Copyright © 2001 by the Genetics Society of America

The Use of Intraallelic Variability for Testing Neutrality and Estimating Population Growth Rate

Montgomery Slatkin* and Giorgio Bertorelle†

**Department of Integrative Biology, University of California, Berkeley, California 94720-3140 and †Sezione di Biologia Evolutiva, Dipartimento di Biologia, Universitá di Ferrara, 44100 Ferrara, Italy*

Manuscript received July 15, 2000
Accepted for publication February 28, 2001

ABSTRACT

To better understand the forces affecting individual alleles, we introduce a method for finding the joint distribution of the frequency of a neutral allele and the extent of variability at closely linked marker loci (the intraallelic variability). We model three types of intraallelic variability: (a) the number of nonrecombinants at a linked biallelic marker locus, (b) the length of a conserved haplotype, and (c) the number of mutations at a linked marker locus. If the population growth rate is known, the joint distribution provides the basis for a test of neutrality by testing whether the observed level of intraallelic variability is consistent with the observed allele frequency. If the population growth rate is unknown but neutrality can be assumed, the joint distribution provides the likelihood of the growth rate and leads to a maximum-likelihood estimate. We apply the method to data from published data sets for four loci in humans. We conclude that the Δ32 allele at *CCR5* and a disease-associated allele at *MLH1* arose recently and have been subject to strong selection. Alleles at *PAH* appear to be neutral and we estimate the recent growth rate of the European population to be ~0.027 per generation with a support interval of (0.017–0.037). Four of the relatively common alleles at *CFTR* also appear to be neutral but ΔF508 appears to be significantly advantageous to heterozygous carriers.

THE age of an allele determines both its frequency and the extent of variation at closely linked marker loci. Allele age itself cannot be observed, but, given assumptions about past selection and population growth, it constrains the relationship between frequency and intraallelic variability. A large discrepancy between allele frequency and the extent of intraallelic variability expected under neutrality provides evidence of past selection. For example, at several loci in the major histocompatibility (MHC) region in humans and other species and at self-incompatibility loci in several plant species, alleles are found in low frequencies yet exhibit substantial variation among different copies of each allele (Richman *et al.* 1996; Hughes and Yeager 1998). That pattern suggests that balancing selection has retained alleles for much longer times than would be expected if they were neutral. A quite different pattern is found at the *CCR5* locus in humans. A 32-bp deletion (Δ32) retards the onset of AIDS in heterozygous carriers and provides resistance to infection by human immunodeficiency virus in homozygous individuals. This deletion is at a frequency of >10% in Europeans (Stephens *et al.* 1998). Yet very strong linkage disequilibrium with two closely linked microsatellite loci indicates the deletion is young, on the order of 1000 years or less, leading Stephens *et al.* (1998) to conclude that this allele has been subject to strong positive selection.

In this article, we develop a formal theory of the relationship between allele frequency and the extent of intraallelic variability under general assumptions about the demographic history of a population. Our theory can be used in two ways. If the pattern and rate of population growth are assumed to be known, our theory provides a statistical test of neutrality. If the past growth rate is unknown but the allele can be assumed to be neutral, our theory provides a way to estimate the past growth rate.

Our results are obtained by combining two models. The first provides the genealogical history of an allele. The second predicts the extent of intraallelic variability given the intraallelic genealogy. We consider three ways of measuring intraallelic variability: (a) the number of chromosomes carrying the ancestral allele at a linked marker locus; (b) the length of a haplotype shared by all copies of an allele; and (c) the number of mutations at one or more closely linked marker loci.

COALESCENT MODEL OF THE INTRAALLELIC GENEALOGY

We assume that we have a sample of n chromosomes from a randomly mating population of diploid individuals. The history of population size is described by a function $N(t)$, where $t = 0$ is the present and t indicates the number of generations in the past. The current

Corresponding author: Montgomery Slatkin, Department of Integrative Biology, University of California, Berkeley, CA 94720-3140. E-mail: slatkin@socrates.berkeley.edu

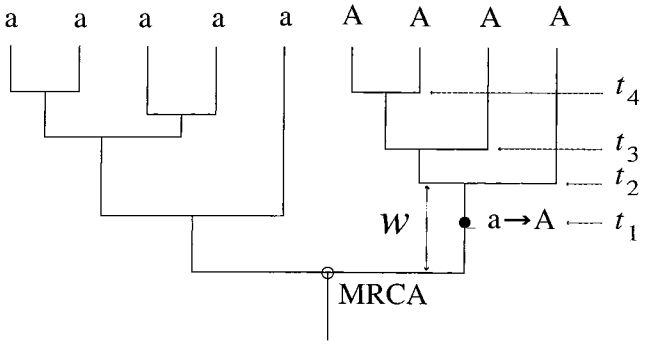

FIGURE 1.—Illustration of a gene genealogy with $n = 9$ tips and an intraallelic genealogy for the mutant allele A found in $i = 4$ copies in the sample. The allele arose at time t_1 in the past. The times t_2, t_3, and t_4 are the intraallelic coalescence times. The time w is the length of the branch joining the intraallelic genealogy to the rest of the gene genealogy and is the weight used to average over replicates, as described in the text (redrawn from SLATKIN and RANNALA 2000). MRCA denotes the most recent common ancestor of the gene genealogy.

population size is $N(0) = N_0$. We assume that i chromosomes in the sample carry an allele A, which is distinguished by a mutation that occurred only once, and hence all i copies of A are identical by descent at the site of the defining mutation. The allele age, t_1, is the time at which the defining mutation occurred. The i A-bearing chromosomes are not necessarily identical at other sites and loci, and the extent of difference among them is the intraallelic variability. How that variability is described and analyzed depends on the data available.

At the site of the mutation defining A, the history of the n chromosomes in the sample can be represented by a gene genealogy that traces ultimately to a single common ancestor. The statistical properties of the gene genealogy are well known (TAVARÉ 1984). For A to have arisen by mutation t_1 generations ago and have exactly i descendants now, it must have occurred on an internal branch that creates a subtree with i terminal branches, as shown in Figure 1. That subtree is the intraallelic genealogy, which necessarily has $i - 1$ internal nodes; we denote the times of those nodes by $t_2, \ldots, t_i$ and call them the intraallelic coalescence times. The time t_k is the time at which the number of lineages in the intraallelic genealogy increases from $k - 1$ to k ($k = 2, \ldots, i$), and it is convenient to define t_{i+1} to be 0, the present. In some coalescent models t_j is used to denote the coalescence times of the entire gene genealogy. In this article, we do not refer to coalescence times other than of the intraallelic genealogy so there is no danger of confusion with other notation.

The extent of intraallelic variability depends on the intraallelic coalescence times. Those times are random variables whose joint distribution is determined by i, n, and $N(t)$. The coalescent model provides a way to generate random sets of intraallelic coalescence times from the correct distribution. By simulating a large number of replicate sets of intraallelic coalescence times and taking appropriate averages over replicates, we can approximate the distribution of intraallelic variability, given i, n, and $N(t)$.

The intraallelic coalescence times are generated as follows. A neutral gene genealogy is simulated using a method similar to that described by HUDSON (1990), but allowing for changes in population size. The procedure is to change the time from t to $\tau(t) = \int_0^t 1/(2N(t'))\,dt'$ and simulate the neutral coalescent model for a constant population size using τ as the independent variable (GRIFFITHS and TAVARÉ 1994). On each simulated gene genealogy, each node is tested to determine whether it has i descendants. If it does, then the set of intraallelic coalescence times is recorded and the length, w, of the branch connecting that subtree to the rest of the gene genealogy is also recorded (see Figure 1). Every node in the gene genealogy has to be tested and, for small values of i, several subtrees with i terminal branches are typically found in each gene genealogy.

For each subtree with i terminal branches, the probability of the observed extent of intraallelic variability is found by combining the intraallelic coalescence times with one of the models described in the next sections. The overall probability is obtained by averaging over a large number of replicates. The average must be weighted by w, the length of the branch connecting that subtree to the rest of the gene genealogy. The reason for using this weighting is that the probability that this subtree represents the genealogical history of A is proportional to the probability that the defining mutation occurred on the branch connecting the subtree to the rest of the gene genealogy, and that probability is proportional to w.

We proceed by describing three different models that generate intraallelic variability. Each model calculates the probability of the data given the set of intraallelic coalescence times, $t_2, \ldots, t_i$.

LINKED BIALLELIC MARKER LOCUS

Mutation and recombination: We assume that a marker locus is linked to A and that the recombination rate between them is c. There are two alleles at the marker, M and m, subject to reversible mutation with rate μ from M to m and ν from m to M. Initially M is on the chromosome carrying A at t_2 so there is perfect linkage disequilibrium between M and A; all A-bearing chromosomes are MA. Subsequently, mutation and recombination create the mA chromosomes. The extent of intraallelic variability is described by the number of MA chromosomes, j. The mathematical problem is to find the probability distribution of j, given the intraallelic coalescence times and the mutation and recombination rates.

We first consider a single lineage of the intraallelic

genealogy. In one generation, the probability that an MA chromosome becomes mA because of mutation and recombination is $u = (1 - q)c + \mu$ and the probability an mA chromosome becomes MA is $v = qc + \nu$, where q is the frequency of M in the population. Because we are concerned with alleles that arose in the recent past, we assume that q remains constant. Slight random variation in q has a negligible effect on the results. Our analysis would not be appropriate if there were substantial systematic changes in q and that problem would require new theoretical analysis.

Still following a single lineage, the probability that an MA chromosome becomes mA after t generations is

$$p_{21} = \frac{u}{u + v}(1 - e^{-(u+v)t}) \tag{1a}$$

and the probability that an mA chromosome becomes MA after t generations is

$$p_{12} = \frac{v}{u + v}(1 - e^{-(u+v)t}). \tag{1b}$$

These equations are derived by applying the standard theory of finite-state continuous-time Markov chains (Karlin and Taylor 1975).

To find the probability distribution of the number of lineages carrying MA in a sample of i chromosomes, we have to account for both the coalescent events and the independent changes on each lineage. Between t_k and t_{k+1} there are k lineages ($t_{i+1} = 0$, the present). When there are k lineages, we can represent the configuration by a vector $p_j^{(k)}$, with $k + 1$ elements representing the probabilities that there are $j = 0, 1, \ldots, k$ MA lineages. If we know these probabilities at $t = t_k$, we can compute them immediately before t_{k+1} by multiplying by a $k + 1$ by $k + 1$ transition matrix denoted by $\mathbf{T}^{(k)}$ whose entries we find from (1) and the assumption that events on different lineages are independent. If there are j MA chromosomes and $k - j$ mA chromosomes, then the number of MA to mA transitions is binomially distributed with probability p_{21} and sample size j, and the number of mA to MA transitions is binomially distributed with probability p_{12} and sample size $k - j$. The elements of $\mathbf{T}^{(k)}$ are found by taking the appropriate convolutions of these binomial distributions. The jj' element of $\mathbf{T}^{(k)}$ is found as follows. Let j_- be the number of MA to mA transitions (*i.e.*, the decrease in j) and let j_+ be the number of mA to MA transitions (*i.e.*, the increase in j). The net change is $j_+ - j_-$ and hence $j' = j + j_+ - j_-$. It follows that

$$T_{jj'}^{(k)} = \sum_{j_-=0}^{\min(j,k-j)} B(j_- + j' - j \mid k - j, p_{12}) B(j_- \mid j, p_{21}),$$

where $B(i|n, p)$ is the binomial probability of i given sample size n and parameter p.

At t_{k+1}, one of the k lineages is chosen randomly to give rise to two descendant lineages. The effect of this event can be modeled by multiplying $P_j^{(k)}$ by a $k + 2$ by $k + 1$ matrix, $\mathbf{S}^{(k)}$, whose jlth element is $1 - j/k$ for $l = j$, j/k, for $l = j + 1$, and 0 otherwise. Multiplying $p_j^{(k)}$ by $\mathbf{S}^{(k)}$ produces $p_j^{(k+1)}$, which has $k + 2$ elements representing the probabilities of the $k + 2$ possible configurations, $j = 0, \ldots, k + 1$.

Given the assumption that initially A was on a chromosome carrying M, $p^{(2)} = (0, 0, 1)$ immediately after t_2. Multiplying successively by $\mathbf{T}^{(k)}$ and $\mathbf{S}^{(k)}$ provides the probabilities of later configurations. In particular, the vector of probabilities that there are j MA chromosomes among the i A-bearing chromosomes in the sample today is

$$p^{(i)} = \mathbf{T}^{(i)} \prod_{k=2}^{i-1} \mathbf{S}^{(k)}\mathbf{T}^{(k)} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \tag{2}$$

The probabilities for each set of intraallelic coalescence times are then averaged using the weighting described in the previous section to obtain an overall probability of j MA chromosomes, which we denote by p_j.

Test for neutrality and estimation of population growth rate: Using the above method we can compute p_j for a given $N(t)$ and, from that distribution, determine whether the observed number of MA chromosomes is consistent with the assumption of neutrality. If j_o is the observed number of MA chromosomes, then the probability P that j is at least as large as j_o under neutrality is found by summing over the tail of the distribution,

$$P = \sum_{j=j_o}^{i} p_j. \tag{3}$$

We can reject neutrality if P is less than a specified level, 0.05 for example.

An alternative is to assume that $N(t)$ is not precisely known and estimate a parameter of it. For example, we could assume that the population has been growing exponentially at rate r in the past to the current size N_0, $N(t) = N_0 e^{-rt}$, and treat r as a parameter to be estimated. If we assume that A is neutral, then p_{j_o} regarded as a function of r is the likelihood. A second demographic model we consider is one in which the rate of exponential growth changes at time t^* in the past. We describe this model as the double exponential model and define the growth between the present and t^* generations in the past to be r_1 and the growth rate before t^* to be r_2. For human populations, a reasonable assumption is $r_1 > r_2$.

Given a single data set, it is impossible to estimate more than one parameter of a model of population growth and hence impossible to know whether the functional form of $N(t)$ assumed is appropriate. But different alleles and different loci all experience the same population growth, so data from different loci can be combined to gain additional information.

LENGTH OF A CONSERVED ANCESTRAL HAPLOTYPE

Model of a continuous chromosome: The general theory of mutation and recombination at several linked marker loci is difficult and not completely developed. At present, a method for efficiently computing configuration probabilities when there are more than two linked markers does not exist, although various approximations have been suggested (Xiong and Guo 1997; Slatkin and Rannala 2000). An alternative is to ignore individual loci and instead model the locations at which recombination has occurred on A-bearing chromosomes. McPeek and Strahs (1999) have developed a theory of this type and applied it to disequilibrium mapping. This approach is appropriate for cases in which all copies of A are on chromosomes carrying the same haplotype at several linked marker loci. This conserved haplotype is assumed to have been on the chromosome on which A arose by mutation and persisted on descendant A-bearing chromosomes because no recombination occurred within the conserved region. The data are then the size l_o, measured in base pairs or in map distance, of the conserved haplotype.

We begin by modeling events on one side of A. At a given position between two bases on the chromosome, the probability that there is a recombination event anywhere on the intraallelic genealogy is the sum of the lengths of the branches (*i.e.*, the tree length, T) multiplied by recombination rate per base pair, denoted by ρ. The probability that there was no recombination between adjacent bases is $1 - \rho T$. The probability that there was no recombination at a distance l bases from A is $(1 - \rho T)^l$ or approximately $e^{-\rho Tl}$. If we let l_1 indicate the distance on one side of A, the cumulative distribution is exponential

$$\Pr(l_1 \leq L) = e^{-\rho TL}, \tag{4}$$

and, hence, the probability distribution is also exponential, as is seen by differentiating (4) with respect to L. The tree length is a simple function of the intraallelic coalescence times:

$$T = t_2 + \sum_{k=2}^{i} t_k. \tag{5}$$

Equation 4 is the cumulative probability that there was no recombination within a distance L on one side of the allele. The total length of an unrecombined segment containing A is the sum of the lengths on the two sides, $l = l_1 + l_2$. The two lengths can be treated as independent random variables provided that the size of the conserved haplotype is sufficiently small that interference in recombination can be ignored. In a small region of a chromosome, the probability of two or more recombination events in a single generation is so low that interference has little effect. The probability distribution of l is then the convolution of two exponential distributions,

$$\Pr(l)\,dl = (\rho T)^2 l e^{-\rho Tl} dl, \tag{6}$$

and hence the cumulative probability distribution is

$$\Pr(l \leq L) = 1 - (1 + \rho TL) e^{-\rho TL}. \tag{7}$$

Test for neutrality and estimation of population growth rate: As in the case of a biallelic linked marker, (6) and (7) provide a basis for either a test of neutrality or a method for estimating a parameter of the population growth model. The observed length of the shared ancestral haplotype containing A is l_o, and that is a minimum estimate because it is based on the detection of a shared haplotype at polymorphic marker loci near A. To test neutrality under the assumption that $N(t)$ is known, we use (5) and (7) to compute $\Pr(l \geq l_o)$ for each set of intraallelic coalescence times and then take the weighted average to obtain an approximation to $\Pr(l \geq l_o)$. If that probability is less than the specified significance level, we reject the hypothesis that the allele is neutral. If instead we assume a parameter of $N(t)$ is unknown, we find the likelihood of that parameter as a function of the data. To do so, we approximate $\Pr(l_o)$ by computing the weighted average of (6) over a large number of replicates.

INFINITE SITES MODEL

A mutation model that is commonly used is the infinite sites model in which a locus is represented by a very large number of completely linked sites. The mutation rate is assumed to be sufficiently small and the number of sites sufficiently large that each site can mutate at most once. Although these assumptions are not valid for most of the nuclear genome, they are convenient for some purposes, particularly when the mutation process at the marker locus or loci is poorly understood. It may be better to assume the infinite sites model and infer the minimum number of mutations that have occurred than to introduce additional uncertainty by assuming an incorrect mutation model. In our applications to *PAH* and *CFTR*, the linked markers are microsatellite loci and the data indicate the minimum number of mutations that have occurred at those loci.

Test for neutrality and estimation of population growth rate: Under the infinite sites model, intraallelic variability is described by the number S of mutations at a locus linked to A on the i chromosomes sampled. Given a set of intraallelic coalescence times, (5) provides the intraallelic tree length, T. The distribution of S is Poisson with parameter μT, where μ is the mutation rate at the marker locus. Taking a weighted average of this Poisson distribution gives us $\Pr(S)$. If S_o is the observed number of segregating sites, then a test of neutrality is obtained by summing the probabilities in the tail of the distribution:

$$P = \Pr(S \leq S_o) = \sum_{S=0}^{S_o} \Pr(S). \tag{8}$$

Either a one- or two-tailed test might be appropriate, depending on prior knowledge. If the population growth rate is unknown, $\Pr(S_o)$ is the likelihood of the growth rate.

COMBINING INFORMATION ACROSS ALLELES AT A LOCUS

Likelihood-ratio test of homogeneity: When data from several alleles at a locus can be analyzed, intraallelic variability is usually assessed at the same linked marker locus or loci. Such a data set provides an opportunity to test whether one of the alleles differs significantly from the others. Because our method provides the likelihood of population growth rate for each allele separately, one test for homogeneity is a likelihood-ratio test, which is done as follows. For the mth allele at a locus, let $L_m(r)$ be the likelihood of r, given the parameters of the model generating intraallelic variability, *i.e.*, the mutation and/or recombination rates. Let $\hat{r}_m$ be the maximum-likelihood estimate (MLE) of r for allele m. Assuming that the alleles are independent, the joint likelihood of r is

$$L_J(r) = \prod_m L_m(r), \tag{9}$$

where the product is taken over all alleles. Let $\hat{r}_J$ be the MLE of r based on the joint likelihood. Finally, define $L_{-m}(r)$ to be $L_J(r)/L_m(r)$, which is the joint likelihood of r based on all but the mth allele, and $\hat{r}_{-m}$ to be the MLE of r based on $L_{-m}(r)$.

In a test for homogeneity, the null hypothesis is that all alleles are equivalent, meaning that the same r applies to each. The likelihood of the data under the null hypothesis is $L_0 = L_J(\hat{r}_J)$. For allele m, we can consider the alternative hypothesis that a different value of r is needed for it and a second value of r can be used for the rest. The likelihood of the data under the alternative hypothesis is $L^* = L_m(\hat{r}_m)L_{-m}(\hat{r}_{-m})$. Under the null hypothesis, $\chi^2 = -2\ln(L_0/L^*)$ is asymptotically distributed as a chi-square deviate with 1 d.f., although the asymptotic theory may not be appropriate for small sample sizes.

Estimate of *r* based on the joint likelihood: If the test of homogeneity does not indicate significant differences among alleles, the joint likelihood provides an MLE of r, $\hat{r}_J$, and that will be a better estimate than is provided by each allele separately. That estimate still depends on the mutation and/or recombination rates and on the assumption of neutrality. Comparisons of estimates across loci can provide additional information.

APPLICATIONS

***CCR5*-Δ*32*:** As discussed in the Introduction, the frequency of the Δ32 allele at *CCR5* exceeds 10% in European populations, yet it appears to be relatively young. Stephens *et al.* (1998) analyzed two microsatellite markers closely linked to *CCR5*, one denoted GAAT and the other AFMB. Stephens *et al.* surveyed 46 chromosomes carrying Δ32 and found that 44 carried the 197 allele at GAAT and 41 carried the 215 allele at AFMB.

We analyze the two markers separately to illustrate the use of our method. For GAAT, $i = 46$, $j_o = 44$, and $q = 0.685$ (the frequency of 197 on chromosomes not carrying Δ32). Stephens *et al.* (1998) estimated the recombination rate between *CCR5* and GAAT to be 0.0021 using a radiation hybrid map. They assumed a mutation rate of $\mu = 0.001$ away from 197. Combining these numbers, $u = c(1 - q) + \mu = 0.0021(1 - 0.685) + 0.001 = 0.00166$, the net rate of loss of 197 alleles on Δ32-bearing chromosomes. The rate of gain is $v = 0.0021 \times 0.685 = 0.00144$, ignoring any gain by mutation. Using these as the two parameters of the model of a biallelic linked marker locus and assuming $r = 0.002$ and $N_0 = 2 \times 10^8$, we find the tail probability, $P = \Pr(j \geq j_o) = 2.2 \times 10^{-12}$. Our analysis strongly supports the conclusion of Stephens *et al.* (1998) that Δ32 is not neutral.

The parameter values needed to carry out this analysis are not known precisely, so it is important to test the sensitivity of the conclusions to variation in their values. In this case, we can reject neutrality for a wide range of parameter values. The P value is most sensitive to changes in u and v. Even if they are reduced by a factor of 10 ($u = 0.000166$ and $v = 0.000144$), P is still only 5.3×10^{-4}. Changes in the model of growth and the rate of recent growth have relatively little effect. For example, if the double exponential model is used with $r_1 = 0.05$, $r_2 = 0.001$, and $t^* = 15$ generations, $P = 1.7 \times 10^{-12}$ with $u = 0.00166$ and $v = 0.00144$.

The data for the other marker locus, AFMB, also allow us to reject neutrality. In this case, $i = 46$, $j_o = 41$, $q = 0.521$, and Stephens *et al.*'s (1998) estimate of c is 0.0093 ($u = 0.00555$ and $v = 0.00485$, assuming mutation at rate 0.001 away from allele 215). With exponential growth and $r = 0.002$, $P = 3.2 \times 10^{-9}$. Reducing u and v each by a factor of 10 increases P to only 8.9×10^{-7}. Assuming double exponential growth with $r_1 = 0.05$, $r_2 = 0.001$, and $t^* = 15$ generations, $P = 3 \times 10^{-9}$.

We conclude, then, that data from the two marker loci closely linked to *CCR5* provide strong evidence that the Δ32 allele has not been neutral in European populations. Slatkin (2001) describes a way to estimate the selection intensity from these data and estimates the selection coefficient in favor of Δ32 to be at least 0.2, comparable with the value estimated by Stephens *et al.* (1998).

***MLH1*:** *MLH1* is one of several DNA mismatch repair genes associated with hereditary nonpolyposis colorectal cancer (HNPCC). Moisio *et al.* (1996) found relatively long conserved haplotypes associated with two recurrent alleles at *MLH1*, denoted *MLH1*1* and *MLH1*2*, in HNPCC families in Finland and concluded that both

were relatively young, 16–43 generations for *MLH1*1* and 5–21 generations for *MLH1*2*. Here, we consider *MLH1*1* in more detail and show that it has probably been selected.

Moisio *et al.* (1996) found almost the same haplotype at four microsatellite markers (loci designated 1612, 1611, 1298, and 3527 in their Figure 1) surrounding *MLH1*1* on 19 chromosomes from unrelated families. Two of these 19 chromosomes differed at one marker locus each, almost certainly as a result of mutation and not recombination. Moisio *et al.* estimated the total recombination distance between these markers to be 7.1 cM. If we assume 1 cM is equivalent to 10^6 bases, the length of the conserved haplotype is $l_o = 7.1 \times 10^6$ bases, and the recombination rate per base, ρ, is 10^{-8}. The actual relationship between centimorgans and base pairs does not affect the results of our analysis, because in our model only the product, ρl_o, which is the total map length spanned by the conserved haplotype (0.071 in this case), enters the calculations. In our notation, i, the number of independent chromosomes sampled, is 19.

The population of Finland has been relatively isolated and has grown rapidly in the recent past (Peltonen *et al.* 1995). It was founded roughly 2000 years ago by a small group and then grew to its present size of ~6 million. Using the result that the effective size of human populations is about one-third of the census size (Felsenstein 1971), we assume a current effective population size of $N_0 = 2 \times 10^6$ and an effective size of the founding population of 1000. If there had been continuous exponential growth for 100 generations, $r = 0.099$. A more accurate model is one that allows for more rapid growth in the recent past. S. K. Service and N. B. Freimer (unpublished data) estimate that the effective population size was ~300,000 in 1700. That would be consistent with the double exponential model with $r_2 = 0.067$ before 1700 ($t^* = 15$ generations) and $r_1 = 0.19$ in the past 300 years.

The frequency of *MHL1*1* is difficult to measure because it is so rare. Aaltonen *et al.* (1998) estimated the frequency to be ~0.0004 based on a screen of individuals with HNPCC. P. Peltomaki (personal communication) found, however, no copies of *MLH1*1* in 2351 healthy anonymous blood donors from the "high risk" region described by Moisio *et al.* (1996), implying that the frequency is <0.0004. In our analysis, we assume $p = 0.0001$, which leads to a conservative test. A higher frequency makes it more likely to reject neutrality.

With these parameter values, our test strongly rejects neutrality of *MLH1*1* for both models of population growth. The tail probabilities under neutrality are $P = 4.5 \times 10^{-7}$ for exponential growth and $P = 1.6 \times 10^{-5}$ for double exponential growth. Even if the estimated length of the conserved haplotype is smaller by a factor of two ($l_o = 3550$ kb), neutrality is still rejected: $P = 0.0048$ for the double exponential model with $r_1 = 0.19$, $r_2 = 0.067$, and $t^* = 15$. And even allowing for more rapid recent growth, $r_1 = 0.3$ (implying a smaller founding population) and assuming $l_o = 3550$ kb, P is still only 0.023.

Another way to explore the robustness of the result is to ask how small l_o has to be to not reject neutrality of *MLH1*1*. We found $P = 0.066$ if $l_o = 2000$ kb, $r_1 = 0.19$, $r_2 = 0.067$, and $t^* = 15$. We can conclude that if estimated size of the conserved haplotype, 7.1 cM, is accurate to within a factor of 2, then *MLH*1* has been selected in the recent past.

***PAH*:** Defective alleles at the locus *PAH* are responsible for phenylketonuria (PKU; Bickel *et al.* 1981). More than 200 causative alleles are known, and many recurrent alleles have been tested for the extent of intraallelic variability at an intronic microsatellite locus. There is no consensus about the effect of alleles associated with PKU on heterozygous carriers. Lewis (1997, p. 248) says that heterozygous women have a lower rate of spontaneous abortion, possibly because of increased resistance to a fungal toxin, but Vogel and Motulsky (1996, p. 289) say that there may be an increased risk of stillbirth and spontaneous abortion.

We analyzed data for 11 recurrent alleles at *PAH*, listed in Table 1. For each allele, we used only those chromosomes with the same combination of restriction fragment length polymorphism (RFLP) and variable number of tandem repeats (VNTR) haplotypes to ensure a single origin. Intraallelic variability was measured at an intronic tetranucleotide microsatellite. To avoid assuming a particular model of mutation at the microsatellite locus, we used the infinite sites model. For each allele, the value of S_o was one less than the number of alleles at the microsatellite locus. The frequency of each allele was estimated from the frequency among PKU chromosomes under the assumption that the overall frequency of PKU chromosomes is 0.01 (Bickel *et al.* 1981). When available, samples from the same allele in different data sets were pooled. The data were assembled from several published surveys and from the *PAH* web site, http://www.mcgill.ca/pahdb. Two different mutation rates for the microsatellite marker were assumed: 0.0021, based on the study by Weber and Wong (1993), and 0.00028, based on the study by Chakraborty *et al.* (1997). When a single growth rate, $r = 0.008$, was assumed, we rejected neutrality for 9 of 11 alleles under the higher mutation rate but for only 1 of 11 alleles at the lower mutation rate (Table 1). Similar results were obtained for the double exponential model with $r_1 = 0.04$, $r_2 = 0.002$, and $t^* = 150$.

We proceeded on the assumption that the lower mutation rate, $\mu = 0.00028$, is more appropriate, assuming in effect that these alleles are neutral or nearly so. We calculated the likelihood of r_1 in the double exponential model (with $r_2 = 0.002$ and $t^* = 150$) for each of the 11 alleles. Figure 2 shows four of the likelihood curves obtained. The rest are similar. We tested the set of 11

TABLE 1

Parameters and results for the *PAH* data sets

Mutation/ RFLP.VNTR background	p ($\times 10^{-4}$)	i	S_o	n ($\times 10^3$)	N_0 ($\times 10^7$)	$\mu = 0.0021$ (single)	$\mu = 0.00028$ (single)	$\mu = 0.0021$ (double)	$\mu = 0.00028$ (double)
F39L/1.8	3.4	27	2	78.8	2.83	1.3×10^{-5}	0.578	0.075	0.955
L48S/4.3	1.5	12	3	82.4	6.00	0.025	0.960	0.648	0.999
I65T/9.8	8.8	75	4	85.2	6.10	5.2×10^{-20}	0.081	3.2×10^{-5}	0.868
R158Q/4.3	4.7	37	3	78.3	2.97	2.4×10^{-7}	0.569	0.041	0.974
G272X/7.8	1.1	6	2	54.1	2.87	0.293	0.983	0.858	0.999
A300S/1.8	0.6	5	1	78.2	4.17	0.221	0.934	0.723	0.991
L348V/9.8	3.8	24	2	63.8	4.10	5.4×10^{-6}	0.542	0.068	0.952
IVS10/6.7	5.5	146	7	265.3	7.50	2.4×10^{-28}	0.074	5.5×10^{-7}	0.942
A403V/1.8	2.4	13	2	54.0	4.07	0.004	0.838	0.370	0.990
R408W/1.8	10.0	34	3	32.6	10.0	6.4×10^{-13}	0.222	0.001	0.890
R408W/2.3	57.7	32	2	5.5	10.0	4.5×10^{-19}	0.032	1.8×10^{-7}	0.369

p, allele frequency; i, number of analyzed chromosomes; S_o, minimum number of mutations in the intraallelic genealogy; n, estimated sample size; N_0, effective population size. The four right-hand columns show the tail probabilities, $P = \Pr(S \geq S_o)$, assuming two mutation rates and assuming the single exponential model with $r = 0.008$ or the double exponential model with $r_1 = 0.04$, $r_2 = 0.002$, and $T = 150$. Approximate population sizes refer to the countries where the data originated and where the mutations are most frequently observed. Given the wide distribution of some alleles, larger population sizes could sometimes be more appropriate.

alleles for homogeneity using the likelihood-ratio test described above. On the 11 tests, 3 resulted in values of χ^2 that would indicate significant heterogeneity at the 5% level assuming a χ^2_1 distribution under the null hypothesis: $\chi^2 = 4.2$ for G272X/7.8, 4.4 for L48S/4.3, and 7.8 for R408W/2.3. The other 8 values were all between 0 and 1.4. Whether these values of χ^2 indicate significant heterogeneity is difficult to assess. There is no theory to tell us whether χ^2 will be distributed as χ^2_1 under the null hypothesis for these sample sizes. Our results from the neutrality tests with $\mu = 0.00028$ suggest that all 11 alleles may be neutral and that the values of χ^2 we find are consistent with no heterogeneity. Because these alleles were sampled in different European countries, some of the variation in χ^2 may reflect differences in the past rate of population growth in different parts of Europe, although other genetic studies of Europeans do not show extensive substructuring that would reflect significant isolation of major groups (Cavalli-Sforza *et al.* 1994).

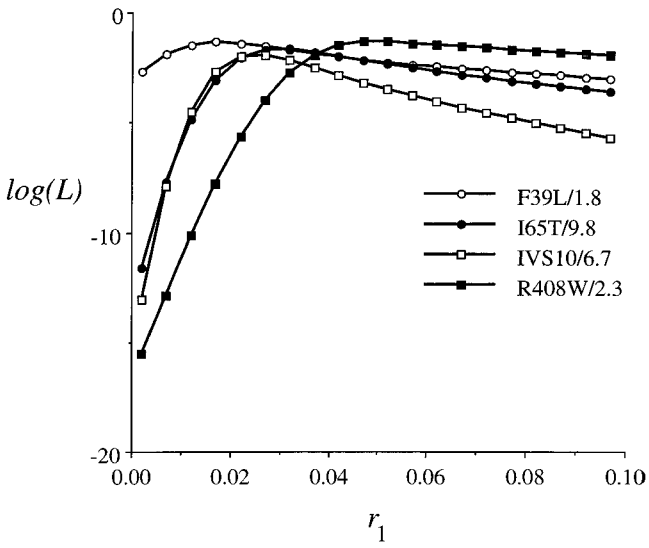

Figure 2.—Log likelihood of r_1 for four alleles at *PAH*, assuming the double exponential model with $t^* = 150$ generations and $r_2 = 0.002$. The values of p, i, S_o, and N_0 are given for each mutation in Table 1. In all cases $\mu = 2.8 \times 10^{-2}$ was assumed. The likelihood curves for the seven other alleles are similar.

If we combine information across alleles, then we get a good estimate of growth rate under the assumption that they are neutral. For $\mu = 0.00028$, and assuming the double exponential model with $r_2 = 0.002$ and $t^* = 150$ generations, the joint likelihood function leads to an MLE of r_1 of $\hat{r}_1 = 0.027$ with a support interval of (0.017–0.037). Figure 3 shows the joint likelihood for two mutation rates at the microsatellite locus. If $\mu = 0.0021$ and the alleles are neutral, then the average recent growth rate, r_1, would have to be $\sim$0.15 per generation, which is probably too high.

Because we do not know the appropriate mutation rate at the linked marker and because our model of past population growth is necessarily simplified, we cannot draw any strong conclusions about alleles at *PAH* or about recent growth of the European populations from which these data were obtained. We present these results to illustrate the use of our method. If the mutation rate at the linked microsatellite is close to 0.00028, then these alleles are neutral or nearly so. When data are combined across alleles, the resulting estimate of the recent growth rate of European populations is not

TABLE 2

Parameters and results for the *CFTR* data sets

Mutation	p ($\times 10^{-4}$)	i	S_o	n ($\times 10^3$)	N_0 ($\times 10^7$)	$\mu = 0.0001$ (single)	$\mu = 0.0005$ (single)	$\mu = 0.001$ (double)	$\mu = 0.0005$ (double)
ΔF508	132	2112	59	160.0	20.0	0.0	0.0	6.0×10^{-73}	6.3×10^{-21}
ΔF508	132	2112	118	160.0	20.0	0.0	5.5×10^{-91}	4.4×10^{-39}	1.0×10^{-4}
G542X	6.8	116	9	170.6	5.82	8.9×10^{-8}	0.012	0.231	0.907
N1303K	5.6	59	7	105.3	5.18	5.9×10^{-4}	0.180	0.639	0.977
1717-1G-A	3.2	24	3	75.0	4.38	0.031	0.370	0.695	0.945
R1162X	2.2	68	4	309.1	5.18	3.4×10^{-4}	0.082	0.313	0.829

Definitions are as in Table 1, except the mutation rates. The ΔF508 mutation was also tested assuming for S_o a value twice as large as the minimum value inferred.

obviously wrong. If the mutation rate at the microsatellite locus is 0.0021, then most or all of these alleles are not neutral and the resulting estimate of r_1 based on the joint likelihood is not accurate because its value reflects both past population growth and past selection. Data from additional markers linked to *PAH* and data from other loci will be necessary to determine which of these conclusions is more nearly correct.

***CFTR*:** Alleles at *CFTR* are responsible for cystic fibrosis (CF). More than 800 causative recessive alleles are known. Until recently CF caused death in early childhood. CF is relatively common in Europeans, with a frequency of 1 in 2500 births implying a frequency of heterozygous carriers of roughly 1/25. Approximately two-thirds of the causative alleles in northwestern Europe are ΔF508, which is a 3-bp deletion resulting in the loss of the 508th amino acid in the CFTR protein (Rommens *et al.* 1989). An important question about alleles at *CFTR*, and in particular about ΔF508, is whether their relatively high frequency results from a selective advantage to heterozygous carriers (Morton 1968).

For our analysis we used intraallelic variation at three intronic microsatellite loci that have been studied extensively. Data were accumulated from several published articles and from the online *CFTR* database, http://www.genet.sickkids.on.ca/cftr. On the basis of the observation that no mutations were detected in 3000 meioses, Kaplan *et al.* (1994) inferred that the combined mutation rate at the three loci together did not exceed 0.001. We assumed the infinite sites model for the three markers together and inferred the minimum number of mutations, S_o, consistent with the data. For all but ΔF508, S_o was one less than the number of distinct marker haplotypes associated with each allele. For ΔF508,

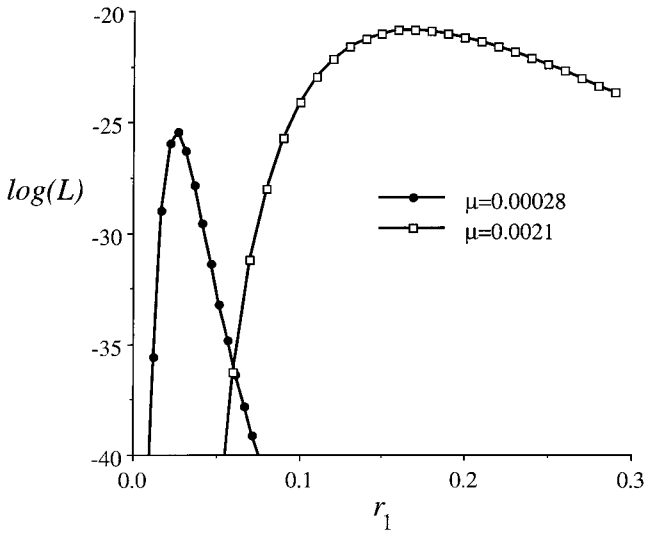

Figure 3.—Log likelihood of r_1 for *PAH* obtained by summing the log-likelihood values from the coalescent simulation over all 11 alleles listed in Table 1, assuming the double exponential model with $t^* = 150$ generations and $r_2 = 0.002$ and assuming two different mutation rates for the infinite sites model.

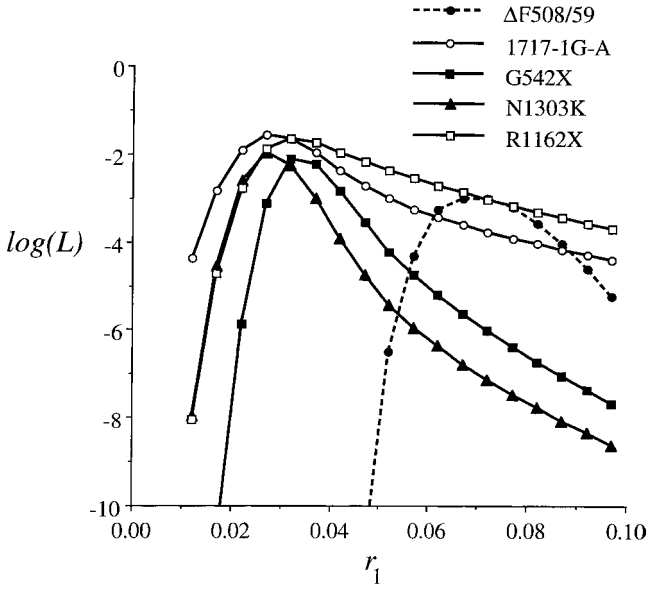

Figure 4.—Log likelihood of r_1 for five alleles at *CFTR*. The results shown were obtained from the coalescent simulation assuming $\mu = 0.00028$ for the infinite sites model; the double exponential model of population growth with $r_2 = 0.002$; $t^* = 150$ generations; and values of i, n, and N_0 given in Table 2.

TABLE 3

Values of $\chi^2 = -2\log(L_0/L^*)$ calculated as described in the text for each allele at *CFTR* compared to the rest

Mutation	$\mu = 0.00028$: $S_o = 59$	$\mu = 0.00028$: $S_o = 118$	$\mu = 0.00084$: $S_o = 59$	$\mu = 0.00084$: $S_o = 118$
ΔF508	20.3	11.0	19.6	4.6
G542X	5.8	1.4	5.4	0.8
N1303K	8.4	3.8	7.9	3.0
1717-1G-A	3.4	0.5	3.5	1.0
R1162X	1.8	0.6	1.6	0

S_o was estimated by using a parsimony algorithm to infer the minimum number of mutations necessary to generate the observed haplotypes. The method was the same as that employed by Morral *et al.* (1994). To allow for the possibility that $S_o = 59$ is too small, we performed the same calculations for $S_o = 118$.

We tested for neutrality of five alleles at *CFTR*. Parameters and results are shown in Table 2. Assuming the lower mutation rate, $\mu = 0.0005$, does not reject neutrality for any of the four less frequent alleles. Even for the higher rate, $\mu = 0.001$, neutrality is rejected only for the model of simple exponential growth. In contrast, neutrality of ΔF508 was rejected in all cases, even with $S_o = 118$.

We computed the likelihood of population growth rate for each allele, as shown in Figure 4. The mutation rate, $\mu = 0.00028$, was used because the joint likelihood for the four alleles other than ΔF508 led to an MLE of r_1 of 0.023, which is roughly the same as that obtained from *PAH*. If these four alleles at *CFTR* and the 11 alleles at *PAH* are neutral, the three markers linked to *CFTR* have a lower per locus mutation rate than the single marker linked to *PAH*. Once again, this conclusion is tentative because of our uncertainty both about the neutrality of the alleles and about the mutation rates of the marker loci.

We used the likelihood-ratio test of homogeneity to determine whether ΔF508 differs significantly from the others. Table 3 shows the values of χ^2 obtained in each of the tests of homogeneity performed for the two mutation rates and two values of S_o. For all four sets of parameter values, the range of χ^2 values for the other four alleles is comparable to that found in the tests of homogeneity for the 11 alleles at *PAH*. And for all four sets of parameter values, χ^2 is substantially larger for ΔF508, which is consistent with it having experienced different selection. That conclusion is consistent with the tests of neutrality summarized in Table 2. Figure 4 shows one set of likelihood curves that illustrates the extent of difference in the likelihood curves. Our conclusion agrees with that of Wiuf (2001), who used a birth-death model to approximate the dynamics of ΔF508 and concluded that it was positively selected in the past few thousand years.

As in the case of *PAH*, our conclusions are tentative because they depend on assumed values of parameters and on a simplified model of past population growth. The results for *CFTR* are compatible with those for *PAH* if a lower per locus mutation is assumed for the microsatellite loci linked to *CFTR* and if only ΔF508 has been subject to significant selection. Additional data will be needed before any firmer conclusions can be drawn.

DISCUSSION AND CONCLUSIONS

We have shown that intraallelic variability combined with allele frequency provides information about the history of selection and population growth experienced by an allele. Our methods, like any methods of historical inference in population genetics, rely on assumptions that cannot be literally true. Populations are not randomly mating and have not undergone continuous exponential growth for long periods of time. The role of theory in this case is to find what models are sufficient to account for observations and to provide tentative conclusions that have to be subject to further testing when new data become available.

One assumption of our analysis is random mating, but that is not an important limitation to the analysis of low-frequency alleles. The dynamics of alleles in low frequency are well approximated by a birth-death process, which assumes that each copy of an allele reproduces independently of all others (Wiuf 2000) and hence independently of which subpopulation it and other copies are in. Therefore, our results are applicable to samples pooled from different geographic regions.

Our analysis shows that *CCR5-Δ32* and *MLH1*1* have probably been selected in the recent past. That conclusion is not sensitive to variation in assumed parameter values. Our result for Δ32 agrees with that of Stephens *et al.* (1998). Our result for *MLH1*1* is similar to that obtained by Slatkin (2000) for two low-frequency alleles at *BRCA1*. Even though these disease-associated alleles at *BRCA1* and *MLH1* are quite rare, the level of intraallelic variability associated with them is too low to be consistent with neutrality.

Our analysis of *PAH* and *CFTR* shows how information can be combined across alleles at a locus and across loci. Differences among alleles at a locus can indicate

the possible effect of selection. If alleles at a locus are neutral, then the joint likelihood provides a good estimate of population growth rate. Neutral alleles at different loci will provide independent estimates of growth rate.

We thank Dr. P. Peltomaki for information about the frequency of mutants of *MLH1* in the Finnish population, and Drs. L. Excoffier and C. Wiuf for helpful comments on an earlier version of this article. This research was supported in part by National Institutes of Health grant GM40282 to M. Slatkin.

LITERATURE CITED

Aaltonen, L. A., R. Salovaara, P. Kristo, F. Canzian, A. Hemminki *et al.*, 1998 Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. N. Engl. J. Med. **338:** 1481–1487.

Bickel, H., C. Bachmann, R. Beckers, N. J. Brandt, B. E. Clayton *et al.*, 1981 Neonatal mass-screening for metabolic disorders—summary of recent sessions of the committee of experts to study inborn metabolic diseases, Public-Health Committee, Council-of-Europe—Review. Eur. J. Pediatr. **137:** 133–139.

Cavalli-Sforza, L. L., P. Menozzi and A. Piazza, 1994 *The History and Geography of Human Genes.* Princeton University Press, Princeton, NJ.

Chakraborty, R., M. Kimmel, D. N. Stivers, L. J. Davison and R. Deka, 1997 Relative mutation rates at di, tri-, and tetranucleotide microsatellite loci. Proc. Natl. Acad. Sci. USA **94:** 1041–1046.

Felsenstein, J., 1971 Inbreeding and variance effective numbers in populations with overlapping generations. Genetics **68:** 581–597.

Griffiths, R. C., and S. Tavaré, 1994 Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. Lond. B **344:** 403–410.

Hudson, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–44.

Hughes, A. L., and M. Yeager, 1998 Natural selection at major histocompatibility complex loci of vertebrates. Annu. Rev. Genet. **32:** 415–435.

Kaplan, N. L., P. O. Lewis and B. S. Weir, 1994 Age of the ΔF508 cystic fibrosis mutation. Nat. Genet. **8:** 216.

Karlin, S., and H. M. Taylor, 1975 *A First Course in Stochastic Processes.* Academic Press, New York.

Lewis, R., 1997 *Human Genetics: Concepts and Applications.* Brown, Dubuque, IA.

McPeek, M. S., and A. Strahs, 1999 Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. Am. J. Hum. Genet. **65:** 858–875.

Moisio, A. L., P. Sistonen, J. Weissenbach, A. de la Chapelle and P. Peltomäki, 1996 Age and origin of two common MLH1 mutations predisposing to hereditary colon cancer. Am. J. Hum. Genet. **59:** 1243–1251.

Morral, N., J. Bertranpetit, X. Estivill, V. Nunes, T. Casals *et al.*, 1994 The origin of the major cystic fibrosis mutation (DELTA-F508) in European populations. Nat. Genet. **7:** 169–175.

Morton, N. E., 1968 Genetic studies on cystic fibrosis in Hawaii. Am. J. Hum. Genet. **20:** 157–169.

Peltonen, L., P. Pekkarinen and J. Aaltonen, 1995 Messages from an isolate—lessons from the Finnish gene pool. Biol. Chem. Hoppe-Seyler **376:** 697–704.

Richman, A. D., M. K. Uyenoyama and J. R. Kohn, 1996 Allelic diversity and gene genealogy at the self-incompatibility locus in the Solanaceae. Science **273:** 1212–1216.

Rommens, J. M., M. C. Iannuzzi, B. Kerem, M. L. Drumm, G. Melmer *et al.*, 1989 Identification of the cystic fibrosis gene: chromosome walking and jumping. Science **245:** 1059–1065.

Slatkin, M., 2000 Allele age and a test for selection on rare alleles. Philos. Trans. R. Soc. Lond. B **355:** 1663–1668.

Slatkin, M., 2001 A method for simulating genealogies of selected alleles in a population of variable size. Genet. Res. (in press).

Slatkin, M., and B. Rannala, 2000 Estimating allele age. Annu. Rev. Genom. Hum. Genet. **1:** 225–249.

Stephens, J. C., D. E. Reich, D. B. Goldstein, H. D. Shin, M. W. Smith *et al.*, 1998 Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. Am. J. Hum. Genet. **62:** 1507–1515.

Tavaré, S., 1984 Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. **26:** 119–165.

Vogel, F., and A. G. Motulsky, 1996 *Human Genetics: Problems and Approaches.* Springer-Verlag, New York.

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

Wiuf, C., 2000 On the genealogy of a sample of neutral rare alleles. Theor. Popul. Biol. **58:** 61–75.

Wiuf, C., 2001 Do Δ*F*508 heterozygotes have a selective advantage? Genet. Res. (in press).

Xiong, M., and S. W. Guo, 1997 Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. Am. J. Hum. Genet. **60:** 1513–1531.

Communicating editor: Y.-X. Fu