Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 15452–15457, December 1998
Evolution

Using rare mutations to estimate population divergence times: A maximum likelihood approach

GIORGIO BERTORELLE*† AND BRUCE RANNALA‡

*Department of Integrative Biology, University of California, Berkeley, CA 94720-3140; and ‡Department of Ecology and Evolution, State University of New York, Stony Brook, NY 11794-5245

Communicated by Henry C. Harpending, University of Utah, Salt Lake City, UT, October 23, 1998 (received for review July 27, 1998)

ABSTRACT **In this paper we propose a method to estimate by maximum likelihood the divergence time between two populations, specifically designed for the analysis of nonrecurrent rare mutations. Given the rapidly growing amount of data, rare disease mutations affecting humans seem the most suitable candidates for this method. The estimator *RD*, and its conditional version *RDc*, were derived, assuming that the population dynamics of rare alleles can be described by using a birth–death process approximation and that each mutation arose before the split of a common ancestral population into the two diverging populations. The *RD* estimator seems more suitable for large sample sizes and few alleles, whose age can be approximated, whereas the *RDc* estimator appears preferable when this is not the case. When applied to three cystic fibrosis mutations, the estimator *RD* could not exclude a very recent time of divergence among three Mediterranean populations. On the other hand, the divergence time between these populations and the Danish population was estimated to be, on the average, 4,500 or 15,000 years, assuming or not a selective advantage for cystic fibrosis carriers, respectively. Confidence intervals are large, however, and can probably be reduced only by analyzing more alleles or loci.**

The amount of genetic divergence between two isolated populations tends to accumulate with time, following their subdivision from a common ancestral population. New alleles are independently generated in each descendent population by mutation, and the frequencies of the pre-subdivision alleles tend to diverge due to the random sampling of genes in each generation (genetic drift). These two processes therefore leave a signature, increasingly evident with time, on the genetic composition of the subdivided populations.

Several methods have been proposed to estimate the time, in the past, when two or more populations arose from a single ancestral population (i.e., the divergence time). Takahata and Nei (1) suggested that the net number of nucleotide substitutions d accumulated between two populations (called also d_A in ref. 2) be used to estimate the time of their divergence. If the two populations have the same constant size, d is expected to increase linearly with the product μT (where μ is the mutation rate and T is the divergence time). The same linear increase with μT is predicted for the genetic distance $(\delta\mu)^2$ (3), computed as the square of the difference between the average allele size observed at microsatellite markers in the diverging populations. In both cases, therefore, an estimation of T can be simply obtained if the mutation rate is known.

Another commonly used approach (see, for example, ref. 4) is to first estimate Wright's F_{st} (5) from allele frequencies and then use this estimate to predict the divergence time. The relationship $F_{st} = 1 - e^{-T/2N}$ (5, 6) can be used to estimate the divergence time T, given that the populations have the same constant and known size N. This estimator is based on a model of genetic drift without mutation, and it is therefore most suitable for populations that have recently separated. However, when the differences between alleles are taken into account by using equivalents of F_{st} that allow for mutation [such as ϕ_{st} (7) or R_{st} (8)], equivalent estimators become feasible for older population subdivision events (8).

More recently, Nielsen *et al.* (9) have proposed a maximum likelihood estimator based on the coalescent process and assuming no mutation. When applied to simulated data from stable populations, this estimator appears to have less bias and lower variance than an F_{st}-based estimator (9).

In this paper we present another likelihood estimator of the time of divergence of two populations, specifically designed for the analysis of rare alleles that have arisen by nonrecurrent mutation. Rare alleles are becoming an important source of information on human populations as more disease mutations are mapped, more effort is focused on the study of population frequencies of disease mutations, and large-scale programs of genetic screening are becoming a realistic possibility. The method we present here is best suited for analyzing data on rare disease mutations, since it assumes that the number of copies of each mutant (at the same or different loci) can be modeled by using a stochastic birth–death process with sampling (10, 11). This assumption, which is satisfied if mutants are rare (12), allows many demographic factors, including selection and population growth, to be introduced into the model in a relatively simple way. This approach also greatly simplifies the analysis of multiallelic and multilocus data.

The Model

We consider a simple model of an ancestral population (labeled population 0) that separates into two descendent populations (labeled populations 1 and 2) at a time T generations in the past. The descendent populations experience no immigration. A rare allele is assumed to have arisen by nonrecurrent mutation at time $T + t$ in the past. When the population split occurs, any copy of the allele in the ancestral population has a probability s and $(1 - s)$ of joining one, or the other, descendent population.

Within each population, we assume that the demography of the mutant lineages can be described by using a birth–death process approximation to the coalescent model, valid when the allele has remained rare (12). Population sizes are allowed to change over time according to an exponential process of growth (or decline) with rates ξ_0 (ancestral population), ξ_1, and ξ_2 (descendent populations).

The Likelihood

A mutant allele arises at time $T + t$ in the past. The probability distribution of the total number of alleles descended from the mutant, k, that existed immediately prior to the population subdivision event T generations ago is

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

© 1998 by The National Academy of Sciences 0027-8424/98/9515452-6$2.00/0
PNAS is available online at www.pnas.org.

Abbreviations: pdf, probability distribution function; pgf, probability generating function; CF, cystic fibrosis.

†To whom reprint requests should be addressed. e-mail: giorgio@mws4.biol.berkeley.edu.

$$\Pr(k) = (1 - \nu_t)\nu_t^{k-2} \qquad k \geq 2, \quad [1]$$

where (see ref. 12) ν_t is defined as

$$\nu_t = 1 - \frac{2\xi_0 e^{-\xi_0 t}}{1 - (1 - 2\xi_0)e^{-\xi_0 t}}. \quad [2]$$

Let j_1 be the number of copies of the mutant that enter descendent population 1 at the population subdivision event and let j_2 be the number that enter descendent population 2. The joint probability distribution function (pdf) of j_1 and j_2, conditional on $k = j_1 + j_2$, is

$$\Pr(j_1, j_2|k) = \binom{k}{j_1} s^{j_1}(1 - s)^{j_2}. \quad [3]$$

The joint probability generating function (pgf) of j_1 and j_2, conditional on k, is then

$$\phi_{j_1,j_2|k}(z_1, z_2) = \sum_{j_1=0}^{k} \binom{k}{j_1} s^{j_1}(1 - s)^{j_2} z_1^{j_1} z_2^{j_2}$$

$$= (z_2 - sz_2 + sz_1)^k. \quad [4]$$

The unconditional joint pgf of j_1 and j_2 is

$$\phi_{j_1,j_2}(z_1, z_2) = \sum_{k=2}^{\infty} (z_2 - sz_2 + sz_1)^k (1 - \nu_t)\nu_t^{k-2}$$

$$= \frac{(z_2 - sz_2 + sz_1)^2(1 - \nu_t)}{1 - \nu_t z_2 + s\nu_t z_2 - s\nu_t z_1}. \quad [5]$$

Using standard techniques (see ref. 13), we obtain the probability of j_1 and j_2 by evaluating the j_1th and j_2th-order partial derivatives with respect to z_1 and z_2 as follows

$$\Pr(j_1, j_2) = \frac{1}{j_1! j_2!} \frac{\partial^{j_1+j_2}\phi_{j_1,j_2}(z_1, z_2)}{\partial z_1^{j_1} \partial z_2^{j_2}}\bigg|_{z_1=z_2=0}. \quad [6]$$

By examining successive terms of the pdf obtained from the pgf in this manner, a general formula for the unconditional pdf of j_1 and j_2 may be obtained

$$\Pr(j_1, j_2) = \binom{j_1 + j_2}{j_1} \frac{(1 - \nu_t)}{\nu_t^2} (s\nu_t)^{j_1}((1 - s)\nu_t)^{j_2}$$

$$\text{for all } j_1 + j_2 \geq 2. \quad [7]$$

We will assume that no additional mutations occur at the site of interest (this condition is satisfied when the product of divergence time T and the mutation rate is small), which implies that each descendent population must have contained at least one copy of the mutant allele after the divergence event so that we must condition on $j_1 > 0$ and $j_2 > 0$. It is easy to show that

$$\Pr(j_1 > 0, j_2 > 0) = \sum_{j_1=1}^{\infty} \sum_{j_2=1}^{\infty} \Pr(j_1, j_2)$$

$$= \frac{(1 - s)s(\nu_t - 2)}{(1 + \nu_t(1 - s))(s\nu_t - 1)}, \quad [8]$$

and the conditional pdf is then

$$\Pr(j_1, j_2|j_1 > 0, j_2 > 0) = \binom{j_1 + j_2}{j_1} (s\nu_t)^{j_1}((1 - s)\nu_t)^{j_2}$$

$$\times \frac{(1 + \nu_t(1 - s))(s\nu_t - 1)(1 - \nu_t)}{\nu_t^2(1 - s)s(\nu_t - 2)}. \quad [9]$$

Let l_i be the number of copies of the mutant found in population i immediately after the divergence event at time T that leave one or more descendents in a present-day sample from population i, where i is either 1 or 2. The pdf of l_i is a binomial of the form

$$\Pr(l_i|j_i) = \binom{j_i}{l_i} Q_{Ti}^{l_i}(1 - Q_{Ti})^{j_i - l_i} \qquad 0 \leq l_i \leq j_i, \quad [10]$$

where Q_{Ti} is the probability that an allele in population i leaves one or more descendents at present and is given by (see ref. 12)

$$Q_{Ti} = \frac{2f_i\xi_i}{f_i - (f_i - 2\xi_i)e^{-\xi_i T}}, \quad [11]$$

where the sampling fraction f_i is the probability that a chromosome in the present-day descendents of population i is sampled.

The pgf of l_i, conditional on j_i, is

$$\phi_{l_i|j_i}(z_i) = \sum_{l_i=0}^{j_i} \binom{j_i}{l_i} Q_{Ti}^{l_i}(1 - Q_{Ti})^{j_i - l_i} z_i^{l_i}$$

$$= (1 - Q_{Ti} + Q_{Ti}z_i)^{j_i}. \quad [12]$$

Because the drift processes in descendent populations 1 and 2 after the subdivision event are independent, the joint pgf of l_1 and l_2, conditional on j_1 and j_2, is a product of the pgfs of l_1 and l_2,

$$\phi_{l_1,l_2|j_1,j_2}(z_1, z_2) = (1 - Q_{T1} + Q_{T1}z_1)^{j_1}(1 - Q_{T2} + Q_{T2}z_2)^{j_2}. \quad [13]$$

The unconditional joint pgf of l_1 and l_2 is then

$$\phi_{l_1,l_2}(z_1, z_2) = \sum_{j_1=1}^{\infty} \sum_{j_2=1}^{\infty} \phi_{l_1,l_2|j_1,j_2}(z_1, z_2)\Pr(j_1, j_2|j_1 > 0, j_2 > 0). \quad [14]$$

Using standard methods (see above), the pdf can be obtained from Eq. **14** and is

$$\Pr(l_1, l_2) = \binom{l_1 + l_2}{l_2} \frac{(-1)^{l_1+1} {Q_{T1}}^{l_1} {Q_{T2}}^{l_2} s^{l_1-1}(s - 1)^{l_2-1} \nu_t^{l_1+l_2-2}(\nu_t - 1)((s - 1)s\nu_t^2 + \nu_t - 1)}{(\nu_t - 2)((Q_{T2}(s - 1) - Q_{T1}s + 1)\nu_t - 1)^{l_1+l_2+1}}. \quad [15]$$

We must once more condition on the fact that at least one copy of the mutant allele leaves descendents in each population sample. This conditional probability is

$$\Pr(l_1, l_2|l_1 > 0, l_2 > 0) = \binom{l_1 + l_2}{l_2} \frac{(-1)^{l_1+1} {Q_{T1}}^{l_1-1} {Q_{T2}}^{l_2-1}(s - 1)^{l_2-1} s^{l_1-1}(\nu_t - 1)\nu_t^{l_1+l_2-2}((Q_{T2}(s - 1) + 1)\nu_t - 1)((Q_{T1}s - 1)\nu_t + 1)}{((Q_{T2}(s - 1) - Q_{T1}s + 1)\nu_t - 1)^{l_1+l_2}((Q_{T2} - Q_{T2}s + Q_{T1}s - 2)\nu_t + 2)}. \quad [16]$$

The pdf of the number of copies of the mutant allele in a sample from population i, denoted as n_i, conditional on l_i is (see, e.g., ref. 10)

$$\Pr(n_i|l_i) = \binom{n_i - 1}{n_i - l_i} (1 - \nu_{Ti})^{l_i} \nu_{Ti}^{n_i - l_i}, \quad [17]$$

where we define (see ref. 12)

$$\nu_{Ti} = 1 - \frac{2\xi_i e^{-\xi_i T}}{f_i - (f_i - 2\xi_i)e^{-\xi_i T}}. \quad [18]$$

Since the drift processes in the two populations are independent, the joint distribution of n_1 and n_2, given l_1 and l_2, is

$$\Pr(n_1, n_2|l_1, l_2) = \binom{n_1 - 1}{n_1 - l_1}\binom{n_2 - 1}{n_2 - l_2}(1 - \nu_{T1})^{l_1}\nu_{T1}^{n_1 - l_1} \times (1 - \nu_{T2})^{l_2}\nu_{T2}^{n_2 - l_2}. \quad [19]$$

The unconditional pdf of n_1 and n_2 is then calculated, using Eq. **16** and Eq. **19**, as

$$\Pr(n_1, n_2) = \sum_{l_1=1}^{n_1}\sum_{l_2=1}^{n_2}\Pr(n_1, n_2|l_1, l_2)\Pr(l_1, l_2|l_1 > 0, l_2 > 0). \quad [20]$$

To simplify the evaluation of the sum Eq. **20** we used the iterative relationship

$$\begin{aligned}\Pr(l_1 = i, l_2 = j|l_1 > 0, l_2 > 0) &= \\ \Pr(l_1 = i - 1, l_2 = j|l_1 > 0, l_2 > 0) &\times \left(1 + \frac{j}{i}\right) \\ \times \frac{-Q_{T1}s\nu_t}{((Q_{T2}(s-1) - Q_{T1}s + 1)\nu_t - 1)} &\\ = \Pr(l_1 = i, l_2 = j - 1|l_1 > 0, l_2 > 0) &\times \left(1 + \frac{i}{j}\right) \\ \times \frac{-Q_{T2}(1-s)\nu_t}{((Q_{T2}(s-1) - Q_{T1}s + 1)\nu_t - 1)}&. \quad [21]\end{aligned}$$

Eq. **20** is the basis for the likelihood function of T used in our analysis:

$$L(T|n_1, n_2; \psi) = \Pr(n_1, n_2|T; \psi), \quad [22]$$

where $\psi = \{t, s, \xi_0, \xi_1, \xi_2, f_1, f_2\}$ is a vector of the additional unknown (nuisance) parameters.

The likelihood function for multiallelic and/or multilocus data can be obtained by multiplying the probability in Eq. **20** for each different rare allele. This is because the birth–death process determining the frequency of each allele is independent. In other words, if r is the number of rare alleles (from the same or from different loci), and $n_{i,z}$ is the number of copies of the zth allele in population i, the probability of observing a configuration $\mathbf{n} = \{\{n_{1,1}, n_{2,1}\}, \{n_{1,2}, n_{2,2}\}, \ldots \{n_{1,r}, n_{2,r}\}\}$ is given by

$$L(T|\mathbf{n}; \psi) = \Pr(\mathbf{n}|T; \psi) = \prod_{z=1}^{r}\Pr(n_{1,z}, n_{2,z}|T; \psi). \quad [23]$$

In the sections that follow, we will briefly analyze the behavior of a maximum likelihood estimator of the divergence time T based on Eq. **23** when applied to hypothetical and to real data. Hereafter we will use the abbreviation *RD* to refer to the *r*are alleles based estimator of *d*ivergence time.

Properties of the Estimator

In this section we will examine several hypothetical data sets to illustrate qualitatively the effects of the parameters of the model on the shape of the likelihood function.

Effect of Population Growth. Suppose first that the total number of copies $n_1 + n_2$ of a rare allele sampled in two divergent populations is 200, and that both populations have grown exponentially. As one might expect, recent and ancient divergence times will often result in similar and different numbers of copies of a rare allele in the descendent populations, respectively. In fact, the estimated divergence time obtained by maximizing Eq. **23** increases when the difference between n_1 and n_2 is increased (see Fig. 1*a*). This is, of course, the most important property of the estimator, and suggests that the number of copies of a rare allele contains some information about the time at which the population was subdivided. Fig. 1*a* also suggests that for very ancient divergence events this information about the divergence time is ultimately lost, since the different possible data configurations become equally probable.

If the growth rate is decreased, the estimated divergence time tends to increase (Fig. 1*b*). This effect is related to the increase with the growth rate of the variance of the number of lineages in a birth–death process (14). After population subdivision, the same difference $n_1 - n_2$ is reached more rapidly in fast growing populations.

The influence of the growth rate on the estimated divergence time appears enormous when considering Fig. 1*b*. However, a variation of the exponential growth rate by a factor of 5 has to be regarded as enormous as well. For example, the final size of a population which started growing 500 generations ago from an initial size of 1,000 is expected to be either less than 15,000 or more than 250 million, depending on whether the growth rate is 0.005 or 0.025. Assuming that the growth rates of the populations can be at least roughly estimated, the variation of *RD* due to such uncertainties may be much smaller. An increase of the population growth rate from 0.015 to 0.025 in Fig. 1*b* (which is still a substantial increase), for example, results only in a decrease of the estimated divergence time from about 300 to 200 generations.

Effect of Mutation Age. Another potential source of uncertainty for the estimator we propose arises from assumptions about the ages of mutations. Allele ages are, in general, not known and must be estimated from the data (for some human diseases it is conceivable that historical information can be used). Different methods exist for estimating allele age (12, 15, 16), but the confidence intervals of these estimates are typically large.

The likelihood curves for several extreme situations in terms of allele age serve to illustrate this point. In Fig. 2*a*, the maximum likelihood estimator *RD*, as well as its support interval (i.e., the radius of curvature of the likelihood) do not seem to be strongly affected by the mutation age. In other words, even if the assumed mutation age varies from 500 to 1500 generations, the estimated divergence time lies within the same region. It can also happen,

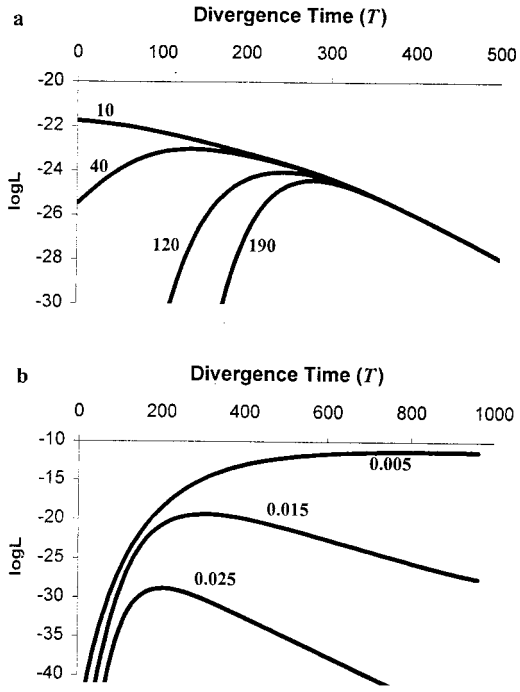

FIG. 1. (*a*) The log-likelihood of four different data sets as a function of the divergence time. In all cases $(n_1 + n_2) = 200$, $f_1 = f_2 = 0.01$, $(T + t) = 1000$, $\xi_0 = \xi_1 = \xi_2 = 0.02$. The plotted numbers correspond to $(n_1 - n_2)$. (*b*) The log-likelihood as a function of the divergence time when $n_1 = 40$, $(n_1 + n_2) = 200$, $f_1 = f_2 = 0.01$, $(T + t) = 1000$. The plotted numbers correspond to the growth rate, assuming $\xi_0 = \xi_1 = \xi_2$.

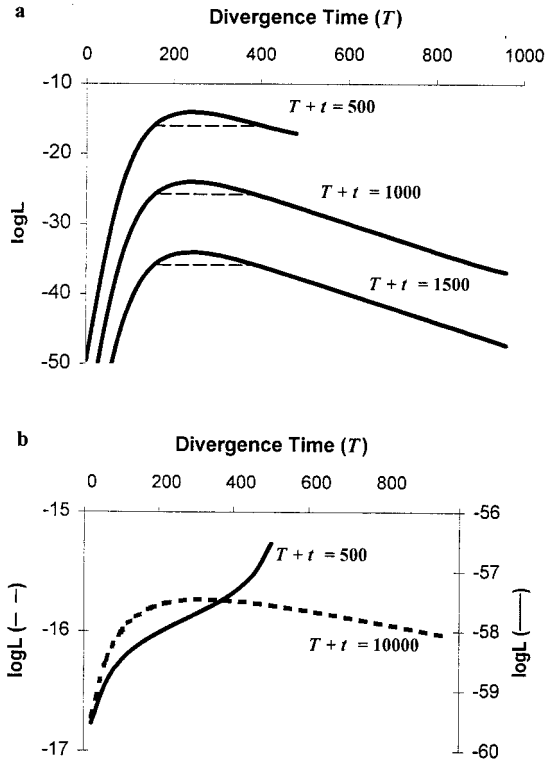

FIG. 2. The log-likelihood as a function of the divergence time for different ages ($T + t$) of the mutant. In all cases, $(n_1 + n_2) = 200$, $f_1 = f_2 = 0.01$. (*a*) $\xi_0 = \xi_1 = \xi_2 = 0.02$, $n_1 = 40$. The dashed lines indicate the 2 log-likelihood units of support intervals. (*b*) $\xi_0 = \xi_1 = \xi_2 = 0.001$, $n_1 = 85$. The model we used assumes that the mutation arose before the split of the ancestral population in two descendent populations. For this reason, some curves terminate earlier than others.

however, that very different assumptions about the age of the allele result in very different estimates of the divergence time (Fig. 2*b*). In addition, as expected (and also shown by Fig. 2*a*), assuming different allele ages affects the absolute value of the likelihood. This behavior implies that, when more alleles are simultaneously used to compute *RD*, a single allele whose age is very poorly estimated could have a differentially strong effect on the likelihood estimate of T.

One possible solution to this problem is to simultaneously estimate the divergence time and the allele ages. In this approach, however, each additional allele would introduce an additional parameter to estimate, thus reducing the power of the method. Instead, we therefore decided to analyze the behavior of a modified version of the estimator *RD*, hereafter called *RDc*. The estimator *RDc* is based on the probability of an observed configuration (n_1, n_2) conditioned on the sum $(n_1 + n_2)$. This probability can be simply derived by using the rule of conditional probabilities as:

$$\Pr(n_1, n_2 | n_1 + n_2) = \frac{\Pr(n_1, n_2)}{\sum_{i=1}^{n_1+n_2-1} \Pr(i, n_1 + n_2 - i)}. \quad [24]$$

In principle, since the expected sum $(n_1 + n_2)$ depends mainly on the allele age, conditioning the probability of a configuration on this sum should reduce the influence of the allele age on the likelihood of T.

The likelihood function based on Eq. **24** for the data sets in Fig. 2 has less curvature because most of the information in the total number of copies is disregarded. The estimator *RDc* obtained by conditioning proved to be largely insensitive to the age of the mutation. For example, if *RDc* is used instead of *RD*, the three data sets used in Fig. 2*a* have only a single likelihood function, and the two data sets in Fig. 2*b* have very similar likelihood functions with the maximum value separated by only about 100 generations (results not shown).

Finally, we used Monte Carlo simulation to examine how uncertainty about the allele age might affect the estimators *RD* and *RDc*. Five hundred samples of either 10 or 50 alleles from two divergent populations were simulated, assuming moderately small growth rates ($\xi_0 = \xi_1 = \xi_2 = 0.005$) and sampling fractions ($f_1 = f_2 = 0.005$). In the first set of simulations (upper part of Table 1), the age of each allele was fixed at 600 generations, whereas in the second set (lower part of Table 1) the age of each allele was randomly assigned from a uniform distribution with lower and upper limits equal to 400 and 800 generations, respectively. The simulated populations were assumed to have diverged $T = 200$ generations ago, and the number of copies of each allele was simulated by using the same birth–death model we used to derive the likelihood function. When these parameters were used: the average number of copies of each allele in each population was 6.8 and 8.6 for the first and second sets of simulations, respectively, and in both cases more than 10% of the alleles were present in a single copy in a population. These are, of course, unrealistically small numbers, but they allowed us to analyze large number of samples. The calculation of the likelihood function for a single data set is quite computationally intensive, especially for the *RDc* estimator.

The maximum likelihood estimate of the divergence time was obtained for each sample by using *RD* and *RDc* and assuming that the growth rates and the sampling fractions were known, as well as the age of each allele (600 generations) for the analysis of the first set of simulations. In the analysis of the second set of simulations, where the real age of the alleles varied between 400 and 800 generations, the *RD* and *RDc* were computed assuming the same fixed age (800 generations) for each allele.

The results (Table 1) show that if the age of the alleles is known, both *RD* and *RDc* are almost unbiased, with *RD* having a slightly lower standard deviation (SD) than does *RDc*. On the other hand, when the age of the alleles is not known, and estimates are calculated by assuming that all alleles have an (incorrect) age equal to that of the maximum possible age, *RDc* has consistently less bias than *RD*, and also lower SD when more rare alleles are considered. The higher SD of *RDc* when data sets of 10 alleles were considered is mainly due to the fact that for a small fraction of the samples the estimated divergence time was equal to the assumed mutation age. This kind of edge effect disappeared when more alleles were sampled.

Application to Cystic Fibrosis

As an example application of the method developed in this paper, we considered three cystic fibrosis (CF) mutations in four human populations from the paper by Estivill *et al.* (17). The three most frequent CF mutations—namely ΔF508, G542X, and N1303K—

Table 1. Results of the simulations

Simulation set	r	Estimator	Mean	SD	Range
First	10	*RD*	188.3	121.6	0.0–600.0
		RDc	196.1	135.5	0.0–600.0
	50	*RD*	194.8	49.6	56.0–314.2
		RDc	197.0	55.4	32.0–334.2
Second	10	*RD*	123.1	108.2	0.0–355.2
		RDc	212.7	154.0	0.0–800.0
	50	*RD*	130.7	82.9	0.0–246.2
		RDc	212.7	57.8	72.1–331.2

In the first set, alleles have all the same age of 600 generations, and the divergence time is estimated assuming that the allele age is known. In the second set, alleles can have any age between 400 and 800 generations with the same probability, and the divergence time is estimated assuming that every allele has the same age of 800 generations. $\xi_0 = \xi_1 = \xi_2 = 0.005$; $f_1 = f_2 = 0.005$; r = number of alleles; actual divergence time $T = 200$.

Table 2. CF data from four European populations

Population	$2N$ ($\times 10^6$)	No. of CF chromosomes			
		Total	ΔF508	N1303K	G542X
Sardinia	3.3	141	82	4	8
Italy	114.8	3,524	1,795	156	156
Denmark	10.6	678	591	7	4
Turkey	112.9	141	49	9	4

N is the present-day population size. Total refers to the number of CF chromosomes in the sample. The numbers of CF chromosomes with specific mutations are given in the last three columns.

will be used to estimate the pairwise divergence times between Italy, Sardinia, Denmark, and Turkey. Since the model used to derive the estimators assumes isolated populations, we expect that gene flow processes would result in an underestimation of the divergence time. This effect, which is probably minor for European populations that experienced reasonably low migration rates (18), should of course be kept in mind.

The population data are shown in Table 2, and the results provided by *RD* when the ages of ΔF508, G542X, and N1303K were set to 50,000, 35,000 and 35,000 years, respectively (16, 17), and the generation time is 20 years are shown in Table 3. Due to the large number of ΔF508 copies, the computation of *RDc* using the complete data sets would be unreasonably slow. Therefore, we computed *RDc* assuming smaller sizes for ΔF508 samples in Italy and Denmark. *RD* and *RDc* provided similar estimates (*RDc* having larger confidence intervals), and we therefore report only the results for *RD*.

Populations are assumed to grow exponentially with a rate of 0.005 per generation, which is compatible with an Upper Paleolithic demographic expansion often suggested for European populations (19–21). We also analyzed these data sets assuming a growth rate of 0.025, which is equivalent to assigning a selective advantage of 0.02 to CF carriers (lethal CF homozygotes are ignored because of their low frequency). Heterozygote advantage has long been proposed as an explanation for high prevalence of CF among Caucasoids (22–24). Recent analyses, however, were either unable to find evidence for such an effect (25, 26), or showed analytically that the high prevalence of some deleterious mutations is not unexpected in expanding populations (27).

The sample size, which is needed to compute the sampling fractions f_i and f_j for each pair of populations *i* and *j*, was estimated from the total number of CF mutants in the samples, assuming an incidence of the disease of 1 in 2,500 newborns (28) in each population. Finally, as we do not have any information about the ratio *s* of the divergent populations at the split, we assumed a ratio equal to the ratio between the present-day population sizes.

The results obtained when the population growth rate is fixed to $\xi = 0.005$ for all populations (thus excluding a selective advantage for CF carriers) suggest an Upper Paleolithic divergence (about 15,000 years ago, on the average) between Denmark and the Mediterranean populations (see Table 3). This result is compatible with several previous genetical analyses of European populations (29), suggesting a relatively recent divergence time even between very distant populations, and also suggesting no major impact in Northern Europe of Neolithic dispersion (30) from the Middle East.

The comparisons among the Mediterranean populations provide estimates of the divergence time ranging from 0 (Italy vs. Sardinia) to 18,000 years (Italy vs. Turkey). In contrast to the comparison with the Danish population, however, very short divergence times cannot be excluded for any pairs of Mediterranean populations due to the large confidence intervals.

Finally, we note that, among all the comparisons, Sardinia and Italy show the closest genetic relationship. Classical markers have often identified Sardinians as a genetical outlier in Europe (31), but the same pattern is not observed when the sequence of the mitochondrial control region is considered (32). In a previous analysis of the relative frequencies of CF alleles in Italy, Rendine *et al*. (33) found a certain level of divergence between Sardinians and other Italian regions, which was, however, mainly due to the private mutation T338I. All in all, it seems, therefore, that strong drift effects and the appearance of some new mutations after the relatively recent colonization of this island (around 9,000 years ago) might explain the peculiarity of the Sardinians. Their relationship with other Europeans, and especially with other Italians, is, however, still evident.

Table 3. Application of *RD* to data from the four European populations

		Divergence time estimates, thousands of years		
		Sardinia	Italy	Denmark
Italy	MLE	0.0		
	2SUI	0.0–11.4		
	3SUI	0.0–13.8		
Denmark	MLE	10.8	16.4	
	2SUI	5.6–17.7	12.5–22.1	
	3SUI	4.5–19.7	11.6–23.8	
Turkey	MLE	10.0	18.4	18.6
	2SUI	0.4–17.2	0.0–28.4	14.2–24.7
	3SUI	0.0–18.9	0.0–31.1	13.2–26.8

MLE is the maximum likelihood estimate of the divergence time. 2SUI is the interval of support computed as the divergence times at two units of support from the best supported value; this interval corresponds roughly to the 95% confidence interval. 3SUI is the interval of support computed as the divergence times at three units of support from the best supported value.

The results we obtained by setting the growth rate to 0.025 (thus assuming a positive selective effect of CF mutations in heterozygotes) gave a maximum likelihood estimate of the divergence time between 1/4 and 1/3 of the previous estimates (of course with the exception of the Italy–Sardinia comparison, which resulted again in a divergence time of 0). In other words, if the CF carriers had experienced a selective advantage of about 2% at any time in the past, our results would be consistent with a more recent (Neolithic) divergence also between populations as distant as Denmark and Turkey. Interestingly, these estimates would support the analysis of 4 microsatellite markers by Chikhi *et al*. (18), who found that the divergence time between pairs of European populations (estimated from the variance in the number of repeats) never exceeded 6,000 years. Only the analysis of other alleles or loci and more accurate estimates of the relevant demographic and selection parameters will clarify this point.

We also analyzed the CF data sets assuming earlier ages for the three mutations. Using linkage disequilibrium patterns, Serre *et al*. (15) estimated the age of ΔF508 between 3,000 and 6,000 years, and Kaplan *et al*. (34) suggested an age of 17,000 years for the same mutation. We assigned the intermediate age of 10,000 years to each allele in our data set. All population comparisons provided a maximum likelihood estimate of the divergence time equal to the age of the mutation. In principle, these estimates cannot be excluded, and due to the large confidence intervals, they are not incompatible with the results we obtained assuming older allele ages. We note, however, that CF allele ages have been estimated assuming a single panmictic European population. Since the age of a nonrecurrent allele shared by two disjunct populations must necessarily be older than the age of the population subdivision event, it is possible that the estimated age for some CF mutations is too recent.

Discussion

In this paper, we have derived some theory for a birth–death process used to describe the population dynamics of rare alleles

in a simple model of population divergence. This theory was used to derive two maximum likelihood estimators, *RD* and its conditional version *RDc*, of the time at which two recently isolated subpopulations diverged from a common ancestral population.§

The appropriate data for applying these estimators are the number of copies of one or more rare alleles at the same or different loci sampled from the descendent populations. It is assumed that each allele has arisen by nonrecurrent mutation in the ancestral population. Data on rare disease mutations in humans that are widespread in several populations seems therefore to be the most suitable for applying the method, since the amount of information on such mutations that is available for population studies is rapidly growing. For example, Estivill *et al*. (17) recently reported the geographic distribution of almost 30,000 CF chromosomes, and equivalent databases for other disease alleles are currently being developed.

In principle, the proposed method could also be applied to alleles that are not rare if the population sizes are not regulated by density dependence (e.g., in rapidly fluctuating island populations or in populations that are far from carrying capacity) (14). In this case, the main assumption of the birth–death process approximation (that individuals reproduce independently of one another) is still valid, since individual birth and death rates do not depend on population density (14).

Maximum likelihood estimators always require some knowledge about the parameters used in the model; the divergence time estimators proposed here are no exception. In particular, information on the demography of the populations and on the age of the alleles considered is needed to compute *RD* or *RDc*. One might therefore suspect that computationally easier methods, such as those based on Wright's F_{st} (5) or on the net number of substitutions (1), should be preferred. These methods, however, rely on strong assumptions about the demography of the populations, and if these assumptions are violated (which is often the case), the results might be difficult to interpret. If reasonable approximations of the parameters of the model are available, it should often be the case that the estimators *RD* and *RDc* we propose here will provide more accurate estimates of the divergence time. Large-scale simulation studies are needed to evaluate the overall performance of the different estimators of population divergence times for a range of biologically reasonable demographic conditions.

Our qualitative analysis of the estimators suggests that the influence of the demographic parameters on *RD* or *RDc* is pronounced only if radically different demographic scenarios are considered. In humans at least, historical or archaeological data may often provide independent information about the demography of a population (35). Alternatively, genetical data can potentially be used to distinguish between stable and growing populations, and, if necessary, to estimate the rate of exponential growth of population (36–40).

The ages of alleles can also affect estimates of the divergence time, even though our results indicate that this influence is probably not great. If the age of the analyzed alleles can be additionally estimated (e.g., see refs. 12, 15, and 16), the simpler and most powerful estimator *RD* is preferable. When, however, the age of an allele cannot be estimated, even approximately, and the data consist of several alleles, each with relatively small sample sizes, *RDc* should be preferred. The computation of RDc is quite computationally intense, but this estimator proved to be less affected by incorrect estimates of allele age than was *RD*.

Finally, when *RD* was applied to three CF mutations in four European populations, we obtained results that appear consistent and compatible with previous studies. In particular, the divergence between Danish and Turkish populations was estimated to have occurred between 15,000 and 25,000 years ago. On the other hand, when three Mediterranean populations (Italy, Sardinia, and Turkey) were compared, the CF data set we analyzed was only able to exclude very old (>30,000 years) divergence times, with an average estimate of about 10,000 years. All these estimates, however, reduced substantially when a selective advantage for CF carriers (22–24) is introduced. This result, unfortunately, does not allow one to distinguish between several current hypotheses (see refs. 18, 41, and 42) on the relative contribution of Paleolithic and Neolithic genes to the present European gene pool. A better understanding of the selection process affecting the CF locus is therefore needed, as well as the analysis of additional rare mutations at additional loci. We expect the amount of these kind of data available for human populations to increase rapidly in the near future.

We thank Monty Slatkin for helpful discussions and for critical reading of this manuscript. G.B. was supported by National Institutes of Health Grant GM28428 to L. L. Cavalli-Sforza. B.R. was supported by National Institutes of Health Grants GM40282 to M. Slatkin and HG01988 to B.R.

1. Takahata, N. & Nei, M. (1985) *Genetics* **110,** 325–344.
2. Nei, M. (1987) *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York).
3. Goldstein, D. B., Ruiz Linares, A., Cavalli-Sforza, L. L. & Feldman, M. W. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 6723–6727.
4. Poloni, E. S., Semino, O., Passarino, G., Santachiara-Benerecetti, A. S., Dupanloup, I., Langaney, A. & Excoffier, L. (1997) *Am. J. Hum. Genet.* **61,** 1015–1035.
5. Wright, S. (1951) *Ann. Eugen.* **15,** 323–354.
6. Cavalli-Sforza, L. L. & Bodmer, W. F. (1971) *The Genetics of Human Populations* (Freeman, San Francisco).
7. Excoffier, L., Smouse, P. & Quattro, J. M. (1992) *Genetics* **131,** 479–491.
8. Slatkin, M. (1995) *Genetics* **139,** 457–462.
9. Nielsen, R., Mountain, J. L., Huelsenbeck, J. P. & Slatkin, M. (1998) *Evolution* **52,** 669–677.
10. Kendall, D. G. (1948) *Ann. Math. Stat.* **19,** 1–15.
11. Nee, S., May, R. M. & Harvey, P. H. (1994) *Phil. Trans. R. Soc. Lond. B* **344,** 305–311.
12. Slatkin, M. & Rannala, B. (1997) *Am. J. Hum. Genet.* **60,** 447–458.
13. Medhi, J. (1994) *Stochastic Processes* (Wiley, New York), 2nd Ed.
14. Rannala, B. (1997) *Heredity* **76,** 417–423.
15. Serre, J. L., Simon-Buoy, B., Mornet, E., Jaume-Roig, B., Balassopolous, A., Schwartz, M., Taillandier, A., Boué, J. & Boué, A. (1990) *Hum. Genet.* **84,** 449–454.
16. Morral, N., Bertranpetit, J., Estivill, X., Nunes, V., Casals, T., Gimenez, J., Reis, A., Varon Mateeva, R., Macek, M., Jr., Kalaydjieva, L., *et al*. (1994) *Nat. Genet.* **7,** 169–175.
17. Estivill, X., Bancells, C., Ramos, C. & the Biomed CF Mutation Analysis Consortium (1997) *Hum. Mutat.* **10,** 135–154.
18. Chikhi, L., Destro-Bisol, G., Bertorelle G., Pascali, V. & Barbujani, G. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 9053–9058.
19. Harpending, H. C., Sherry, S. T., Rogers, A. R. & Stoneking, M. (1993) *Curr. Anthropol.* **34,** 483–496.
20. Rogers, A. R. & Jorde, L. B. (1995) *Hum. Biol.* **67,** 1–36.
21. Comas, D., Calafell, F., Mateu, E., Perez-Lezaun, A., Bosch, E. & Bertranpetit, J. (1997) *Hum. Genet.* **99,** 443–449.
22. Quinton, P. M. (1982) in *Fluid and Electrolytes Abnormalities in Exocrine Glands in Cystic Fibrosis*, eds. Quinton, P. M., Martinez, R. J. & Hopfer, U. (San Francisco Press, San Francisco), pp. 53–76.
23. Romeo, G., Devoto, M. & Galietta, L. J. (1989) *Hum. Genet.* **84,** 1–5.
24. Bertranpetit, J. & Calafell, F. (1996) in *Variation in the Humane Genome*, eds. Chadwick, D. & Cardew, G. (Wiley, Chichester, U.K.), pp. 97–114.
25. De Vries, H. G., Collée, J. M., de Walle, H. E. K., van Veldhuizen, M. H. R., Smit Sibinga, C. Th., Scheffer, H. & ten Kate, L. P. (1997) *Hum. Genet.* **99,** 74–79.
26. Bertorelle, G. & Barbujani, G. (1997) *Ann. Hum. Genet.* **61,** 532–533.
27. Thompson, E. A. & Neel, J. V. (1997) *Am. J. Hum. Genet.* **60,** 197–204.
28. Boat, T. F., Welsh, M. J. & Beaudet, A. L. (1989) in *The Metabolic Basis of Inherited Disease*, eds. Scriver, C. R., Beaudet, A. L., Sly, W. S. & Valle, D. (McGraw-Hill, New York), pp. 2649–2680.
29. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. (1994) *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton, NJ).
30. Ammerman, A. J. & Cavalli-Sforza, L. L. (1984) *The Neolithic Transition and the Genetics of Populations in Europe* (Princeton Univ. Press, Princeton, NJ).
31. Piazza, A., Cappello, N., Olivetti, E. & Rendine, S. (1988) *Ann. Hum. Genet.* **52,** 203–213.
32. Stenico, M., Nigro, L., Bertorelle, G., Calafell, F., Capitanio, M., Corrain, C. & Barbujani, G. (1996) *Am. J. Hum. Genet.* **59,** 1363–1375.
33. Rendine, S., Calafell, F., Cappello, N., Gagliardini, R., Caramia, G., Rigillo, N., Silvetti, M., Zanda, M., Miano, A., Battistini, F., *et al*. (1997) *Ann. Hum. Genet.* **61,** 411–424.
34. Kaplan, N. L., Lewis, P. O. & Weir, B. S. (1994) *Nat. Genet.* **8,** 216–218.
35. Mussi, M. (1990) in *The World at 18,000 BP: High Latitudes*, eds. Soffle, O. & Gamble, C. (Allen & Unwin, London), pp. 126–147.
36. Roger, A. R. & Harpending, H. (1992). *Mol. Biol. Evol.* **9,** 552–569.
37. Slatkin, M. & Hudson, R. R. (1991) *Genetics* **129,** 555–562.
38. Griffiths, R. C. & Tavaré, S. (1994) *Phil. Trans. R. Soc. Lond. B* **344,** 403–410.
39. Polanski, A., Kimmel, M. & Chakraborty, R. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 5456–5461.
40. Kuhner, M. K., Yamato, J. & Felsenstein, J. (1998) Genetics **149,** 429–434.
41. Richards, M., Corte-Real, H., Forster, P., Macaulay, V., Wilkinson-Herbots, H., Demaine, A., Papiha, S., Hedges, R., Bandelt, H.-J. & Sykes, B. (1996) *Am. J. Hum. Genet.* **59,** 185–203.
42. Barbujani, G., Bertorelle, G. & Chikhi, L. (1998) *Am. J. Hum. Genet.* **62,** 488–491.

§A computer program to compute the divergence time estimators proposed in this paper can be downloaded from the WWW site at the URL http://mw511.biol.berkeley.edu.