Estimating the Time to the Most Recent Common Ancestor for the Y chromosome or Mitochondrial DNA for a Pair of Individuals

Bruce Walsh

Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721

Manuscript received July 22, 2000
Accepted for publication March 22, 2001

ABSTRACT

Bayesian posterior distributions are obtained for the time to the most recent common ancestor (MRCA) for a nonrecombining segment of DNA (such as the nonpseudoautosomal arm of the Y chromosome or the mitochondrial genome) for two individuals given that they match at k out of n scored markers. We argue that the distribution of the time t to the MRCA is the most natural measure of relatedness for such nonrecombining regions. Both an infinite-alleles (no recurring mutants) and stepwise mutation model are examined, and these agree well when n is moderate to large and k/n is close to one. As expected, the infinite alleles model underestimates t relative to the stepwise model. Using a modest number (20) of microsatellite markers is sufficient to obtain reasonably precise estimates of t for individuals separated by 200 or less generations. Hence, the multilocus haplotypes of two individuals can be used not only to date very deep ancestry but also rather recent ancestry as well. Finally, our results have forensic implications in that a complete match at all markers between a suspect and a sample excludes only a modest subset of the population unless a very large number of markers (>500 microsatellites) are used.

MOLECULAR marker information has proven an invaluable tool for assessing the degree of relatedness between individuals. To date, most uses of marker information have been concerned with zero-, one- or two-, or deep-generation relatives. By zero-generation, we mean matching/rejecting a forensic sample and a suspect. One- or two-generation assessment includes paternity testing and assessing the degree of relatedness between individuals in natural populations. Typically, these tests have very low power for detecting relatives more distant than sibs and first cousins (Queller and Goodnight 1989; Lynch and Ritland 1999). Finally, human geneticists have been very successful in using marker information to assess very deep relationships, on the order of hundreds (or more typically thousands) of generations. Many of these deep relationship studies have used the haplotypes of nonrecombining chromosomes, such as the nonpseudoautosomal arm of the Y (*e.g.*, Hammer 1995; Deka *et al.* 1996; Skorecki *et al.* 1997; Bianchi *et al.* 1998; Kittles *et al.* 1998; Wilson and Balding 1998; Thomas *et al.* 2000) and mitochrondial DNA (*e.g.*, Torroni *et al.* 1994; Merriwether *et al.* 1995; Forster *et al.* 1996; Brown *et al.* 1998; Stone and Stoneking 1998; Torroni *et al.* 1998). Here we show how haplotype information can also be used to estimate the age of the common ancestor for individuals sharing an intermediate ancestry (tens to hundreds of generations) with reasonable precision. Since our comparison is restricted to the time to the common ancestor for a particular pair of individuals of interest, the fine details of the population history and structure do not enter into the analysis, other than very weakly through the mean of the assumed prior (as discussed below).

For unlinked markers, the product rule (multiplying single-locus genotype probabilities together to obtain a multilocus genotype probability) applied to highly polymorphic loci allows just a few (5–10) unlinked markers to be quite sufficient for identifying individuals that share a common relative in the last generation (such as parent-offspring or sibs). However, because of recombination, unlinked markers have very weak power for distinguishing individuals sharing deeper common ancestry. While there is a growing body of literature on estimating relatedness of two individuals given autosomal marker information (Thompson 1975; Queller and Goodnight 1989; Blouin *et al.* 1996; Ritland 1996; Marshall *et al.* 1998; Lynch and Ritland 1999), it is less clear how to proceed when using markers from nonrecombining DNAs. The product rule does not hold for such regions, as the markers are inherited as a single block.

The key to assessing the amount of relatedness using markers on a nonrecombining chromosome is that any two individuals (indeed, the entire population) will have a most recent common ancestor for that region. This follows from coalescence theory (Hudson 1991; Donnelly and Tavaré 1995), which shows that in a demographically stable population the expected time back to this most recent common ancestor (MRCA) follows a geometric distribution with parameter $\lambda = 1/N_e$, the

Author e-mail: jbwalsh@u.arizona.edu

inverse of the appropriate measure of effective population size [the effective numbers N_m and N_f of males and females for the Y and mitochondrial (mt)DNAs, respectively]. Using marker information, we can estimate the time t to the MRCA, and it is the distribution of t that provides a natural metric for describing the relatedness (at least for that region) between two individuals. It is important to stress this difference between the MRCA of a region of DNA and the MRCA for two individuals. Two individuals can easily share an ancestor that is more recent than their MRCA for a particular DNA region. Hence, estimates of the time to the MRCA using a particular DNA region provide an upper bound on the time back to the most recent common ancestor shared by two individuals.

Here we develop a Bayesian estimator for t, obtaining the complete posterior distribution for the time t to the MRCA. We assume that n markers are scored on the nonrecombining chromosome of interest (either the Y or mtDNA) and that we observe matches in allelic state at k of these. We start by assuming the infinite alleles model, where each mutation is assumed to be unique. We then modify our results by assuming a (symmetric single-step) stepwise mutational model, which is a more appropriate descriptor for microsatellite markers. As we show, when k/n is close to one and n moderate to large, the two different mutational models give essentially the same distribution for t.

THE INFINITE ALLELES MODEL

Suppose we score the allelic states at n defined markers on a nonrecombining segment of DNA for two individuals and we wish to estimate the time back to their MRCA (for this segment). We first assume for each marker that (at most) only a single mutation has occurred over both lineages leading from the MRCA to the two individuals being considered. This is not an unreasonable assumption if the individuals match at most markers. We refer to this model as the infinite alleles model, as our development is also exact for the situation where each mutation is unique so that matches occur only when the marker locus in both lines has not mutated. Finally, we assume that the markers are exchangeable in the sense that the per-generation mutation rate μ is the same for each locus. We relax these assumptions below.

Let $q(t)$ be the probability of a match (at any given single marker) between two individuals with a most recent common ancestor t generations ago. The number of matches (k) out of n loci follows a binomial distribution, with

$$\Pr(k) = \frac{n!}{(n-k)!k!} q(t)^k [1 - q(t)]^{n-k}. \quad (1)$$

Ignoring matches created by parallel and/or back mutations, the probability that a marker matches after t generations is simply $(1-\mu)^{2t}$, as the probability of no mutations occurring in the lineage from the ancestor to individual one is $(1-\mu)^t$, with a similar probability from the MRCA to individual two. Hence,

$$q(t) = (1-\mu)^{2t} \simeq e^{-2\mu t} = e^{-\tau}, \quad (2a)$$

where

$$\tau = 2\mu t \quad (2b)$$

is the time scaled as total generations of divergence times the mutation rate. Note that the expression given by Equation 2a is a lower bound for the probability of a match, as if back mutations occurred in one (or both) lineages, or, if both lineages experienced parallel mutations, we also observe a match. These types of matches require at least two mutational events, and hence from the first two terms of the Poisson distribution their probability is bounded above by $1 - \exp(-2\mu t)(1 + 2\mu t)$, *i.e.*, on the order of $(2\mu t)^2\exp(-2\mu t)$. Thus if $2\mu t \ll 1$, the effects of such multiple mutations have only a trivial effect on increasing $q(t)$ over the value assuming no mutations (see Figure 4). When a specific value of μ is required, we generally use $\mu = 1/500 = 0.20\%$, motivated by estimates for Y chromosome microsatellites of 0.28% (Kayser *et al.* 2000) and 0.21% (Heyer *et al.* 1997), which are very similar to the estimated mutation rates of 0.1 to 0.21% for autosomal microsatellites (Weber and Wong 1993; Brinkmann *et al.* 1998). We note that these mutation rate estimates are generally done by scoring microsatellites already known to be polymorphic, which introduces a slight ascertainment bias. However, since we assume the markers being scored are also chosen because they are known to be polymorphic (in the population as a whole), then these potentially biased estimates of mutation rates are appropriate for our analysis.

From Equations 1 and 2a, the resulting likelihood for the time t back to the MRCA given that we observe k out of n matches is

$$L(t|n, k) = \frac{n!}{(n-k)!k!} e^{-k\tau}(1 - e^{-\tau})^{n-k}. \quad (3)$$

Setting the derivative of $\ln(L)$ equal to zero gives the maximum-likelihood estimate (MLE) for $\hat{\tau} = 2\hat{t}\mu$ as

$$\hat{\tau} = 2\hat{t}\mu = \ln\left(\frac{n}{k}\right). \quad (4)$$

Hence, the MLE for the time back to the MRCA becomes

$$\hat{t} = \frac{1}{2\mu} \ln\left(\frac{n}{k}\right). \quad (5)$$

Note that the MLE is not especially informative, as the distribution for t is highly positively skewed, resulting in a considerable variance and highly asymmetric confidence intervals about the MLE. In particular, note that

the MLE is zero for all values of n when there are no mismatches ($k = n$), which tells us nothing about the possible restrictions on the maximal time back to the MRCA (see Fu and Li 1996 and Donnelly *et al.* 1996 for a related discussion).

BAYESIAN POSTERIOR DISTRIBUTIONS FOR TIME TO MRCA

While the MLE describes one feature of the distribution of t (the mode), the most complete picture is given by the full posterior distribution of t, which can be obtained by a Bayesian analysis (*e.g.*, Lee 1997). Such an analysis proceeds from Bayes' theorem, with the posterior distribution $p(t|k)$ being proportional to the product of a prior distribution $p(t)$ for t and a likelihood $L(t|n, k)$ given the data (k out of n matches),

$$p(t|k) \propto L(t|n, k)p(t). \tag{6}$$

The main objection to a Bayesian analysis raised by non-Bayesians is that the choice of a prior is often very subjective and, as such, this can greatly bias the posterior. For the time back to the MRCA, an appropriate prior naturally follows from standard coalescence theory, as the expected time back to a MRCA under pure drift in an effective population size of N_e follows the geometric distribution with success parameter N_e^{-1} (*e.g.*, Wilson and Balding 1998). The parameter is $1/N_e$ in this case [as opposed to $1/(2N_e)$ for an autosomal gene] because the uniparental inheritance means that both mtDNA and the Y chromosome are essentially haploid. As summarized by Hammer (1995), estimates for N_e based on the standing level of variation at presumably neutral markers are on the order of 5000 in humans for both mtDNA and the male-specific region of the Y chromosome.

Treating time as continuous, the geometric prior is equivalent to using an exponential distribution with hyperparameter $\lambda = N_e^{-1}$. Thus, the natural prior for the time to MRCA (in the absence of any marker information) is to use

$$p(t) = \lambda \exp(-\lambda t), \quad \text{where } \lambda = N_e^{-1}. \tag{7}$$

Taking the limit as $\lambda \to 0$ gives an (improper) flat prior. As we will shortly see, the actual value of λ used has at best a trivial effect on the posterior distribution unless most markers do not match ($k \ll n$) and the effective population size is extremely small. Thus the prior is both well motivated and the choice of the prior hyperparameter (λ) has very little effect on the final (posterior) distribution in most cases.

Recalling Equation 3, the resulting posterior distribution becomes

$$p(t|k) \propto L(t|n, k)p(t) = \exp[-(2\mu k + \lambda)t]\ (1 - \exp[-(2\mu t)])^{n-k} \tag{8a}$$

so that

$$p(t|k) = \frac{\exp[-(2\mu k + \lambda)t](1 - \exp[-(2\mu t)])^{n-k}}{I(\mu, k, n, \lambda)}, \tag{8b}$$

where the normalizing constant is given by

$$I(\mu, k, n, \lambda) = \int_0^\infty \exp[-(2\mu k + \lambda)t](1 - \exp[-(2\mu t)])^{n-k}\, dt. \tag{9}$$

The impact of the choice of the hyperparameter λ for the prior immediately follows from Equation 8a. Provided $2\mu k \gg \lambda$, alternate choices of λ have little effect on the posterior distribution. Since $\lambda = N_e^{-1}$, this rearranges to $2N_e\mu k \gg 1$. For a mutation rate of $\mu = 1/500$, the choice of N_e (and hence λ) has essentially no effect on the prior provided $N_e k \gg 250$, which is a very mild restriction.

Returning to the posterior distribution, the normalizing constant is easily computed by expanding the $(1 - e^{-2\mu t})^{n-k}$ term, noting that we can express the function being integrated as

$$\exp[-(2\mu k + \lambda)t]\left(\sum_{i=0}^{n-k}(-1)^i\frac{(n-k)!}{i!(n-k-i)!}\exp[-(2\mu ti)]\right)$$
$$= \sum_{i=0}^{n-k}(-1)^i\frac{(n-k)!}{i!(n-k-i)!}\exp[-(2\mu(k+i)+\lambda)t]. \tag{10}$$

Thus

$$I(\mu, k, n, \lambda) = \sum_{i=0}^{n-k}(-1)^i\frac{(n-k)!}{i!(n-k-i)!}\int_0^\infty \exp[-(2\mu(k+i)+\lambda)t]\,dt$$
$$= \sum_{i=0}^{n-k}(-1)^i\frac{(n-k)!}{i!(n-k-i)!}\frac{1}{2\mu(k+i)+\lambda}$$
$$= \frac{2^{n-k}(n-k)!\mu^{n-k}}{\prod_{i=0}^{n-k}[\lambda + 2\mu(n-i)]}. \tag{11a}$$

The last step can be shown either by induction or by using a standard symbolic algebra package (such as Mathematica).

With a flat prior ($\lambda = 0$) the normalizing term further simplifies to

$$I(\mu, k, n, 0) = \frac{(n-k)!(k-1)!}{(2\mu)n!}. \tag{11b}$$

Hence, the posterior density becomes

$$p(t|k, \lambda) = \left(\frac{\prod_{i=0}^{n-k}[\lambda + 2\mu(n-i)]}{2^{n-k}(n-k)!\mu^{n-k}}\right)\frac{(1 - \exp[-2\mu t])^{n-k}}{\exp[t(2\mu k + \lambda)]}. \tag{12}$$

For zero marker mismatches ($k = n$), the posterior is simply an exponential distribution with parameter $\lambda + 2n\mu$,

$$p(t|k = n) = (\lambda + 2n\mu)\exp[-(2\mu n + \lambda)t]. \tag{13}$$

It immediately follows that the mean (μ_t) and standard deviation (σ_t) for the time to MRCA when there are no mismatches are

$$\mu_t = \sigma_t = \frac{1}{\lambda + 2n\mu} \simeq \frac{1}{2n\mu}. \quad (14a)$$

Likewise, the cumulative probability distribution for the time back to the MRCA is just

$$\Pr(t \le T) = \int_0^T p(t|k = n)\,dt = 1 - \exp(-(2\mu n + \lambda)T). \quad (14b)$$

In particular, the time T_α satisfying $\Pr(t \le T_\alpha) = \alpha$ is given by

$$T_\alpha = \frac{-\ln(1 - \alpha)}{2\mu n + \lambda}. \quad (14c)$$

Assuming a flat prior ($\lambda = 0$), if two individuals are identical at all n marker loci, there is a 90% probability they shared a MRCA in the last $1.15/(\mu n)$ generations. The values for 50, 95, and 99% are 0.347, 1.498, and 2.303 $(\mu n)^{-1}$ generations, respectively.

The posterior distributions with one or more mismatches also follow from Equation 12. For example, for one ($k = n - 1$) and two ($k = n - 2$) mismatches, the posteriors are

$$p(t|k = n - 1) = \left(\frac{(\lambda + 2n\mu)(\lambda + 2\mu[n - 1])}{2\mu}\right)\frac{1 - \exp[-(2\mu t)]}{\exp[t(2\mu(n - 1) + \lambda)]}$$

and

$$p(t|k = n - 2) = \left(\frac{(\lambda + 2n\mu)(\lambda + 2\mu[n - 1])(\lambda + 2\mu[n - 2])}{8\mu^2}\right) \times \frac{(1 - \exp[-(2\mu t)])^2}{\exp[t(2\mu(n - 2) + \lambda)]}.$$

The mean and variance for the posterior distribution for any value of k again follow by expanding the $(1 - e^{-2\mu t})^{n-k}$ term. Define

$$h(\mu, k, n, \lambda) = \int_0^\infty t \exp[-(2\mu k + \lambda)t](1 - \exp[-2\mu t])^{n-k}\,dt \quad (15a)$$

and

$$g(\mu, k, n, \lambda) = \int_0^\infty t^2 \exp[-(2\mu k + \lambda)t](1 - \exp[-2\mu t])^{n-k}\,dt. \quad (15b)$$

Expanding and term-by-term integration gives

$$h(\mu, k, n, \lambda) = \sum_{i=0}^{n-k}(-1)^i \frac{(n - k)!}{i!(n - k - i)!}\frac{1}{(2\mu(k + i) + \lambda)^2} \quad (16a)$$

$$g(\mu, k, n, \lambda) = \sum_{i=0}^{n-k}(-1)^i \frac{(n - k)!}{i!(n - k - i)!}\frac{2}{(2\mu(k + i) + \lambda)^3}. \quad (16b)$$

Hence, the mean and variance for the time t to the MRCA, given k of n marker loci match, a prior with hyperparameter λ, and a per marker mutation rate of μ, are given by

$$\mu_t = \frac{h(\mu, k, n, \lambda)}{I(\mu, k, n, \lambda)} \quad (17a)$$

and

$$\sigma^2(t) = \frac{g(\mu, k, n, \lambda)}{I(\mu, k, n, \lambda)} - \mu_t^2. \quad (17b)$$

Figures 1–3 plot the posterior distributions corresponding to different numbers of matches (k) for n = 5, 10, 20, 50, and 100 markers under the assumption that $\mu = 1/500$ and there is a flat prior ($\lambda = 0$). Provided that $N_e \gg 250$, the results are the same for any other prior using $\lambda = 1/N_e$.

Table 1 summarizes the mean, standard deviation (SD), and mode (the MLE) for the posterior distributions as well as the 2.5, 50, 90, and 97.5% cutoff values. Note that a 95% credible region for t is given by the 2.5 and 97.5% values. As expected, the distribution of t is highly skewed, as mode < median < mean. The distribution becomes increasingly skewed to the right as the number of mismatches increases, which is reflected in not only an increase in the mean but also in the variance. However, note that the mean/SD ratio declines with increasing numbers of mismatches, so that the coefficient of variation declines as k decreases.

Finally, note that the resolution for t offered by using $n = 5$ markers is very poor, but rather fine precision is offered by using 100 markers. While scoring the latter number of markers may be unrealistic, using 20 markers is both feasible as well as offering reasonable precision.

DIFFERENTIAL MUTATION RATES

As our knowledge of the parameters associated with the mutational process continues to improve, it is likely that we may find significant differences in the mutation rates at different markers. Fortunately, it is straightforward to modify the likelihood function (Equation 3) to take this into account. Suppose n markers are examined, generating the matching data $x_1, \ldots, x_n$, where the x_i are coded as

$$x_k = \begin{cases} 1 & \text{match at marker } k \\ 0 & \text{no match at marker } k. \end{cases}$$

The likelihood becomes

$$L(x_1, \ldots, x_n|t) = \prod_{k=1}^{n} q_k(t)^{x_k}[1 - q_k(t)]^{1-x_k}, \quad (18a)$$

where

$$q_k(t) = (1 - \mu_k)^{2t} \simeq e^{-2t\mu_k}, \quad (18b)$$

giving

$$L(x_1, \ldots, x_n|t) = \exp\left[-2t\sum_{k=1}^{n}\mu_k x_k\right]\prod_{k=1}^{n}[1 - e^{-2t\mu_i}]^{1-x_k}. \quad (18c)$$

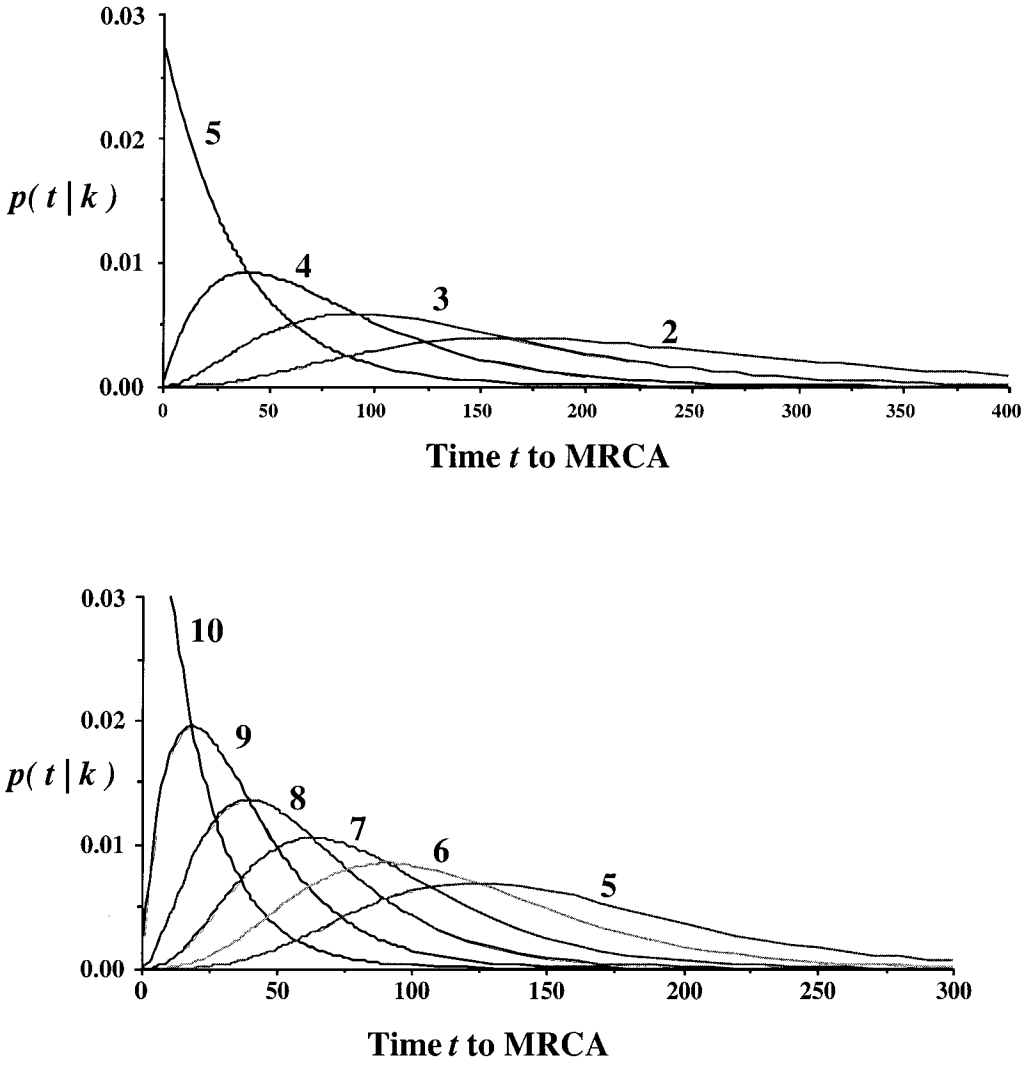


FIGURE 1.—The posterior distributions for the time to the most recent common ancestor (MRCA) between two individuals assuming 5 (top) and 10 (bottom) marker loci were scored and k matches are observed. A flat prior ($\lambda = 0$) and a mutation rate of $\mu = 1/500$ were assumed. Numbers indicate number of marker alleles k that match between the two individuals. Values for another mutation rate μ^* are given by scaling the values by μ^*/μ. Time is measured in generations.

Using an exponential prior with hyperparameter λ, the posterior is proportional to

$$p(t|\mathrm{x}) \propto \exp\left[-t\left(\lambda + 2\sum_{k=1}^{n}\mu_k x_k\right)\right]\prod_{k=1}^{n}[1 - e^{-2t\mu_i}]^{1-x_k}. \quad (19)$$

In any particular data set, the normalization constant is easily obtained by expanding the product involving $1 - e^{-2t\mu_i}$, which generally involves only a few terms unless there are a significant number of mismatches.

For the simplest case of no mismatches ($k = n$), the posterior for the time to the MRCA becomes

$$p(t|\text{no mismatches}) = (\lambda + 2n\bar{\mu})\exp[-t(\lambda + 2n\bar{\mu})], \quad (20a)$$

where $\bar{\mu}$ is the mean mutation rate across all markers. Equations 14a–c hold with μ replaced by the mean mutation rate $\bar{\mu}$. Likewise, for one mismatch (say marker i), the posterior becomes

$$\frac{[2(n-1)\bar{\mu}_{-i} + \lambda][2(n\bar{\mu}) + \lambda]}{2\mu_i}\exp[-t(\lambda + 2(n-1)\bar{\mu}_{-i})][1 - e^{-2t\mu_i}], \quad (20b)$$

where $\bar{\mu}_{-i}$ is the mean mutation rate for all markers, excluding marker i.

CORRECTING FOR MULTIPLE HITS: THE STEPWISE MUTATION MODEL

Microsatellites, given their higher mutation rates compared to single nucleotide polymorphisms (SNPs), are clearly the marker of choice for estimating the time to MRCA when the individuals are assumed to be at least modestly related. As microsatellite "alleles" correspond to different lengths of the repeat unit in the microsatellite array, the infinite alleles model assumed previously is not appropriate if multiple mutations are expected in any given marker. Since mutations change the number of repeats (and hence the size) of an array, two (or more) mutations can recover the initial state found in the MRCA. Likewise, parallel mutations in both lineages leading from the MRCA can also lead to the two individuals sharing the same allelic state, even though mutations have occurred. Thus, using an infinite

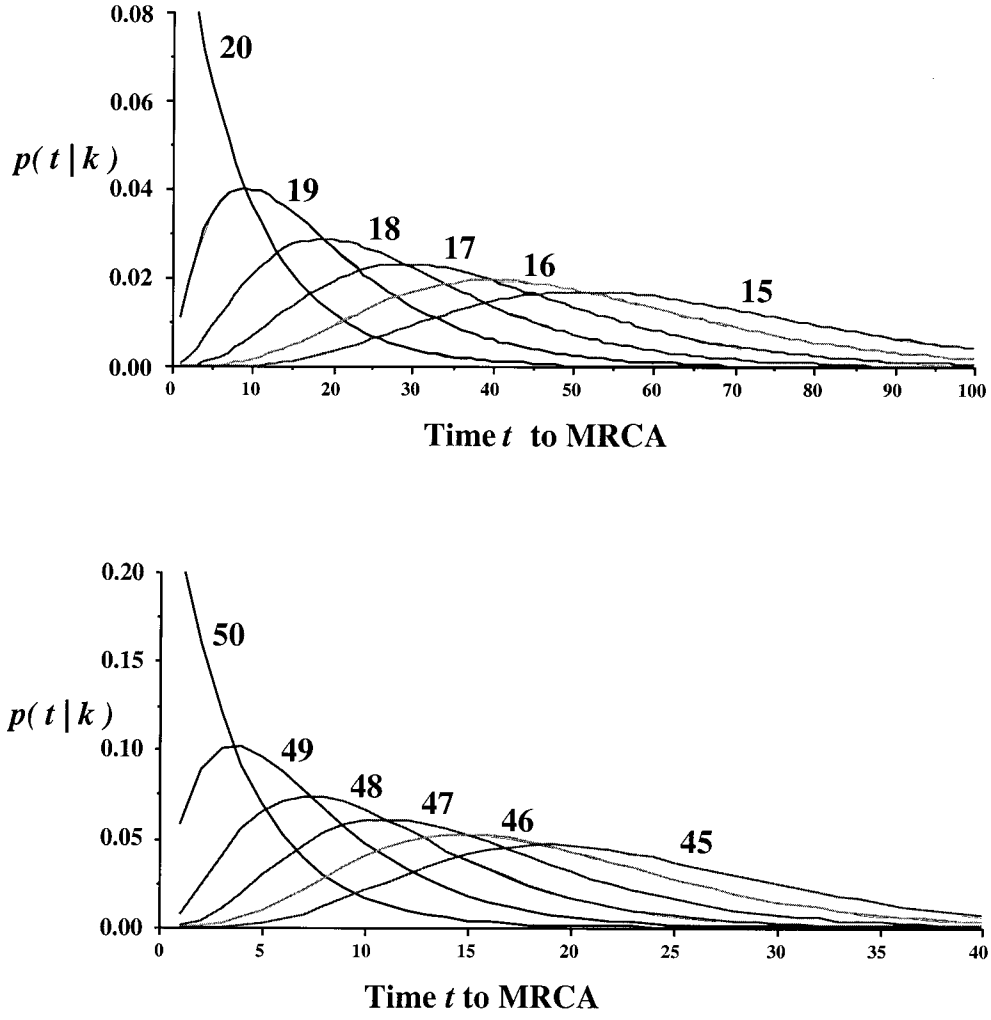


FIGURE 2.—Posterior distributions for time to MRCA for 20 (top) and 50 (bottom) markers. Details are as in Figure 1.

alleles model for microsatellites will tend to underestimate the time to the MRCA. To examine the severity of this underestimation, we consider the divergence, assuming a stepwise mutational model (SMM) assuming equal probabilities of an up (increase array size by one) or down (decrease array size by one) move (the symmetric single-step SMM). The roots of the SMM trace from OHTA and KIMURA's (1973) model of charge differences in electrophoretically scored proteins. A number of workers have shown this model to be a good fit for microsatellites in both indirect studies examining the distribution of array sizes in natural populations (EDWARDS *et al.* 1992; SHRIVER *et al.* 1993; VALDES *et al.* 1993; DI RIENZO *et al.* 1994) and direct studies looking at actual mutations arising in pedigrees (BRINKMANN *et al.* 1998; KAYSER *et al.* 2000). In the latter two studies, the vast majority (35 out of 37) of new mutations were single step, while the remaining (2 out of 37) were two step.

Denote the allelic state (array size) in the MRCA as state 0, and let X_t denote the array size at time t. As before we assume a per-generation mutation rate of μ. Under this model, the transition probabilities between states become

$$\Pr(X(t+1) = i-1 | X(t) = i) = \Pr(X(t+1) = i+1 | X(t) = i) = \frac{\mu}{2}$$

$$\Pr(X(t+1) = i | X(t) = i) = 1 - \mu$$

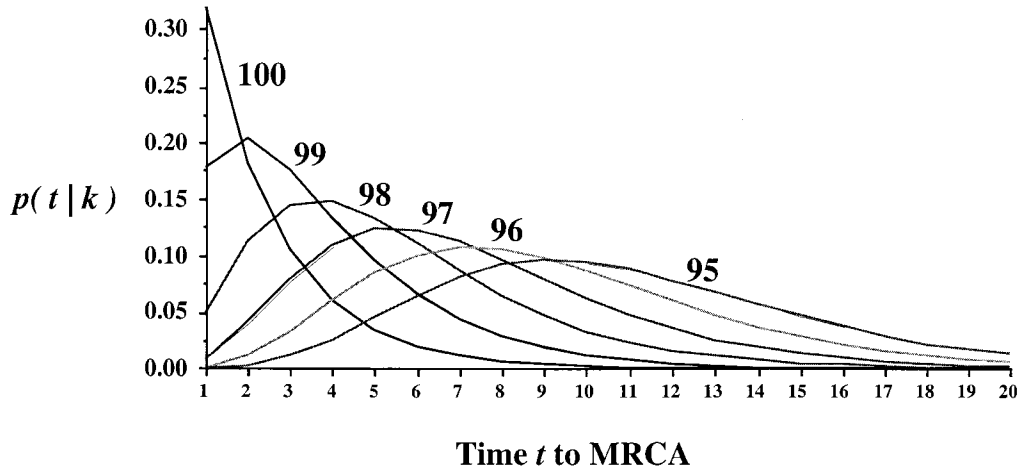
$$\Pr(|X(t+1) - X(t)| \geq 2 | X(t) = i) = 0. \qquad (21)$$

To apply this model, we need to compute $q(t)$, the probability of a match between two lineages sharing a MRCA t generations ago. A little thought shows that for the one-step model allelic states in two lineages can only match if an even number ($2M$) of mutations have occurred. The APPENDIX shows that

$$\Pr(\text{match} | 2M \text{ moves}) = \frac{1}{2^{2M}} \binom{2M}{M} = \frac{1}{2^{2M}} \frac{(2M)!}{(M!)^2}. \qquad (22)$$

For example, after a total (between both lineages) of 2, 4, 6, 8, and 10 mutations, the probabilities that the marker allelic states match are 0.5, 0.375, 0.313, 0.273, and 0.246 (respectively). Thus, under this model there is a one in four chance that the two lineages share the same allelic state even after a total of 10 mutations have occurred.

Since the probability of $2M$ total mutations along both



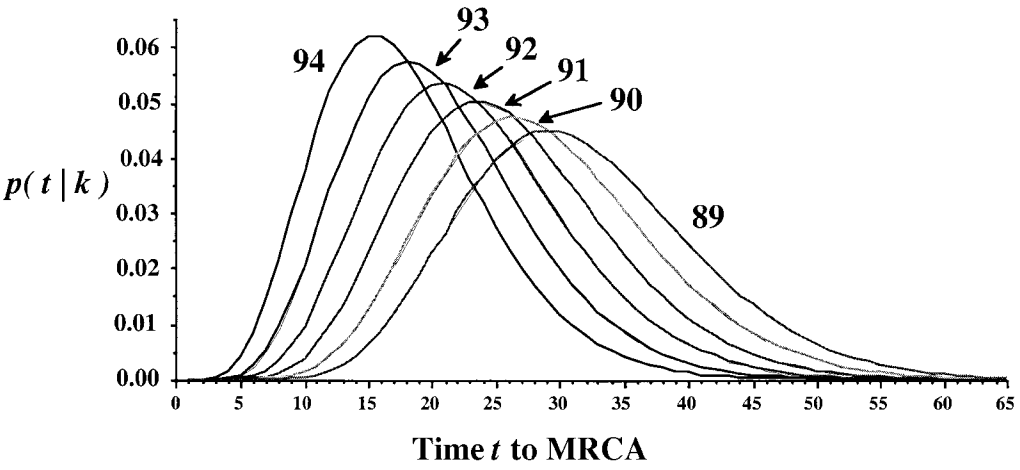


FIGURE 3.—Posterior distributions for time to MRCA for 100 markers. Details are as in Figure 1.

lineages assuming t generations back to the MRCA is given by a Poisson distribution with parameter $2t\mu$, we have the probability of a match conditioned on t as

$$\begin{aligned} p(t) &= \sum_{M=0}^{\infty} \Pr(\text{match}|2M \text{ moves})\Pr(2M \text{ moves}|t) \\ &= \sum_{M=0}^{\infty}\left(\frac{1}{2^{2M}}\frac{(2M)!}{(M!)^2}\right)\left(\frac{(2\mu t)^{2M}}{(2M)!}\right)\exp(-2t\mu) \\ &= \exp(-2t\mu)\left(\sum_{M=0}^{\infty}\frac{(\mu t)^{2M}}{(M!)^2}\right). \end{aligned} \quad (23)$$

For example, considering only the first 10 mutations,

$$p(t) = \exp(-2t\mu)\left(1 + (t\mu)^2 + \frac{(t\mu)^4}{4} + \frac{(t\mu)^6}{36} + \frac{(t\mu)^8}{576} + \frac{(t\mu)^{10}}{14{,}400}\right).$$

The general solution to Equation 23 follows by noting that

$$\sum_{k=0}^{\infty}\frac{(x)^{2k}}{(k!)^2} = \mathrm{I}_0(2x), \quad (24)$$

where I_0 denotes the zero-order modified type I Bessel function (OLVER 1964). Hence, under the single-step mutational model, the probability of a match after $\tau = 2\mu t$ generations is

$$q(\tau) = \exp(-\tau)\mathrm{I}_0(\tau). \quad (25)$$

Figure 4 compares the probability of a match as a function of τ for the infinite allele and one-step models. The match probabilities under both models are rather similar for $\tau < 0.5$ (125 generations with $\mu = 1/500$), but they diverge rather quickly after that. Note that even after 20τ generations (5000 generations with $\mu = 1/500$), the match probability under the stepwise model is still nontrivial (0.09), reflecting the very slow decrease in the probability of a match as the total number of mutations increases. By contrast, the corresponding match probability is essentially zero (2.06×10^{-9}) under the infinite alleles model, as a single mutation causes the allelic states to diverge and subsequent back (and/or parallel) mutations are not allowed.

As shown in Table 2, both the mean and variance (measured by the standard deviation) of the distribution for time to the MRCA are larger under the stepwise model than the infinite alleles model. This is certainly

TABLE 1

Summary of the posterior distribution $p(t|k, n)$, where t is the time back to the most recent common ancestor (MRCA) for two individuals that match at k of n markers

k	MLE	Mean	SD	Median	$t_{0.9}$	$t_{0.025}$	$t_{0.975}$
			$n = 5$ markers				
5	0.0	50.0	50.0	34.7	115.1	1.3	184.4
4	55.8	112.5	80.0	94.2	219.2	13.6	315.0
3	127.7	195.8	115.5	173.3	480.0	39.6	480.0
2	229.0	320.8	170.2	289.7	546.8	83.4	736.6
			$n = 10$ markers				
10	0.0	25.0	25.0	17.3	57.5	0.6	92.2
9	26.3	52.8	37.4	44.3	102.7	6.4	147.2
8	55.8	84.0	48.7	74.8	149.3	17.3	203.0
7	89.2	119.7	60.4	109.7	200.6	32.4	264.2
6	127.7	161.4	73.4	150.2	259.5	51.8	334.5
5	173.3	211.4	88.8	198.7	329.8	76.1	419.0
4	229.1	273.9	108.6	258.9	418.4	106.8	526.9
			$n = 20$ markers				
20	0.0	12.5	12.5	8.7	28.8	0.3	46.1
19	12.8	25.7	18.1	21.5	49.9	3.1	71.5
18	26.3	39.5	22.9	35.2	70.2	8.1	95.3
17	40.6	54.3	27.2	49.8	90.7	14.7	119.1
16	55.8	69.9	31.3	65.2	111.8	22.6	143.5
15	71.9	86.5	35.5	81.7	134.0	31.6	168.9
14	89.2	104.4	39.7	99.4	157.5	41.8	195.7
13	107.7	123.6	44.1	118.4	182.5	53.1	224.2
12	127.7	144.5	48.8	138.9	209.5	65.5	255.0
11	149.5	167.2	53.8	161.3	238.8	79.3	288.6
10	173.3	192.1	59.4	185.9	271.0	94.7	325.5
			$n = 50$ markers				
50	0.0	5.0	5.0	3.5	11.4	0.1	18.5
49	5.1	10.1	7.1	8.5	19.7	1.2	28.2
48	10.2	15.3	8.8	13.7	27.2	3.2	36.9
47	15.5	20.6	10.3	18.9	34.5	5.6	45.2
46	20.8	26.0	11.7	24.3	41.7	8.5	53.4
45	26.3	31.6	12.9	29.9	48.9	11.6	61.5
44	32.0	37.3	14.1	35.5	56.1	15.0	69.6
43	37.7	43.1	15.3	41.3	63.4	18.6	77.8
42	43.6	49.1	16.4	47.3	70.9	22.4	86.0
41	49.6	55.2	17.5	53.3	78.4	26.4	94.4
40	55.8	61.4	18.6	59.6	86.1	30.6	102.8
39	62.1	67.8	19.6	65.9	93.9	35.0	111.4
38	68.6	74.4	20.7	72.5	101.9	39.5	120.2
37	75.3	81.2	21.8	79.2	110.0	44.3	129.2
36	82.1	88.1	22.9	86.1	118.4	49.2	138.3
35	89.2	95.2	24.0	93.2	127.0	54.3	147.7
34	96.4	102.6	25.1	100.6	135.7	59.6	157.3
33	103.9	110.2	26.2	108.1	144.8	65.1	167.2
32	111.6	118.0	27.3	115.9	154.1	70.8	177.4
31	119.5	126.1	28.5	123.9	163.7	76.7	187.9
30	127.7	134.4	29.7	132.2	173.6	82.8	198.7

(*continued*)

expected, as the effect of the stepwise mutational model is to allow for a match following two (or more) mutations, while a match is assumed to never recover following a mutation under the infinite alleles model. Table 2 gives the ratios of the means and standard deviations for the distribution of t under the stepwise model compared to the same statistic under the infinite alleles model. When the number of markers is large (20 or greater) and the number of mismatches is small, these ratios are close to one. The ratio of standard deviations

TABLE 1

(Continued)

k	MLE	Mean	SD	Median	$t_{0.9}$	$t_{0.025}$	$t_{0.975}$
			$n = 100$ markers				
100	0.0	2.5	2.5	1.7	5.8	0.1	9.2
99	2.5	5.0	3.6	4.2	9.8	0.6	14.0
98	5.1	7.6	4.4	6.8	13.5	1.6	18.3
97	7.6	10.2	5.1	9.3	17.0	2.8	22.3
96	10.2	12.8	5.7	11.9	20.4	4.1	26.1
95	12.8	15.4	6.3	14.5	23.8	5.6	29.9
94	15.5	18.0	6.8	17.2	27.2	7.3	33.7
93	18.1	20.7	7.3	19.9	30.5	8.9	37.4
92	20.8	23.5	7.8	22.6	33.9	10.7	41.1
91	23.6	26.2	8.3	25.3	37.2	12.6	44.8
90	26.3	29.0	8.7	28.1	40.6	14.5	48.5
89	29.1	31.8	9.2	30.9	44.0	16.4	52.2
88	32.0	34.6	9.6	33.7	47.4	18.4	55.9
87	34.8	37.5	10.0	36.6	50.8	20.5	59.6
86	37.7	40.4	10.4	39.5	54.3	22.6	63.3
85	40.6	43.4	10.9	42.5	57.7	27.8	67.1
84	43.6	46.3	11.2	45.4	61.2	27.0	70.9
83	46.6	49.3	11.6	48.4	64.7	29.2	74.7
82	49.6	52.4	12.0	51.5	68.3	31.5	78.5
81	52.7	55.5	12.4	54.6	71.9	33.9	82.4
80	55.8	58.6	12.8	57.7	75.5	36.3	86.3
79	58.9	61.8	12.2	60.8	79.2	38.7	90.2
78	62.1	65.0	12.6	64.0	82.9	41.2	94.2
77	65.3	68.2	14.0	67.3	86.6	43.7	98.2
76	68.6	71.5	14.3	70.5	90.4	46.2	102.3
75	71.9	74.8	14.7	73.9	94.2	48.8	106.4
74	75.3	78.2	15.1	77.2	98.1	51.5	110.5
73	78.7	81.6	15.5	80.7	102.0	54.2	114.7
72	82.1	85.1	15.9	84.1	106.0	56.9	119.0
71	85.6	88.6	16.3	87.6	110.0	59.7	123.3
70	89.2	92.2	16.7	91.2	114.1	62.5	127.6
69	92.8	108.9	22.2	107.0	138.1	71.0	157.6
68	96.4	113.7	23.0	111.7	143.9	74.5	164.2
67	100.1	118.7	23.8	116.6	150.0	78.0	171.0
66	103.9	123.7	24.6	121.6	156.2	81.7	178.0
65	107.7	129.0	25.5	126.8	162.6	85.4	185.2
64	111.6	134.4	26.5	132	169.2	89.3	192.7
63	115.5	139.9	27.4	137.5	176.0	93.2	200.4
62	119.5	145.6	28.4	143.1	183.0	97.3	208.4
61	123.6	151.6	29.5	148.9	190.3	101.5	216.8
60	127.7	157.7	30.6	154.9	197.9	105.8	225.5

A flat (improper) prior was used ($\lambda = 0$) and a mutation rate of $\mu = 1/500$ was assumed. Results for any other mutation rate μ^* follow by multiplying the appropriate table entry by $\mu^*/\mu = 500 \cdot \mu^*$. MLE, maximum-likelihood estimate (which is also the mode of the posterior under a flat prior); SD, standard deviation; t_α satisfies $P(t \leq t_\alpha \mid k, n) = \alpha$. The median corresponds to $t_{0.5}$, while a 95% credible region is given by ($t_{0.025}$, $t_{0.975}$).

is always larger than the means ratio, reflecting the longer tail (relative to that for the infinite alleles model) generated under the stepwise model. As the number of mismatches increases, the mean and SD ratios increase, reflecting increasingly larger probabilities for t under the stepwise relative to the infinite alleles model.

Also note from Table 2 that for the same fraction of observed matches (say 4/5, 8/10, 16/20, 40/50, and 80/100), the ratios of the means and standard errors under the stepwise *vs.* the infinite alleles model decrease toward 1 as the number n of markers scored increases. For example, for 80% observed matches, assuming a flat prior ($\lambda = 0$), the mean ratios are 4.8, 1.34, 1.2, 1.1, and 1.1 (for $n = 5$, 10, 20, 50, and 100 markers). Likewise, the ratios of standard deviations are 34.2, 2.0, 1.3, 1.2, and 1.2. The reason for this can be seen by considering the case where we observe what appears to be a complete match at all n markers. In this case, there

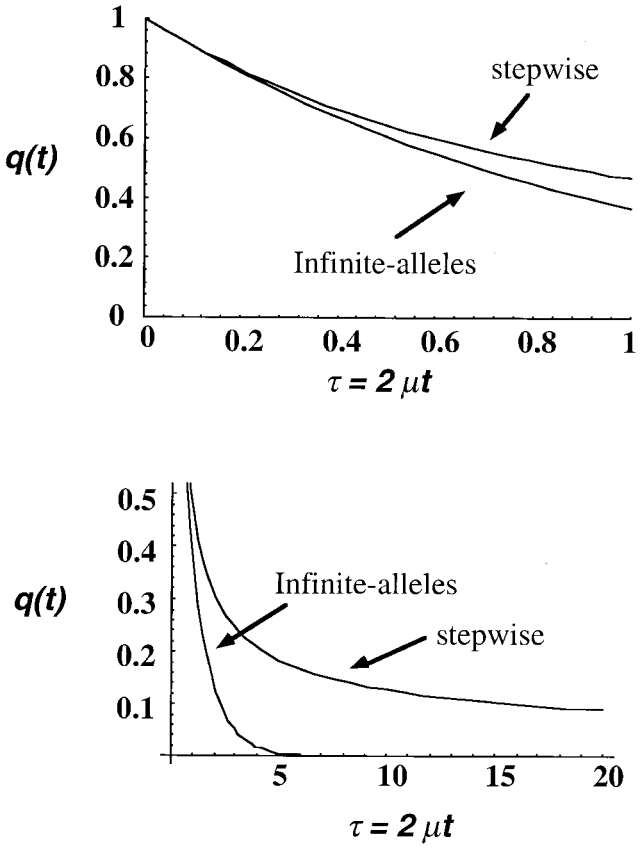


FIGURE 4.—The probability of a match in allelic state between two lineages with a MRCA t generations ago under the infinite alleles and stepwise models. Top, generations 0 to $\tau = 2\mu t = 1$, which corresponds to 250 generations for $\mu = 1/500$. Bottom, up through 20τ (5000 generations under this value of μ).

is some small chance that two (or more) mutations have occurred in one (or more) markers. However, if indeed a total of two mutations have occurred across both lineages, the probability that both occurred in the same marker is just $1/n$ (assuming the same mutation rate across markers). Hence, as the number of markers increases, the probability that a multiple mutation is masked decreases due to scoring changes over more loci.

The final observation from Table 2 is that, unlike for the infinite alleles model, the choice of the hyperparameter λ for the prior can significantly affect the posterior distribution. This is true when the number of markers is very small and/or the fraction of mismatches is nontrivial. The most extreme difference between the two assumed mutational models is seen under a flat prior ($\lambda = 0$). As the value of λ increases (corresponding to a decrease in the assumed effective population size as $\lambda = 1/N_e$), the ratios of the means and standard deviations under the two models become increasingly similar. Almost all of this difference is due to significant decreases in the mean and variance under the distribution of t under the stepwise model as λ increases. This arises because the effect of increasing λ is to shorten the distributions of times to MRCA in the prior (decreasing N_e decreases the coalescent times), which in turn down-weights matches under the stepwise model from individuals with assumed very long times to the MRCA.

Our analysis under the SMM model considers markers only as showing a match or mismatch, ignoring any additional information on the differences between marker alleles when a mismatch has occurred. Over short time scales (on the order of $1/2\mu < 1$) we are likely not losing much information, as most markers will likely have at most one mutation and hence scoring a match *vs.* no match is sufficient. Over longer time scales, we are clearly losing information. In such cases, a logical extension of our model would be replacing the probability of a match with the probability that the two microsatellite alleles in the individuals being compared differ by r repeats. This is accomplished as follows. Consider the probability of an even number ($2k$) of differences between microsatellite array sizes. Again, a little thought shows that this can occur only with an even number of total mutations ($2M$). With a total of $2M$ mutations, the probability that the array sizes differ by $2k$ is the probability that the number of up (+) mutations is either $M + k$ or $M - k$ (by symmetry, these two probabilities are the same under the SMM). Thus, the probability of a difference (in absolute value) of $\Delta = 2k$ given $2M$ total mutations is

$$\Pr(\Delta = 2k|2M) = 2\left(\frac{1}{2}\right)^{2M}\binom{2M}{M-k} \quad \text{for } k \le M. \tag{26}$$

Since M follows a Poisson distribution with mean $2\mu t$, the probability of a difference of $2k$ given the time t to the MRCA is

$$\begin{aligned}\Pr(\Delta = 2k|t) &= \sum_{M=k}^{\infty}\Pr(\Delta = 2k|2M)\Pr(2M|t)\\ &= \sum_{M=k}^{\infty}2\left(\frac{1}{2}\right)^{2M}\binom{2M}{M-k}e^{-2\mu t}\frac{(2\mu t)^{2M}}{(2M)!}\\ &= 2e^{-2\mu t}\sum_{M=k}^{\infty}\frac{(\mu t)^{2M}}{(M-k)!(M+k)!}\\ &= 2e^{-2\mu t}\mathrm{I}_{2k}(2\mu t)\end{aligned} \tag{27a}$$

where I_s denotes the s-order modified type I Bessel function (OLVER 1964).

Using the same logic, for a difference of $2k + 1$ an odd number ($2M + 1$) of total mutations are required, and following the same steps leading to Equation 27a gives

$$\Pr(\Delta = 2k + 1|t) = 2e^{-2\mu t}\,\mathrm{I}_{2k+1}(2\mu t). \tag{27b}$$

Hence, the probability that the array sizes for two alleles differ by j after $\tau = 2\mu t$ generations is

$$q^{(j)}(\tau) = 2\exp(-\tau)\mathrm{I}_j(\tau) \quad \text{for } j \ge 1. \tag{28}$$

Figure 5 plots the probability for 0, 1, 2, 3, and ≥ 4

TABLE 2

The increase in the mean and standard deviation (SD) for the time to the MRCA under the stepwise mutational model as compared to the infinite alleles model (Table 1)

	$\lambda = 0$		$\lambda = 1/20{,}000$		$\lambda = 1/500$	
k	Mean	SD	Mean	SD	Mean	SD
			$n = 5$ markers			
5	1.65	6.56	1.59	4.09	1.30	1.62
4	4.77	34.26	3.34	14.88	1.51	2.17
3	25.05	108.19	9.91	38.56	1.75	2.67
2	64.45	148.81	20.63	57.99	1.86	2.77
			$n = 10$ markers			
10	1.14	1.25	1.14	1.24	1.12	1.21
9	1.21	1.47	1.21	1.46	1.18	1.34
8	1.34	2.01	1.34	1.96	1.25	1.56
7	1.60	3.79	1.59	3.36	1.36	1.88
6	2.28	10.31	2.17	7.21	1.51	2.32
5	4.56	31.32	3.65	16.34	1.68	2.80
4	12.89	80.04	7.15	32.93	1.85	3.18
			$n = 20$ markers			
20	1.06	1.09	1.06	1.09	1.05	1.09
19	1.08	1.13	1.08	1.13	1.07	1.12
18	1.10	1.19	1.10	1.19	1.09	1.17
17	1.13	1.25	1.13	1.25	1.12	1.23
16	1.16	1.33	1.16	1.33	1.14	1.29
15	1.20	1.44	1.20	1.44	1.18	1.38
14	1.25	1.59	1.25	1.58	1.22	1.49
13	1.31	1.81	1.31	1.80	1.26	1.63
12	1.40	2.16	1.40	2.15	1.32	1.82
11	1.53	2.76	1.53	2.72	1.39	2.06
10	1.73	3.83	1.72	3.71	1.49	2.37

(*continued*)

differences in array size as a function of τ. F. Rousset (personal communication) kindly pointed out that Equation 28 can be found buried in Wehrhahn (1975), who obtained this result using the method of generating functions (also see Li 1976; Wilson and Balding 1998).

The likelihood function follows from the multinomial distribution. If a total of n markers are scored, and n_i denotes the number of markers differing in size by i (with the largest difference being k), then

$$L(t|n_0, \ldots, n_k) = \frac{n!}{n_0!n_1! \ldots n_k!} \prod_{j=0}^{k} [q^{(j)}(2\mu t)]^{n_j}. \quad (29)$$

Again using an exponential prior, the resulting posterior distribution is proportional to

$$p(t|n_0, \ldots, n_k) \propto \prod_{j=0}^{k} [q^{(j)}(2\mu t)]^{nj} e^{-\lambda t} = 2^{n-n_0} e^{-(\lambda+2\mu n)t} \prod_{j=0}^{k} [\mathrm{I}_j(2\mu t)]^{n_j}. \quad (30)$$

The full distribution is recovered by numerical integration to normalize the posterior, *viz.*,

$$p(t|n_0, \ldots, n_k) = \frac{e^{-(\lambda+2\mu n)t} \prod_{j=0}^{k} [\mathrm{I}_j(2\mu t)]^{n_j}}{\int_0^{\infty} e^{-(\lambda+2\mu n)t} \prod_{j=0}^{k} [\mathrm{I}_j(2\mu t)]^{n_j} dt}. \quad (31)$$

As an example of applying Equation 31, consider haplotypes 1 and 3 from Thomas *et al.*'s (2000) study on the Lemba and the Cohen (Y chromosome) modal haplotype. Six microsatellite markers were scored and, of these, both alleles match at four markers, while one marker differs by one repeat and another by two repeats. In this case, Equation 31 becomes

$$p(t|4, 1, 1) \propto e^{-(\lambda+2\mu 6)t} [\mathrm{I}_0(2\mu t)]^4 \cdot \mathrm{I}_1(2\mu t) \cdot \mathrm{I}_2(2\mu t).$$

Table 3 compares the estimated parameters under this model with those estimated using the infinite alleles and SMM matching models. Figure 6 plots the resulting posterior distributions. We use $\mu = 0.245\%$ (the average of the Kayser *et al.* 2000 and Heyer *et al.* 1997 estimates) and a prior of $\lambda = 1/5000$ (from Hammer's 1995 estimate of N_e for the Y) for these results.

While Equation 31 provides the foundation for a full Bayesian analysis, we caution that its usefulness depends on accurately capturing the mutation model for the markers in question. While the stepwise mutation model seems a reasonable fit, the fine details are still unclear. For example, there are suggestions that mutation rate may increase with array size (Brinkmann *et al.* 1998;

TABLE 2

(Continued)

	λ = 0		λ = 1/500			λ = 0		λ = 1/500	
k	Mean	SD	Mean	SD	*k*	Mean	SD	Mean	SD
				n = 50 markers					
50	1.02	1.03	1.02	1.03	39	1.11	1.22	1.10	1.21
49	1.03	1.04	1.03	1.04	38	1.12	1.25	1.11	1.24
48	1.03	1.06	1.03	1.06	37	1.13	1.27	1.12	1.26
47	1.04	1.07	1.04	1.07	36	1.14	1.30	1.13	1.29
46	1.05	1.09	1.05	1.08	35	1.15	1.34	1.15	1.32
45	1.05	1.10	1.05	1.10	34	1.17	1.38	1.16	1.36
44	1.06	1.12	1.06	1.11	33	1.18	1.42	1.17	1.39
43	1.07	1.13	1.07	1.13	32	1.20	1.46	1.90	1.44
42	1.08	1.15	1.08	1.15	31	1.22	1.52	1.21	1.48
41	1.09	1.17	1.08	1.17	30	1.24	1.58	1.22	1.53
40	1.10	1.20	1.09	1.19					
				n = 100 markers					
100	1.01	1.02	1.01	1.02	79	1.08	1.17	1.08	1.17
99	1.01	1.02	1.01	1.02	78	1.09	1.18	1.09	1.18
98	1.02	1.03	1.02	1.03	77	1.09	1.19	1.09	1.19
97	1.02	1.03	1.02	1.03	76	1.10	1.21	1.10	1.20
96	1.02	1.04	1.02	1.04	75	1.10	1.22	1.10	1.21
95	1.02	1.04	1.02	1.04	74	1.11	1.23	1.11	1.23
94	1.03	1.05	1.03	1.05	73	1.11	1.24	1.11	1.24
93	1.03	1.06	1.03	1.06	72	1.12	1.26	1.12	1.25
92	1.03	1.06	1.03	1.06	71	1.12	1.27	1.12	1.26
91	1.04	1.07	1.04	1.07	70	1.13	1.29	1.13	1.28
90	1.04	1.08	1.04	1.08	69	1.14	1.30	1.13	1.29
89	1.04	1.09	1.04	1.08	68	1.14	1.32	1.14	1.31
88	1.05	1.09	1.05	1.09	67	1.15	1.33	1.15	1.33
87	1.05	1.10	1.05	1.10	66	1.16	1.35	1.15	1.34
86	1.05	1.11	1.05	1.11	65	1.16	1.37	1.16	1.36
85	1.06	1.12	1.06	1.12	64	1.17	1.39	1.17	1.38
84	1.06	1.13	1.06	1.12	63	1.18	1.41	1.17	1.40
83	1.07	1.13	1.07	1.13	62	1.19	1.43	1.18	1.42
82	1.07	1.14	1.07	1.14	61	1.19	1.45	1.19	1.44
81	1.07	1.15	1.07	1.15	60	1.2	1.48	1.20	1.46
80	1.08	1.16	1.08	1.16					

When the number of markers and/or matches is low, the value of $\lambda = 1/N_e$ chosen for the prior distribution can have a nontrivial effect on the results under the stepwise model. For example, with two of five matches, the mean time under the stepwise model is over 64 times that of the infinite alleles model (assuming a flat prior, $\lambda = 0$, on both), but only a twofold (1.86) difference is seen when the prior assumes a hyperparameter of $\lambda = 1/500$ (corresponding to $N_e = 500$).

Fu and Chakraborty 1998) but also observations that suggest this is not the case (Valdes *et al.* 1993). There are also observations suggesting a bias toward increased array size (Kayser *et al.* 2000; Fu and Chakraborty 1998), and although most mutational steps change array size by one, mutations of two or more steps are seen (Brinkmann *et al.* 1998; Kayser *et al.* 2000). Finally, there may be multiple molecular processes operating at microsatellites, such as a majority of small changes against a background of rare major changes (Di Rienzo *et al.* 1994) or independent deletions (Walsh 1987). Until these details are sorted out, the risk run by using a model to correct for multiple mutations is that at least as much bias could be introduced by assuming an incorrect mutation model as is removed by accounting for multiple hits. One approach would be to use Fu and Chakraborty's (1998) minimum chi-square approach for estimating the generalized stepwise mutation model and compare the minimal (best) fitting parameters with those for the symmetric single-step model that we have assumed.

DISCUSSION

Both the Y chromosome and mtDNA have been successfully used for assessing deep ancestry, while unlinked autosomal markers have proven much more valuable for determining very recent ancestry. Here we

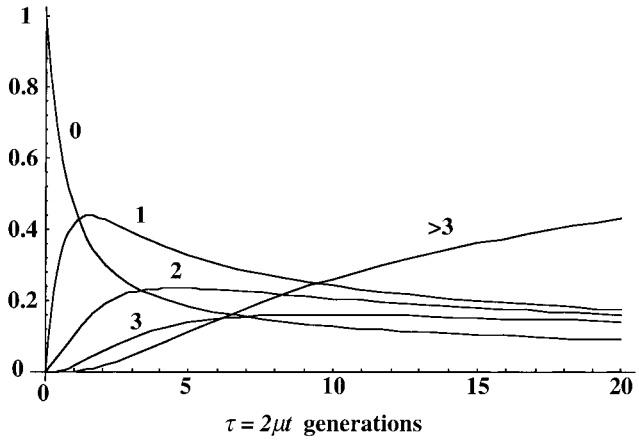


FIGURE 5.—Probabilities that two microsatellite alleles that have been separated for a total of $\tau = 2\mu t$ generations differ in array size by 0, 1, 2, 3, and >3 repeats (computed using Equation 28). The single-step symmetric stepwise mutational model is assumed.

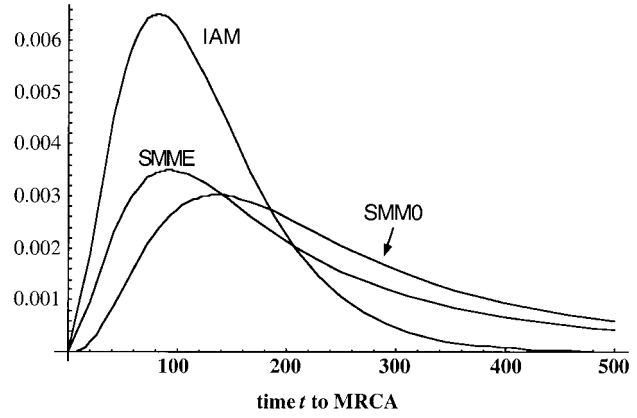


FIGURE 6.—Posterior distributions for t, the estimated time to MRCA, between Y chromosome haplotypes 1 and 3 of THOMAS *et al.* (2000). IAM, infinite alleles model; SMM0, stepwise mutational model only scoring matches *vs.* no match; SMME, stepwise mutation model scoring number of differences between array sizes. See the Table 3 legend for further details.

examine the effectiveness of using multilocus haplotypes from a region of nonrecombining DNA (such as the majority of Y chromosome or mtDNA) to estimate the time t to the MRCA of two individuals for that region. We argue that, to assess the relatedness of individuals on the basis of their haplotypes in a region of nonrecombining DNA, the time to the MRCA is the natural replacement for probability calculations based on the product rule used for unlinked markers.

We assume n prechosen markers are examined and are (initially) coded as either matching (agreeing in allelic state) or not matching. Using a Bayesian approach, the resulting posterior distribution for t is a function of n, the number of matches k, the per-marker mutation rate μ, and the hyperparameter $\lambda = 1/N_e$ (the reciprocal of the effective population size) of the assumed prior distribution of t (the assumed distribution for the time to the MRCA for two random individuals in the absence of any marker information). We examined two rather different mutation models—infinite alleles (each mutation is unique, with a match implying no mutation at that marker since the MRCA) and the SMM (which allows for matches created by several parallel mutations).

TABLE 3

Estimated time to MRCA between haplotypes 1 and 3 of THOMAS *et al.* (2000)

Model	Mean	Median	2.5%	97.5%
IAM	124.8	110.7	25.4	303.8
SMM0	394.4	193.5	33.2	2129.1
SMME	352.5	235.8	53.4	1380.5

Six microsatellite markers were scored, four of which were exact matches, one differed by one, and one by two. The parameter values used were $\lambda = 1/5000$, $\mu = 0.0025$. IAM, infinite alleles model; SMM0, stepwise mutational model only scoring matches *vs.* no match; SMME, stepwise mutation model scoring number of differences between array sizes.

Our results show that it is possible to use Y or mtDNA marker information to provide reasonable estimates for t when individuals share a MRCA of intermediate age (tens to hundreds of generations when $\mu = 10^{-3}$), provided a sufficient number of markers are scored (Table 1, Figures 1–3). Estimates using only 5 markers have distributions that are highly skewed for large t values, especially when the possibility of back mutations regenerating a match is taken into account (*i.e.*, the SMM model). As the number of markers increases, the width of the credible intervals around an estimate of t decreases. While 10 markers give a reasonable interval, 20 markers seems a more workable tradeoff between cost and precision. Our results also suggest that the forensic use of either the Y or mtDNA is rather limited. They can be used for exclusion, but make only weak probability statements about a sample and a suspect when there is a complete match. Using Equation 14c, a complete match at 10 out of 10 markers between a sample and a suspect implies only that the individual generating the sample and the suspect have a 90% chance of sharing a MRCA no more than 58 generations ago (assuming a typical microsatellite mutation rate of $\mu = 1/500$). With a complete match, the number of markers that need to be scored to have a 90% probability that the sample and suspect have a MRCA no more than 1 generation ago is ~580, an order of magnitude more than the number of currently known Y-linked markers (M. HAMMER, personal communication). A 50% probability that the MRCA does not exceed 1 generation requires roughly 340 microsatellite markers.

We obtained the posterior distributions for t given the marker data by using a Bayesian analysis, which requires a prior distribution for t. From standard population-genetics theory, t follows a geometric distribution with success parameter $1/N_e$ (as the Y and mtDNA behave as haploids). Hence, the functional form of the prior is well motivated, while its exact shape is determined by the hyperparameter $\lambda = 1/N_e$ used for any particular prior. For the infinite alleles model, the actual value of λ had essentially no effect unless we used a very small value for N_e (<200) and/or there were a very significant number of mismatches ($k/n \ll 1$). Thus, we used a flat prior ($\lambda = 0$, corresponding to an infinite effective population size) for the general results tabulated, although the equations given cover arbitrary values of $\lambda \geq 0$. Conversely, under the stepwise model, the value of λ often had a significant effect on the posterior, especially when the number of markers n is small and the number of matches is modest or poor ($k/n \ll 1$). In such cases, the use of a flat prior gave the largest difference in the posteriors for the two mutational models. As the value of λ is increased (corresponding to decreasing N_e), the posteriors under the two different models become increasingly similar (Table 2). This occurs because, as we decrease N_e, we make the initial assumption that individuals with long times to the MRCA become increasingly unlikely. It is these individuals that still have a modest probability of showing a match under the stepwise model, greatly inflating the estimated times to the MRCA relative to the infinite alleles estimate. In cases where there was a recent population expansion (or contraction) or a selective sweep, the distribution of times to MRCA may deviate from a geometric distribution. However, even in such cases, we expect there to be little dependence on the prior except in cases where changes in λ under a geometric prior have a significant effect on the posterior.

One potential issue of concern is whether our results are somehow biased by scoring only markers known to be polymorphic in the human population as a whole (but not necessarily in the particular pair of individuals being contrasted). When the goal is to estimate the coalescent time for an entire population, using markers known (in advance) to be polymorphic in the population of interest creates an ascertainment bias that needs to be corrected (*e.g.*, Nielsen 2000). In our analysis, μ is the mutation rate for microsatellites conditional on their being polymorphic. However, since the estimates of microsatellite mutation rates used are also ascertained by scoring microsatellites known to be polymorphic, we have corrected for this effect. When considering random microsatellites, the mutation rate μ in our analysis is replaced by $c\mu$, where $c > 1$ is an ascertainment correction that can be specified only by knowing how the polymorphic markers were ascertained.

While we have implicitly framed much of our discussion in terms of microsatellites [simple tandem repeats (STRs)], SNP data can be included by using the infinite alleles model and the appropriate mutation rates. Equation 18a allows the method to handle mixed (STRs and SNPs) marker data by using the appropriate expression for $q_k(t)$, the probability of a match for marker k at time t. For SNPs, the infinite alleles model is used, $q_k(t) = \exp(-2\mu_k t)$, while microsatellites use Equation 25, $q_k(t) = \exp(-2\mu_k t)I_0(2\mu_k t)$, which corrects for multiple hits under the stepwise mutation model (I_0 denoting the zero-order modified type I Bessel function; Olver 1964). In either case, μ_k is the mutation rate for marker locus k. More generally, if microsatellite comparisons are coded by differences in size (as opposed to match/no match), Equation 28 gives the probability that two individuals differ by j steps at the kth microsatellite as $q_k^{(j)}(\tau) = 2\exp(-2\mu_k t)I_j(2\mu_k t)$, where I_j is the jth-order modified type I Bessel function.

Our approach is easily modified to estimate the time between an individual and a particular (known or inferred) ancestral haplotype. When comparing an individual's haplotype against a fixed standard, we are following only one branch from the MRCA, so that $q(t) = (1 - \mu)^t \sim \exp(-\mu t)$, as opposed to $q(t) \sim \exp(-2\mu t)$ when looking at mutations over both branches. Hence, we simply replace the mutation rate μ by $\mu/2$, and all of our previous results apply.

Estimation of the time to the MRCA for a sample of individuals (as opposed to our simpler setting of just two individuals) has been examined by Fu (1996) and Tavaré *et al.* (1997), under the assumption of an infinite sites model (Watterson 1975). These analyses assume a Poisson likelihood for the number of segregating sites, while we assumed a binomial likelihood for number of segregating sites under the assumption that the n sites to be scored in the two individuals were fixed in advance. Given that the rough figure is one common polymorphism every 10 kb for the human Y (M. Hammer, personal communication), focusing on specific sites known to be polymorphic in the population as a whole (as opposed to sequencing large regions) is not an unreasonable approach. When n is large and μ small, the two different likelihoods for estimating t should give very similar results.

As we tried to stress in several places, the major caveat to this (or any other) approach for estimating the time t to MRCA is our uncertainty about both the mutational process and rates. The Bayesian framework allows us to incorporate these uncertainties; for example, if $p(\mu)$ is some assumed prior for the mutation rates, then the marginal posterior (after integrating out μ) can be used to estimate t,

$$p(t|\lambda, \text{marker information}) \propto \int L(t|\text{marker information})\, p(\lambda)\, p(\mu)\, d\mu.$$

However, practical application requires a reasonable prior for μ. An even more serious problem is the exact form the stepwise mutational model used for microsatellite loci. Using an inappropriate model can potentially introduce more bias than it corrects.

Thanks go to Mike Lynch, Mike Hammer, Monty Slatkin, Francois Rousset, and Yun-Xin Fu for useful comments and advice and to Bennet Greenspan and Max Rothschild for dragging me into this problem in the first place. Special thanks go to Jay Taylor for very useful discussions on the probability of a match in two independent Markov chains. Curses go to Danny Gianola for planting the seed of Bayesian statistics and then adding plenty of fertilizer and to Bruce Southey, Sandra Rodriguez-Zas, and Dan Sorensen for useful tutoring. Two anonymous reviewers provided very useful comments and criticisms.

LITERATURE CITED

Bianchi, N. O., C. I. Catanesi, G. Bailliet, V. L. Martinez-Marignac, C. M. Bravi *et al.*, 1998 Characterization of ancestral and derived Y-chromosome haplotypes of new world native populations. Am. J. Hum. Genet. **63:** 1862–1871.

Blouin, M. S., M. Parsons, V. Lacaille and S. Lotz, 1996 Use of microsatellite loci to classify individuals by relatedness. Mol. Ecol. **5:** 393–401.

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

Brown, M. D., S. H. Hosseini, A. Torroni, H.-J. Bandlet, J. C. Allen *et al.*, 1998 mtDNA haplogroup X: an ancient link between Europe/Western Asia and North America? Am. J. Hum. Genet. **63:** 1852–1861.

Deka, R., L. Jin, M. D. Shriver, L. Mei Yu, N. Saha *et al.*, 1996 Dispersion of human Y chromosome haplotypes based on five microsatellites in global populations. Genome Res. **6:** 1177–1184.

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

Donnelly, P., and S. Tavaré, 1995 Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. **29:** 401–421.

Donnelly, P., S. Tavaré, D. J. Balding and R. C. Griffiths, 1996 Estimating the age of the common ancestor of men from the *ZFY* intron. Science **272:** 1357–1359.

Edwards, A. H., A. Hammond, L. Jin, C. T. Caskey and R. Chakraborty, 1992 Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics **12:** 241–253.

Forster, P., R. Harding, A. Torroni and H.-J. Bandelt, 1996 Origin and evolution of Native American mtDNA variation: a reappraisal. Am. J. Hum. Genet. **59:** 935–945.

Fu, Y.-X., 1996 Estimating the age of the common ancestor of a DNA sample using the number of segregating sites. Genetics **144:** 829–838.

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

Fu, Y.-X., and W.-H. Li, 1996 Estimating the age of the common ancestor of men from the *ZFY* intron. Science **272:** 1356–1357.

Hammer, M. F., 1995 A recent common ancestry for human Y chromosomes. Nature **378:** 376–378.

Heyer, E., J. Puymirat, P. Dieltjes, E. Bakker and P. De Knijff, 1997 Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum. Mol. Genet. **6:** 799–803.

Hudson, R. R., 1991 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, edited by D. J. Futuyama and J. Antonovics. Oxford University Press, Oxford.

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

Kittles, R. A., M. Perola, L. Peltonen, A. W. Bergen, R. A. Aragon *et al.*, 1998 Dual origins of Finns revealed by Y chromosome haplotype variation. Am. J. Hum. Genet. **62:** 1171–1179.

Lee, P. M., 1997 *Bayesian Statistics: An Introduction*, Ed. 2. Arnold, London.

Li, W.-H., 1976 Electrophoretic identity of proteins in a finite population and genetic distance between taxa. Genet. Res. **28:** 119–127.

Lynch, M., and K. Ritland, 1999 Estimation of pairwise relatedness with molecular markers. Genetics **152:** 1753–1766.

Marshall, T. C., J. Slate, L. E. B. Kruuk and J. M. Pemberton, 1998 Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. **7:** 639–655.

Merriwether, D. A., F. Rothhammer and R. E. Ferrell, 1995 Distribution of the four-founding lineage haplotypes in Native Americans suggests a single wave of migration for the New World. Am. J. Phys. Anthropol. **98:** 411–430.

Nielsen, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154:** 931–942.

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

Olver, F. W. J., 1964 Bessel functions of integer order, pp. 355–434 in *Handbook of Mathematical Functions*, edited by M. Abramowitz and I. A. Stegun. National Bureau of Standards, Washington, DC.

Queller, D. C., and K. F. Goodnight, 1989 Estimating relatedness using genetic markers. Evolution **53:** 258–275.

Ritland, K., 1996 Estimators for pair-wise relatedness and individual inbreeding coefficients. Genet. Res. **67:** 175–185.

Shriver, M. D., L. Jin, R. Chakraborty and E. Boerwinkle, 1993 VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. Genetics **134:** 983–993.

Skorecki, K., S. Selig, S. Blazer, R. Bradman, N. Bradman *et al.*, 1997 Y chromosomes of Jewish priests. Nature **385:** 32.

Stone, A. C., and M. Stoneking, 1998 mtDNA analysis of a prehistoric Oneota population: implications for the peopling of the New World. Am. J. Hum. Genet. **62:** 1153–1170.

Tavaré, S., D. J. Balding, R. C. Griffiths and P. Connelly, 1997 Inferring coalescent times from DNA sequence data. Genetics **145:** 505–518.

Thomas, M. G., T. Parfitt, D. A. Weiss, K. Skorecki, J. F. Watson *et al.*, 2000 Y chromosomes traveling south: the Cohen modal haplotype and the origins of the Lemba—the "Black Jews of Southern Africa." Am. J. Hum. Genet. **66:** 674–686.

Thompson, E. A., 1975 The estimation of pair-wise relationship. Ann. Hum. Genet. **39:** 173–188.

Torroni, A., J. V. Neel, R. Barrantes, T. G. Schurr and D. C. Wallace, 1994 Mitochondrial DNA "clock" for the Amerinds and its implications for timing their entry into North America. Proc. Natl. Acad. Sci. USA **91:** 1158–1162.

Torroni, A., H.-J. Bandelt, L. D'Urbano, P. Lahermo, P. Morla *et al.*, 1998 mtDNA analysis reveals a major late paleolithic population expansion from southwestern to northeastern Europe. Am. J. Hum. Genet. **62:** 1137–1152.

Valdes, A. M., M. Slaktin and N. B. Freimer, 1993 Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics **133:** 737–749.

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

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **10:** 256–276.

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

Wehrhahn, C. F., 1975 The evolution of selectively similar electrophoretically detectable alleles in finite natural populations. Genetics **80:** 375–394.

Wilson, I. J., and D. J. Balding, 1998 Genealogical inference from microsatellite data. Genetics **150:** 499–510.

Communicating editor: Y.-X. Fu

APPENDIX

Suppose a total of S mutations have occurred over the two lineages, with n mutations in lineage 1 and $S - n$ in lineage 2, where n is a random variable. Let X_1 and X_2 denote the changes in array size from the common ancestor. What is the distribution for $d = X_1 - X_2$ (the

difference in array size between the two lineages) conditioned on the total number of mutations S in both lineages? Jay Taylor (Department of Ecology and Evolutionary Biology, University of Arizona) suggested the following approach, which is considerably more elegant than a more brute force method I initially used. Define the random variable Z_i as the change (± 1) in array size generated by the ith mutation. Under the symmetric single-step model, $\Pr(Z_i = +1) = \Pr(Z_i = -1) = \frac{1}{2}$. Further,

$$X_1 = \sum_{i=1}^{n} Z_i \quad \text{and} \quad X_2 = \sum_{i=1}^{S-n} Z_{n+i}.$$

Noting that, in distribution, $Z_i = -Z_n$, it immediately follows that, in distribution,

$$X_1 - X_2 = \sum_{i=1}^{n} Z_i - \sum_{i=1}^{S-n} Z_{n+i} = \sum_{i=1}^{n} Z_i + \sum_{i=1}^{S-n} Z_{n+i} = \sum_{i=1}^{S} Z_i.$$

Hence the probability that the difference in array size is d given a total of S mutations over both lineages is simply the probability that the random walk given by the symmetric single-step model starting from state zero is in state d after S steps.

Using Taylor's result, the probability that two arrays are the same size given that a total of $2M$ mutations have occurred between them equals the probability of M up moves and M down moves, or

$$\Pr(\text{match}|2M \text{ mutations}) = \frac{(2M)!}{(M!)^2}\frac{1}{2^{2M}}.$$