sample is in fact an unbiased estimator of ASD for the population. It is well known, however, that sample variances are not unbiased estimators of the parametric variance. Therefore, an unbiased estimator of $(\delta\mu)^2$ is easily obtained by substituting unbiased estimators for V_x and V_y. That is, in practice $(\delta\mu)^2$ can be calculated as

$$(\delta\mu)^2 = \text{ASD} - \tilde{V}_x - \tilde{V}_y = \sum_i \sum_j (i - j)^2 x_i y_j - \frac{n_x}{n_x - 1} \sum_i (i - m_x)^2 x_i - \frac{n_y}{n_y - 1} \sum_i (i - m_y)^2 y_i \tag{C3}$$

where $\tilde{V}$ denotes the usual unbiased estimate of the parametric variance based on the variance in the sample, x_i and y_j, now represent the frequency of alleles i and j in the samples from populations x and y, m_x and m_y are the sample means, and n_x and n_y are the number of alleles sampled from each population. Equation (C3) shows that the difference between the unbiased estimate of $(\delta\mu)^2$ and the squared difference between sample means [that is, $(\delta\mu)^2$ calculated directly for the sample] will be very slight unless both the sample size and the level of differentiation is small. The amount of differentiation matters because the bias enters in only through the estimation of the variance. If the expected difference between the means is very small relative to the variance, even a few percent change in the variance will be large as a proportion of the distance. In those cases where differentiation is very slight relative to the population variance, however, nonstepwise distances should not be used (see above). In practice, therefore, we do not expect that the bias corrected version of $(\delta\mu)^2$ will be noticeably different from the uncorrected version.

Zhivotovsky and Feldman (1995) were able to derive the variance of $(\delta\mu)^2$ under the unbounded SMM model, and it is $2(2\beta\tau)^2$. Thus, the expected standard deviation for $(\delta\mu)^2$ averaged over equivalent loci is the square root of twice the expectation of $(\delta\mu)^2$ divided by the square root of the number of loci. This makes it clear why many loci are needed to achieve an acceptable degree of accuracy with these measures. We would predict that reasonable reliability would generally require between 50 and 100 loci as a minimum. The requirement of a large set of loci becomes even more pronounced when one considers the variation among loci in mutation rate and other basic properties. Finally, the large drift variances of these genetic distance measures necessitate that empirical confidence intervals for them be derived by bootstrapping over loci. It is never appropriate to use bootstrapping over individuals as a substitute for bootstrapping over loci in assessing the reliability of these distances.

mator, the difference between minimum and maximum allele sizes, is extremely inaccurate when only one or a few independent populations are available. Corrections developed have increased accuracy (Pollock DD, Bergman A, Feldman MW, and Goldstein DB, submitted), but assume that the populations are sufficiently diverged that mean allele sizes are no longer correlated. Reasonable adjustments for phylogenetic relatedness may eventually be developed for these estimators, but in the meantime it is probably best to make estimates based on well-diverged populations for application to closer populations (assuming the microsatellites have not been degraded). Under the conditions analyzed in Pollock DD, Bergman A, Feldman MW, and Goldstein DB (submitted), relative mutation rates estimated from the allelic variance (Feldman et al. 1996; Nauta and Weissing 1996) were somewhat more accurate than when estimated via the least squares distance methods they introduced. The variance-based methods may be affected more by population size fluctuations and selection, however, and the least squares methods may improve faster as more taxa are introduced. It will be particularly interesting to see how well these different methods perform on real datasets, in particular whether they succeed in dramatically improving predictions concerning long-term microsatellite locus evolution and whether they can be used to effectively partition loci according to usefulness in addressing particular phylogenetic questions.

For very recently separated populations, the distances that make use of the product of allele frequencies shared between populations are most accurate (although not linear with time), and they are more accurate when allelic variance, and thus the mutation rate, is high (Takezaki

Box D. D_L, D_{LS}, D_{GLS}

D_L, D_{LS}, and D_{GLS} were developed using a simple model in which the range is restricted by reflecting boundaries at low and high values, and otherwise is identical to the standard SMM model. D_L takes the form

$$\log\left[1 - \sum_l (\delta\mu)^2{}_l/LM\right],$$

where $(\delta\mu)^2$ is defined in Box C for an individual locus, L is the number of loci, and M is the average value of $(\delta\mu)^2$ at maximal divergence (Goldstein et al. 1995a; Feldman et al. 1996). $M = (R^2 - 1)/6 - 2V$, where R is the number of states the microsatellite can assume, and V is the allelic variance. When all loci have identical range restrictions (R) and stepwise mutation rates (β), D_L will increase nearly linearly with time for much longer than $(\delta\mu)^2$. If the range and rate do not differ excessively, the average M can be substituted and D_L will still be relatively accurate, although not increasing exactly linearly with time. Under conditions of arbitrary range and rate variation, linear distance corrections can be obtained by the least squares methods resulting in the distances D_{LS} and D_{GLS}, which are obtained by finding the minimum with respect to time, τ, of

$$\sum_l \{(\delta\mu)^2{}_l - E[(\delta\mu)^2{}_l]\}^2 \Big/ W_l$$

where the expectation of $(\delta\mu)^2$ for a particular locus l is given by $E[(\delta\mu)^2_l = M(1 - \exp[-4\beta\tau + 4\beta\tau\cos(\pi/R)\tau])$. W_l is either one for D_{LS}, or σ_l for D_{GLS}, where σ^2_l is the variance at locus l. These distances require estimation of R and/or β for each locus, and different methods for making those estimations are discussed in Pollock DD, Bergman A, Feldman MW, and Goldstein DB (submitted). Assuming correct knowledge of R and β, when β varies among loci under range constraints, D_{GLS} is considerably more accurate than $(\delta\mu)^2$, D_L, or D_{LS}. When R varies among loci, it is slightly less accurate than the others (Pollock DD, Bergman A, Feldman MW, and Goldstein DB, submitted).

and Nei 1996). For more distant populations, these distances become much less accurate than the distances which make use of the degree of separation between alleles. The $(\delta\mu)^2$-based distances are not sensitive to levels of allelic variance in the absence of range constraints, but with range constraints, loci with the lowest mutation rates will remain accurate longer. Thus, when extending phylogenetic analysis using microsatellites beyond the subspecies level, it is preferable to select those loci with lower allelic variation; exactly the opposite of the preference for studying differentiation of subpopulations within species.

Discussion

With all the difficulties itemized, we wish to emphasize that for certain phylogenetic problems microsatellites remain the most promising approach and it seems well worth the effort of improving methods for their analysis. For example, a method of "genetic absolute dating" based on microsatellites has recently been introduced (Goldstein et al. 1995b). The novelty of this method is that the expected rate of differentiation can be estimated by studying microsatellite mutations in pedigrees, which removes the requirement of rate calibration using uncertain paleontological dates. Moreover, since microsatellite analyses can easily collect information from a number of different genomic regions, it is possible to model the divergence of populations as opposed to the genealogical history of particular genomic regions. For some applications, such as the study of human evolutionary history, inferences about population differentiation are of particular importance. In addition to these advantages, the rapid rate of microsatellite evolution also means that reliable information may be gained even for taxa so closely related that it would be impractical to collect enough sequence information to work out their relationships.

For these reasons it is important to determine whether the incorporation of more details about microsatellite behavior can lead to more accurate inferences. We are especially encouraged in this regard by a simple comparison with the use of sequence variation. Just as genes with evolutionary rates and properties appropriate to a particular phylogenetic problem must be carefully selected, we might expect that microsatellite loci appropriate to particular phylogenetic problems must be screened and selected.

A great deal of empirical work remains to be done in evaluating how best to employ microsatellites in phylogeny reconstruction. The methods for assessing range constraints and mutation rates need to be applied to real data from different taxonomic groups and types of microsatellite loci. Clustering loci with similar ranges and mutation rates could be very useful, but the statistical considerations involved in such procedures remain to be elucidated. One of our major goals here is to encourage the collection of the data necessary to estimate the relevant details of microsatellite behavior, that is, data on both the length and sequence structure of homologous microsatellites in sets of related species. Once multiple observations for different types of microsatellites are available it will be possible to determine whether a priori characteristics (e.g., motif size/type) correlate with key features such as range constraints, longevity, and mutation rates.

References

Akkaya MS, Shoemaker RC, Specht JE, Bhagwat AA, and Cregan PB, 1995. Integration of simple sequence repeat DNA markers into a soybean linkage map. Crop Sci 35: 1439–1445.

Amos W and Rubinsztein DC, 1996. Microsatellites are subject to directional evolution. Nat Genet 12: 13–14.

Amos W, Sawcer SJ, Feakes RW, and Rubinsztein DC, 1996. Microsatellites show mutational bias and heterozygote instability. Nat Genet 13:390–391.

Ashley CT and Warren ST, 1995. Trinucleotide repeat expansion and human disease. Annu Rev Genet 29:703–728.

Beckmann JS and Weber JL, 1992. Survey of human and rat microsatellites. Genomics 12:627–631.

Blouin MS, Parsons M, Lacaille Y, and Lotz S, 1996. Use of microsatellite loci to classify individuals by relatedness. Mol Ecol 3:393–401.

Bowcock AM, Linares AR, Tomfohrde J, Minch E, Kidd

JR, and Cavalli-Sforza LL, 1994. High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455–457.

Broun P and Tanksley SD, 1996. Characterization and genetic mapping of simple sequences in the tomato genome. Mol Genet 250:29–49.

Causse MA, Fulton TM, Cho YG, Ahn SN, Chunwongse J, Wu KS, Xiao JH, Yu ZH, Ronald PC, Harrington SE, Second G, McCouch SR, and Tanksley SD, 1994. Saturated molecular map of the rice genome based on an interspecific backcross population. Genetics 138:1251–1274

Cavalli-Sforza LL and Edwards AWF, 1967. Phylogenetic analysis: models and estimation procedures. Am J Hum Genet 19:233–257.

Chakraborty R and Nei M, 1977. Bottleneck effects on average heterozygosity and genetic distance with the stepwise mutation model. Evolution 31:347–356.

Coote T and Bruford MW, 1996. Human microsatellites applicable for analysis of genetic variation in apes and Old World monkeys. J Hered 87:406–410.

Crawford AM, Dodds KG, Ede AJ, Pierson CA, Montgomery GW, Garmonsway HG, Beattie AE, Davies K, Maddox JF, Kappes SW, Stone RT, Nguyen TC, Penty JM, Lord EA, Broom, JE, Buitkamp J, Schwaiger W, Epplen, JT, Matthew P, Matthews ME, Hulme DJ, Beh KJ, McGraw, RA, and Beattie CW, 1995. An autosomal genetic linkage map of the sheep genome. Genetics 140:703–724.

Crooijmans RPMA, Vanoers PAM, Strijk JA, Vanderpoel JJ, and Groenen MAM, 1996. Preliminary linkage map of the chicken (*Gallus domesticus*) genome based on microsatellite markers—77 new markers mapped. Poult Sci 75:746–754.

Crouau-Roy B, Service S, Slatkin M, and Freimer N, 1996. A fine-scale comparison of the human and chimpanzee genomes—linkage, linkage disequilibrium and sequence-analysis. Hum Mol Genet 5:1131–1137.

Ellegren H, Primmer CR, and Sheldon BC, 1995. Microsatellite evolution: directionality or bias in locus selection. Nat Genet 11.360–362

Deka R, Jin L, Shriver MD, Yu LM, Decroo S, Hundrieser J, Bunker CH, Ferrell RE, and Chakraborty R, 1995. Population genetics of dinucleotide $(dC\text{-}dA)_n \cdot (dG\text{-}dT)$ polymorphisms in world populations. Am J Hum Genet 56: 461–474.

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, Lathrop M, Gyapay G, Morissette J, and Weissenbach J, 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature 380:152–154.

Dietrich WF, Miller J, Steen R, Merchant MA, Damron-boles D, Husain Z, Dredge R, Daly MJ, Ingalls KA, O'Connor TJ, Evans CA, DeAngelis MM, Levinson DM, Kruglyak L, Goodman N, Copeland NG, Jenkins NA, Hawkins TL, Stein L, Page DC, and Lander ES, 1996. A comprehensive genetic map of the mouse genome. Nature 380:149–152.

Feldman MW, Bergman A, Pollock DD, and Goldstein DB, 1996. Microsatellite genetic distances with range constraints: analytic description and problems of estimation. Genetics 145:207–216.

FitzSimmons NN, Moritz C, and Moore SS, 1995. Conservation and dynamics of microsatellite loci over 300-million years of marine turtle evolution. Mol Biol Evol 12:432–440.

Garza JC and Freimer NB, 1996. Homoplasy for size at microsatellite loci in humans and chimpanzees. Genome Res 6:211–217.

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

Goldstein DB and Clark AG, 1995. Microsatellite variation in North American populations of *Drosophila melanogaster*. Nucleic Acids Res 23:3882–3886.

Goldstein DB, Linares AR, Cavalli-Sforza LL, and Feldman MW, 1995a. An evaluation of genetic distances for use with microsatellite loci. Genetics 139:463–471.

Goldstein DB, Linares AR, Cavalli-Sforza LL, and Feldman MW, 1995b. Genetic absolute dating based on microsatellites and the origin of modern humans. Proc Natl Acad Sci USA 92:6723–6727.

Goldstein DB and Pollock DD, 1994. Least-squares estimation of molecular distance—noise abatement in phylogenetic reconstruction. Theor Popul Biol 45:219–226.

Goldstein DB, Zhivotovsky LA, Nayar K, Linares AR, Cavalli-Sforza LL, and Feldman MW, 1996. Statistical properties of the variation at linked microsatellite loci—implications for the history of human Y-chromosome. Mol Biol Evol 13:1213–1218.

Henderson ST and Petes TD, 1992. Instability of simple sequence DNA in *Saccharomyces cerevisiae*. Mol Cell Biol 12:2749–2757.

Jame P and Lagoda PJL, 1996. Microsatellites, from molecules to populations and back. Trends Ecol Evol 11.424–430.

Jeffreys AJ, Royle NJ, Wilson V, and Wong Z, 1988. Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. Nature 332:278–281.

Kimmel M, Chakraborty R, Stivers DN, and Deka R, 1996. Dynamics of repeat polymorphisms under a forward-backward mutation model—within-population and between-population variability at microsatellite loci. Genetics 143:549–555.

Kimura M and Crow JF, 1964. The number of alleles that can be maintained in a finite population Genetics 49· 725–738.

Levinson G and Gutman GA, 1987. High frequency of short frameshifts in poly-CA/GT tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. Nucleic Acids Res 15:5323–5338.

Ma RZ, Russ I, Park C, Heyen DW, Beever JE, Green CA, Lewin HA, 1996. Isolation and characterization of 45 polymorphic microsatellites from the bovine genome. Anim Genet 27:43–47.

Moran PAP, 1975. Wandering distributions and the electrophoretic profile. Theor Popul Biol 8:318–330.

Nauta MJ and Weissing FJ, 1996 Constraints on allele size at microsatellite loci: implications for genetic differentiation. Genetics 143:1021–1032.

Nei M, Tajima F, and Tateno Y, 1983. Accuracy of estimated phylogenetic trees from molecular data. J Mol Evol 19:153–170.

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

Pollock DD and Goldstein DB, 1995. A comparison of two methods for constructing evolutionary distances from a weighted contribution of transition and transversion. Mol Biol Evol 12:713–717.

Postlethwait JH, Johnson SL, Midson CN, Talbot WS, Gates M, Ballinger EW, Africa D, Andrews R, Carl T, Eisen JS, Horne S, Kimmel CB, Hutchinson M, Johnson M, and Rodriguez A, 1994. A genetic linkage map for the zebrafish. Science 264:699–703.

Primmer CR, Ellegren H, Saino N, and Møller AP, 1996. Directional evolution in germline microsatellite mutations. Nat Genet 13:391–393.

Rohrer GA, Alexander LJ, Hu ZL, and Smith TPL, Keele JW, and Beattie CW, 1996. A comprehensive map of the porcine genome. Genome Res 6:371–391.

Rubinsztein DC, Amos W, Leggo J, Goodburn S, Jain S, Li SH, Margolis RL, Rose CA, and Fergusonsmith MA, 1995. Microsatellites are generally longer in humans compared to their homologs in non-human primates. evidence for directional evolution at microsatellite loci. Am J Hum Genet 57: 214.

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

Shriver MD, Jin R, Boerwinkle E, Deka R, Ferrell RE, and Chakraborty R, 1995. A novel measure of genetic distance for highly polymorphic tandem repeat loci. Mol Biol Evol 12:914–920.

Slatkin M, 1995a. Hitchhiking and associative overdominance at a microsatellite locus. Mol Biol Evol 12:473–480.

Slatkin M, 1995b. A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457–462.

Stephens JC, Bilbert DA, Yuhki N, and O'Brien SJ, 1992. Estimation of heterozygosity for single-probe multilocus DNA fingerprints. Mol Biol Evol 9:729–743.

Su XZ and Willems TE, 1996. Toward a high-resolution *Plasmodium falciparum* linkage map—polymorphic markers from hundreds of simple sequence repeats. Genomics 33:431–444.

Tajima M and Takezaki N, 1994. Estimation of evolutionary distance for reconstructing molecular phylogenetic trees. Mol Biol Evol 11:278–286.

Takezaki N and Nei M, 1996. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. Genetics 144:389–399.

Taramino G and Tingey S, 1996. Simple sequence repeats for germplasm analysis and mapping in maize. Genome 39.277–287.

Valdes AM, Slatkin M, and Freimer NB, 1993. Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics 133:737–749.

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

Zhivotovsky LA and Feldman MW, 1995. Microsatellite variability and genetic distances. Proc Natl Acad Sci USA 92:11549–11552.

Received January 2, 1997
Accepted March 31, 1997

Corresponding Editor: Stephen J. O'Brien