Annu. Rev. Genomics Hum. Genet. 2000. 01:225–49

ESTIMATING ALLELE AGE

Montgomery Slatkin[1] and Bruce Rannala[2]
[1]*Department of Integrative Biology, University of California, Berkeley, California 94720–3140; e-mail: slatkin@socrates.berkeley.edu*
[2]*Department of Ecology and Evolution, State University of New York, Stony Brook, New York*

Key Words coalescent theory, population genetics, linkage disequilibrium, gene genealogy

■ **Abstract** The age of an allele can be estimated both from genetic variation among different copies (intra-allelic variation) and from its frequency. Estimates based on intra-allelic variation follow from the exponential decay of linkage disequilibrium because of recombination and mutation. The confidence interval depends both on the uncertainty in recombination and mutation rates and on randomness of the genealogy of chromosomes that carry the allele (the intra-allelic genealogy). Several approximate methods to account for variation in the intra-allelic genealogy have been derived. Allele frequency alone also provides an estimate of age. Estimates based on frequency and on intra-allelic variability can be combined to provide a more accurate estimate or can be contrasted to show that an allele has been subject to natural selection. These methods have been applied to numerous cases, including alleles associated with cystic fibrosis, idiopathic torsion dystonia, and resistance to infection by HIV. We emphasize that estimates of allele age depend on assumptions about demographic history and natural selection.

INTRODUCTION

Geneticists have (almost) created a time machine. It is now possible to work backwards from contemporary observations of genetic variation to inferences about past processes. Such inferences will never be easy, and complete resolution of historical events affecting genetic variation will never be achieved, but, for much of the history of humans and other species, it is our only chance. Genetic analysis has already told us much about the past growth and dispersal of human populations. The same principles can be used to infer the ages of individual alleles.

The age of an allele is the time since it was created by mutation. Recent interest in estimating allele age, sometimes referred to as "dating" an allele, stems from the extensive DNA sequencing and marker typing being done to map and clone alleles that cause genetic diseases. Variability at closely linked polymorphic markers and allele frequency itself provide estimates of allele age. We review methods

1527-8204/00/0728-0225$14.00

for estimating allele age, by considering variation at marker loci and then allele frequency. Next we discuss how different estimates can be combined to provide a single estimate or contrasted to provide additional information. We emphasize that all estimates of allele age depend on assumptions about past genetic and demographic processes. Finally, we review several examples in which allele ages have been estimated, choosing those that best illustrate various aspects of the problem. Because of space limitations, we restrict our discussion to estimating the age of alleles associated with diseases or distinctive phenotypes. We do not discuss the separate class of problems that arise in estimating the ages of mutations found on gene genealogies of nonrecombining regions of the genome (particularly mitochondrial and Y chromosomes; 13, 33, 50).

Estimating allele age is done partly out of curiosity and partly from the desire to make additional use of data that are gathered for other purposes. Curiosity and thoroughness are not bad reasons, but there may be others as well. Different kinds of data provide different information about allele age and may point to an important role for natural selection or other processes, which were not originally envisioned. Estimating ages of several alleles at the same locus or of alleles at different loci may help to sort out demographic processes and aid in reconstructing population histories (4). Most studies of allele age have been carried out independently, but, as such studies become commonplace, they can be usefully combined and examined to identify broader patterns.

WHAT IS AN ALLELE?

For our purposes, an allele is defined as an alteration in DNA sequence—a substitution, deletion, or insertion—at a single nucleotide position. We call this alteration the defining mutation. For example, the defining mutation of the $\Delta F508$ allele of the *CFTR* locus is the deletion of the three nucleotides that code for amino acid 508 in the CFTR protein (19). By assumption, the defining mutation obeys the rules of Mendelian inheritance. With this definition, an allele does not have to be associated with a phenotype, but, in practice, questions about allele age are usually posed for alleles that have obvious phenotypic effects. Different copies of an allele carry the same defining mutation but do not otherwise have to be identical in sequence. In fact, differences in sequence among different copies, what we call intra-allelic variability, provide important information about allele age. As an example, Morral et al (29) found extensive variation among copies of $\Delta F508$ at three intronic microsatellite loci.

Our definition of an allele excludes those that are distinguished from one another by differences at two or more distinct sites, as is often found, for example, at loci in the major histocompatibility complex (MHC) in mammals. When two or more alterations are required, defining allele age becomes problematic, particularly because intragenic recombination can create and recreate the same allele. Restricting our focus to a single defining mutation ensures that the meaning of allele age is

unambiguous; it is the time since the occurrence of the defining mutation inherited by all later copies.

Our definition of an allele does not assume that it can arise only once by mutation. Some alleles, including the *S* allele of the beta-globin locus that causes sickle cell anemia, have arisen more than once (34). Methods for estimating allele age have to be applied separately to copies that descend from different mutations, but these methods can be adapted to estimating the number of independent origins.

INTRA-ALLELIC VARIABILITY

Moment Estimator

The relationship between allele frequency and allele age was first analyzed by population geneticists in the 1970s, but recent estimates of the ages of disease-associated alleles have been based primarily on intra-allelic variability. In the first example of this kind of analysis, Serre et al (41) surveyed two restriction fragment length polymorphism (RFLP) sites that are closely linked to the $\Delta F508$ allele of *CFTR*. They assayed haplotypes at these two sites, treated as two diallelic loci, in a pooled sample of 240 French families. When only the *E* locus was considered, 90.3% of the $\Delta F508$ chromosomes carried the marker allele designated *E2*, whereas only 28.2% of the normal chromosomes carried that allele. The excess of *E2* on $\Delta F508$ chromosomes is attributed to the recent origin of $\Delta F508$ on a chromosome carrying *E2*. Subsequent recombination with normal chromosomes then created the few *E1*-$\Delta F508$ chromosomes. As shown in Equation 1, the theory of recombination provides a simple relationship between the frequencies of *E1* and *E2* on $\Delta F508$ chromosomes and *t*, the time since $\Delta F508$ arose:

$$x(t) - y = (1 - c)^t(1 - y), \qquad 1.$$

where *c* is the recombination rate, $x(t)$ is the expected frequency of *E2*-$\Delta F508$ in generation *t*, and *y* is the frequency of *E2* on normal chromosomes, which is assumed to be constant during the time since the mutation occurred (41). The recombination rate *c* is assumed to be known, and $x(t)$ and *y* are obtained from the genetic survey, so Equation 1 can be solved to yield an estimate of *t*, the allele age.

$$t = \frac{1}{\ln(1 - c)} \ln\left(\frac{x(t) - y}{1 - y}\right). \qquad 2.$$

In this example, $x(t) = 0.903$, and $y = 0.282$, so, if *c* is 0.0008, $t = 181.4$ generations.

This method for estimating allele age was first used by Serre et al but gained popularity after the study by Risch et al (39). Technically, Equation 2 is a method-of-moments estimator, which we will call a moment estimator, of *t*, because the

estimate results from equating the observed proportion of nonrecombinant chromosomes with the proportion expected if the true value of t is the estimated value. (An alternative discussed below is a maximum-likelihood estimator.)

We have assumed so far that the ancestral marker allele can be identified unambiguously, but, in reality, it is unknown. Usually, a strong association between a particular marker allele and the defining mutation suggests which marker allele was ancestral, but in some cases that is not so. For example, at one of the marker loci (*D9S64*) in the Risch et al (39) study of idiopathic torsion dystonia (ITD), three alleles were found in substantial frequency (0.333, 0.278, and 0.194), making it difficult to know which was initially linked to the disease-associated allele. This situation can be dealt with in several ways. One approach is to estimate the age by assuming that each marker allele was ancestral and then to average the estimates, weighted by the marker frequency on normal chromosomes. Another approach is to estimate the age for each ancestral marker allele and present the results separately (27, 39).

Equations 1 and 2 show the close relationship between estimating allele age and disequilibrium mapping, for which the age t is assumed to be known and the problem is to estimate c, the recombination rate between the marker and the allele that causes a genetic disease (22). The moment estimate of c is obtained by solving Equation 1 for c instead of t. For disequilibrium mapping in an isolated population, such as the population of Finland, t is assumed to be the number of generations since the founding of the population ($\sim$100 for Finland), under the assumption that most or all copies of the disease-associated allele are descended from a single copy in the newly founded population (15). One of the difficulties in disequilibrium mapping in heterogeneous populations is that both allele age and map position are unknown.

CONFIDENCE INTERVAL

The moment estimator of age is appealing because it requires no population genetic and demographic assumptions, and it is easy to compute. It assumes only the exponential decay of initially perfect linkage disequilibrium (i.e. $x(0) = 1$), because of recombination. The question with the moment or any other estimator is how much confidence can be placed in it. That answer is much harder to obtain and requires an understanding of possible sources of uncertainty in the estimate. One source of uncertainty is in the genetic parameters—the recombination and mutation rates. These rates are small and consequently difficult to estimate. Many marker loci are so closely linked that the recombination rate cannot be estimated by genetic analysis and instead must be inferred from a radiation hybrid map. Serre at al (41) and others assume a range of values that are consistent with available data and report the resulting range in estimated ages obtained.

Another source of uncertainty is the intrinsic unpredictability of recombination and mutation. The probability of recombination can be calculated, but, on a

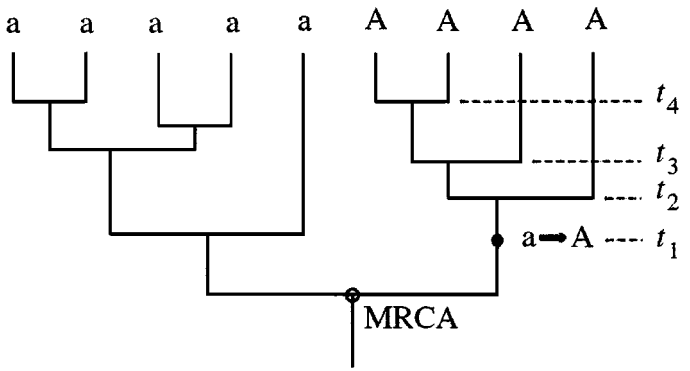

Figure 1 Illustration of gene genealogy for nine chromosomes and the intra-allelic genealogy for the four chromosomes carrying the *A* allele that arose by mutation (denoted by the *closed circle*) at time t_1 in the past. The point "MRCA" (denoted by the *open circle*) indicates the time of the most recent common ancestor of the gene genealogy. Time increases from the present ($t = 0$) to the past. The intra-allelic coalescence times are denoted by t_4, t_3, and t_2 in increasing order.

particular lineage, recombination either does or does not occur. Accounting for the unpredictability of recombination requires assuming or computing the numbers of genetic lineages that carried the allele at different times in the past. A way to visualize the problem is in terms of the genealogical history of a sample of chromosomes, some of which carry the allele, as illustrated in Figure 1. This genealogy represents the history of the site at which the defining mutation occurred. In all nine chromosomes, this site is descended from a most recent common ancestor (MRCA) at some time in the past. Each branch of the genealogy represents the sequence of chromosomes ancestral to those in the sample, and each node represents the appearance of two lineages from a single ancestral lineage. The times at which two or more lineages descend from a single ancestral lineage are called the coalescence times. With n tips of the genealogy representing n chromosomes in a sample, there are $n - 1$ coalescence times. We have drawn the genealogy as having only two lineages at each node (a bifurcating tree) because that is by far the most likely case, but we can allow for three or more lineages at a node by assuming that two or more of the coalescence times are equal.

The defining mutation occurs on one lineage at time t_1 in the past, and it is carried by all descendant lineages unless back mutation is allowed for. We call the part of the genealogy that carries the defining mutation the intra-allelic genealogy. Although allelic genealogy seems appropriate for this purpose, that term is already used in the study of MHC variations for the genealogical relationship among different alleles. The intra-allelic genealogy represents the history of the i mutant chromosomes in the sample. We denote the coalescence times of the intra-allelic genealogy by $t_i \ldots t_2$. The numbers of lineages that carried the allele at different

times in the past represent the net opportunity for recombination to alter the initially perfect association of the allele and markers carried by the ancestral chromosome at t_1.

Before discussing methods for taking account of the intra-allelic genealogy, we note that there is some ambiguity in the definition of allele age. In Figure 1, t_1 is the true age, which is the time at which the defining mutation occurred. The time t_2 is the time of the most recent common ancestor of all copies of the allele in the sample. Between t_1 and t_2, only one lineage carrying the defining mutation had any descendants in the sample. There were probably other copies of the allele present during that time interval, but they left no descendants in the sample. Different samples of the same allele might have different values of t_2, but they would have the same value of t_1. Furthermore, any recombination between the defining mutation and linked marker loci or any mutation at those markers would change the marker alleles on the chromosome that carried the defining mutation at t_2. Intra-allelic diversity can be generated between t_2 and the present, but not before t_2. Therefore, Equation 2 estimates t_2 and not t_1. For many purposes, estimating t_2 may be the goal, because that is the time after which intra-allelic variability arose, but in general the age of the most recent common ancestor is not the same as the age of the allele.

RANDOMNESS OF THE INTRA-ALLELIC GENEALOGY

The coalescence times of the intra-allelic genealogy determine the opportunity for generating intra-allelic variability. Unfortunately, the frequency of nonrecombinant haplotypes $[x(t)]$ in a sample does not itself determine the coalescence times, so the data used in Equation 2 are not sufficient to compute the confidence interval or other properties of the moment estimator, without making further assumptions. The moment estimator requires no assumptions about the intra-allelic genealogy, because the frequency of nonrecombinant haplotypes in the sample is proportional to the sum of the number of nonrecombinant haplotypes (either 0 or 1) for each chromosome in the sample. The expectation of a sum of random variables is equal to the sum of the expectations, even if the variables are not independent, so the expected frequency of nonrecombinant haplotypes depends only on the expectation for each chromosome, which is what Equation 1 provides.

The coalescence times of the intra-allelic genealogy are to some extent random, and their probability distribution depends on other factors, including allele frequency, the demographic history of the population, and the effects of natural selection, if any, on the allele. The complete distribution of intra-allelic coalescence times can be obtained only by a simulation method that relies on coalescent theory (12) and, even when the distribution of coalescence times is known, it is very difficult to obtain the general statistical properties of allele age estimates. The complete problem has not yet been solved, but numerous approximations have been proposed. It is important to distinguish between the intra-allelic

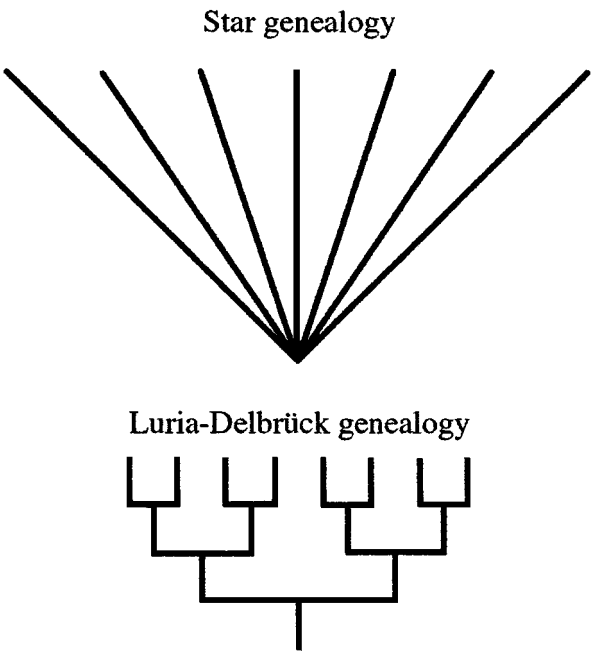

Figure 2 Two idealized shapes assumed for intra-allelic genealogies, as discussed in the text. A star genealogy assumes that all lineages arose at the same time in the past. A Luria-Delbrück genealogy assumes that every lineage bifurcates at regular time intervals.

coalescence times ($t_2, \ldots, t_i$ in Figure 1) and the coalescence times of the entire gene genealogy. The two sets of coalescence times may have similar distributions, as assumed by McPeek & Strahs (27) and by Goldstein et al (11), but in general they are not the same (43).

A very simple approximation for the intra-allelic genealogy is to assume it is a star genealogy, as shown in Figure 2. In a star genealogy, all lineages are descended independently from the ancestral chromosome at t_2, and all coalescence times are the same. If the intra-allelic genealogy is not a star genealogy, then different lineages are to some extent correlated, because they share some common ancestry between t_2 and the present ($t = 0$). Assuming a star genealogy eliminates all randomness in the numbers of ancestral lineages, so that the only uncertainty arises from whether recombination occurs on each lineage. Equation 1 tells us the probability for each lineage, and the independence of events on different lineages lets us multiply probabilities and obtain a confidence interval on the age. It is straightforward to show that, under the assumption of a star genealogy, the moment estimator is also the maximum-likelihood estimator. Risch et al (39) used this procedure to determine the confidence interval on estimated age of an allele that causes ITD in Ashkenazi Jews. The assumption of a star genealogy is justified

because, in a very rapidly growing population, the whole gene genealogy of the locus would be starlike (44); therefore, the genealogy of any subsample would also be starlike (38). The difficulty is that, although the intra-allelic genealogy may be starlike, slight differences from a perfect star genealogy can lead to magnified effects in a way similar to the process modeled by Luria & Delbrück (24).

Additional assumptions are needed to allow for variation in the number of ancestral lineages. Labuda et al (21) used the Luria-Delbrück theory to show that the moment estimator is biased under that model and to suggest that the moment estimate be increased by $-(1/r)\ln[ce^r/(e^r-1)]$, where c is the recombination rate, and r is the rate of past exponential population growth per generation. To illustrate, 7 of 54 recombinants were found between one locus (*ASS*) in the Risch et al (39) study of ITD, with $c = 0.018$. Equation 2 then yields an estimated age of 8.4 generations. The growth rate for the Ashkenazi Jewish population is $\sim$0.4055 per generation (39), so the Labuda et al correction is 7.2 generations, almost doubling the estimated age. The Luria-Delbrück model assumes a synchronously bifurcating genealogy (shown in Figure 2), which allows for change in the numbers of ancestral lineages but not for any randomness in those numbers.

Several other methods assume a stochastic model that generates the intra-allelic coalescence times. Some of these methods do not try to refine the moment estimator but instead compute the likelihood of the age as a function of the data. The likelihood function provides both a maximum-likelihood estimate (MLE) of age and a support interval that can be interpreted with suitable qualifications as a confidence interval.

We have found a simple approximate method for characterizing the allelic genealogy when the allele occurs at low frequency (45). We adapted existing results derived from the theory of linear birth-death processes to approximate the distribution of intra-allelic coalescence times. We assumed that the population grew exponentially at rate r in the past and that there was additive selection of strength s on the mutant allele. The coalescence times are found by drawing $i-1$ independent numbers from the probability distribution:

$$b(x) = \frac{[P(0,x)]^2 e^{-\xi x}}{2[f - P(0,t_1)e^{-t_1\xi}]}, \qquad 3.$$

where $P(0,x) = 2f\xi/(f-(f-2\xi)e^{-\xi t})$, $\xi = r+s$, and f is the fraction of the mutant chromosomes found in the sample. The distribution of coalescence times depends on the parameters only through $b(x)$, so its shape can tell us how similar the intra-allelic genealogy is to a star genealogy. If $b(x)$ is very narrow, then only a small range of coalescence times is possible, and the genealogy is nearly a star. If it is broad, then coalescence times are quite different from one another. The overall shape of $b(x)$ depends most strongly on ξ, the combined effects of population growth and selection. Increasing ξ reduces the width, as shown in Figure 3. Our results also show that, with rapid growth, $b(x)$ becomes nearly independent of t_1 as t_1 increases, and hence the extent of intra-allelic variability alone makes it difficult to exclude large values of t_1.

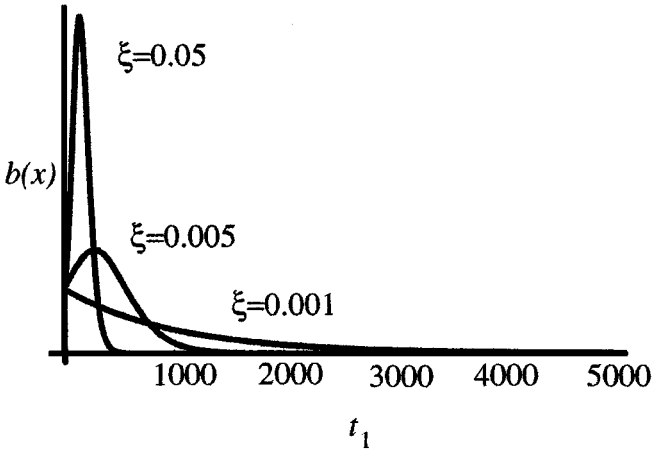

Figure 3 Graphs of the function $b(x)$ defined by Equation 3 in the text. The curves shown are for a fraction 0.002 of the population sampled.

The distribution of coalescence times can be combined with a model of recombination on each branch to allow efficient computation of the likelihood of t_1 as a function of the data (in this case the number of nonrecombinant haplotypes in the sample). The likelihood estimator provides an MLE of t_1, not t_2, because the model generates a distribution of $t_1 - t_2$, the time during which only one ancestral lineage was present. The method is easily modified to estimate t_2 instead. As an example, the data of Risch et al, cited above, led to an estimate of t_1 (and t_2 also) of 17.8 (39), compared with estimates of 8.4 generations from Equation 1 and of 15.6 generations after adding the correction factor of Labuda et al (21). In this case, $t_1 - t_2$ is nearly 0. A computer program to carry out this analysis (DMLE) is available from the State University of New York, Stony Brook (http://allele.bio.sunysb.edu/software.html).

Reich & Goldstein (38) used a different method to generate intra-allelic coalescence times. They assumed that the distribution of intra-allelic coalescence times is the same as the entire gene genealogy for a neutral locus, which is then easily simulated by using, for example, the C program described by Hudson (17). Reich & Goldstein justified their approximation by arguing that the intra-allelic genealogy will be nearly a star genealogy in a rapidly growing population.

Guo & Xiong (14) derived an approximate likelihood method that is applicable to data from one or more loci. Their method relies on the expected means, variances, and covariances of haplotype frequencies that are obtained from a diffusion approximation of a model of genetic drift. From these expectations, Guo & Xiong computed either a linear or quadratic Taylor series approximation to the likelihood surface. Their linear approximation is equivalent to assuming a star intra-allelic genealogy. They noted that this method leads to very rapid computations, and they

applied it to several published data sets. Their method does not allow for population growth because the corresponding diffusion equations cannot be solved analytically.

McPeek & Strahs (27) introduced another approximation to account for the dependence among lineages, created by the intra-allelic genealogy. They were concerned primarily with disequilibrium mapping but noted that the same approach can be taken to estimating allele age. They assumed that the correlation between all lineages is the same as the average computed from a coalescent model of ancestry. The result is that the estimated age is the same as that obtained by assuming a star intra-allelic genealogy but the confidence interval is wider by a calculable amount, representing a reduction in the sample size because of the nonindependence of lineages.

Multiple Marker Loci

In most published studies, data are available for more than one linked marker locus. With several marker loci, the most common approach is to analyze each separately, using one of the methods described in the previous section. Loci that are found in perfect disequilibrium with the defining mutation, meaning that no recombinants are present ($x(t) = 1$), are ignored. Instead, loci with a few recombinants are chosen, and the results of analyzing each such locus are presented. In addition, the size in centimorgans of the region sharing the "ancestral haplotype" is noted. For example, Moisio et al (28) found a shared haplotype spanning $\sim$8 cmorgans surrounding a mutation in a mismatch repair gene, *MLH1*. Such large shared haplotypes provide additional evidence for a relatively small age for the mutation, $\sim$20 generations for two mutations at *MLH1*.

Combining information from different marker loci is difficult, and the theory is not yet complete. One approach is based on generalizing the moment estimator for one locus. Serre et al (41) derived equations that are satisfied by the moment estimator when there are two linked loci. Guo & Xiong (14) corrected minor errors in Serre et al's formulas and provided an analytical solution to their equations. Several papers have derived estimators based on moments of haplotype frequencies for two or more linked marker loci, allowing for recombination as well as mutation (14, 27, 32, 38).

Guo & Xiong (14) described a method for calculating the expected haplotype frequencies at two linked loci in a sample of chromosomes. They assumed that the population is of constant size and that map distances between the markers and the disease allele are known. They suggested that a multipoint estimate (i.e. one that combines information across markers) of allele age could be obtained by minimizing the sum of the squared differences between expected and observed haplotype frequencies. They noted that this approach could be generalized to arbitrary numbers of markers, although they did not present an explicit theory for the general case. Guo & Xiong also suggested that multilocus haplotype frequencies could provide estimates of both the age and the location of the defining mutation.

An alternative approach is to use the likelihood of the observed multilocus haplotypes directly to estimate allele age. In this case, exact calculations do not appear possible. Guo & Xiong (14) discussed several approximate methods. One method uses the so-called composite likelihood, which is obtained by assuming that the joint likelihood is the product of marginal likelihoods. For estimating allele age, there are two kinds of marginal likelihood. One kind is the marginal likelihood that is calculated for each locus separately. This marginal likelihood is then multiplied across loci to obtain the composite likelihood. This approach was taken by Terwilliger (48) and Devlin et al (7) for disequilibrium mapping. The other kind of marginal likelihood is that of the multilocus haplotype computed for each chromosome separately. The compound likelihood is computed by multiplying across chromosomes. The second approach is equivalent to assuming a star intra-allelic genealogy because it assumes the independence of all of the chromosomes since t_2. The method proposed by Neuhausen et al (32) uses both approximations, multiplying marginal likelihoods across loci and across chromosomes. The accuracy and efficiency of composite likelihood estimators are difficult to predict because no general theory exists. It is clear that confidence intervals obtained from composite likelihoods are too narrow, because a composite likelihood assumes independence where it does not exist and, in effect, inflates the sample size.

McPeek & Strahs (27) introduced another method for combining information from several marker loci. They modeled the decay of the size of the region containing shared haplotypes and applied their method to both disequilibrium mapping and estimating allele age. Their results are comparable to those obtained by Guo & Xiong (14).

At this stage, methods for analyzing multilocus marker haplotypes must be regarded as preliminary because there are many important unanswered questions, especially related to the effects of past population growth, the utility of finding additional markers, and the overall effects of sample design.

ESTIMATES OF AGE BASED ON FREQUENCY

The frequency of an allele in a population does not change in an arbitrary manner; instead, it is governed by natural selection, genetic drift, mutation, and gene flow. Understanding the effects of these processes allows the estimation of allele age from only its current frequency. Although much of the underlying theory was developed >20 years ago, it has been largely ignored in recent discussions of allele age.

Neutral Alleles in Populations of Constant Size

Kimura & Ohta (20) were the first to consider the relationship between age and frequency. They showed that for a neutral allele with frequency p in a large randomly mating population of constant effective size N, the expected age t_1 is

approximately

$$E(t_1) = \frac{-2p}{1-p}\ln(p), \qquad 4.$$

where time is measured in units of $2N$ generations. An estimate of age is obtained by inserting the observed allele frequency on the right-hand side. For example, the expected age of an allele with frequency 2% is 0.16 in scaled time units. If $N = 10{,}000$, which is often regarded as a minimum estimate of the effective population size of modern humans during the period before recent growth, that would imply an age of 1600 generations or, assuming a generation time of 20 years, roughly 32,000 years.

To obtain a confidence interval on the age, we could use the variance in allele age, which was obtained by Li (23). Numerical analysis shows that the variance is quite large compared with the mean, implying that estimates obtained from Equation 4 have wide confidence limits. A simpler way to obtain a confidence interval is from the probability distribution of ages, which has been obtained by Watterson (52), Griffiths & Tavaré (12), and others. The distribution has a complicated mathematical form, but there is a simple approximation that we use here. Griffiths & Tavaré (12) proved that the cumulative probability distribution of allele age is

$$P(t_1 \le t) = E\left[(1-p)^{n(t)-1}\right], \qquad 5.$$

where E indicates expectation, p is the allele frequency in a sample of size n, and $n(t)$ is the number of lineages that are ancestral to the sample at t. The function $n(t)$ is random, and its distribution can be derived from coalescent theory. Evaluating the expectation requires extensive calculations or simulations, but we can write a simple approximation based on the result that, to a very good approximation, $n(t) \approx n/(1+nt/2)$ (45). Substituting this expression into Equation 5 gives

$$P(t_1 \le t) \cong (1-p)^{-1+n/(1+nt/2)}. \qquad 6.$$

Taking the derivative for t provides the probability distribution of t_1. The approximate distributions for $p = 0.01$ and $p = 0.1$ are shown in Figure 4. The distribution is skewed, implying that the expectation, given by Equation 4, is different from the MLE of t_1, which, from Equation 6, is approximately $-\ln(1-p) - 2/n$. For example, with $p = 0.01$, the $E(t_1) = 0.093$, but the MLE is 0.010, and, if $p = 0.1$, $E(t_1) = 0.510$ and the MLE is 0.105, all in units of $2N$ generation, if n is large.

We can use Equation 6 to find an approximate confidence interval. With $x = 0.01$, there is a 95% chance that the age is in the interval 0.0034–0.566 (in units of $2N$ generations) and a 99% chance that the age is in the interval 0.0018–1.33. The relatively wide and asymmetric confidence interval reflects the fact that a neutral allele might be in low frequency because it arose recently or because it arose much earlier and was in decline from a previous high frequency.

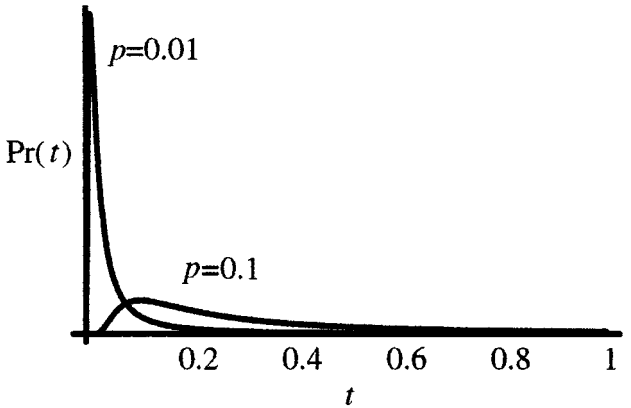

Figure 4 Graphs of the approximate distribution of allele ages in a population of constant size. These curves are obtained from Equation 6 in the text, by assuming a sample of $n = 1000$ chromosomes. Time is measured in units of $2N$ generations.

Past Population Growth

The sizes of most human populations have not been constant but instead have undergone various changes. There is no formula for the expected allele age, corresponding to Equation 1, for a population of variable size, but it is easy to find the approximate distribution of allele ages by using the Griffiths & Tavaré (12) theory. Variation in population size is equivalent to an expansion or contraction of the time scale, because genetic drift is stronger in smaller populations and weaker in larger ones. A scaled time $\tau(t)$ represents the net effect of changes in population size *N(t)*, at time *t* generations in the past:

$$\tau(t) = \int_0^t \frac{dt'}{2N(t')}.$$

The key result is that the distribution of allele ages in a population of variable size is the same as that in a population of constant size, provided that $\tau(t)$ replaces t (12).

To illustrate the effect of population growth, assume that the population has been growing exponentially at rate r in the past, $N(t) = N_0 e^{-rt}$ and, $\tau(t) = (e^{rt} - 1)/(2N_0 r)$, where N_0 is the current effective size. When time is measured in units of N_0 generations, the distribution of ages depends on p and the composite parameter $R = 2N_0 r$. Many studies are concerned with the ages of alleles found in populations in western Europe that have grown during the past 10,000 years, since the introduction of agriculture, and have grown very rapidly during the past 500 years. If we assume a current population size of 300 million, N_0 is ~100 million (16). If such a population had been randomly mating during its period of growth, even a small value of r would imply an enormous value of R. For example, if we assume continuous exponential growth from an effective population size of 10,000

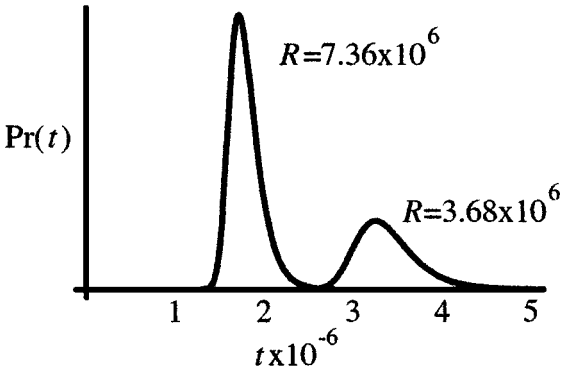

Figure 5 Graphs of the approximate distribution of allele ages in a growing population for an allele found at 2% frequency. Time is measured in units of $2N_0$ generations, where N_0 is the current effective population size. $R = 2N_0r$, where r is the population growth rate. The values of R shown correspond to an effective size of 10,000 and exponential growth rates of 0.0184 and 0.0368, as discussed in the text.

individuals ∼500 generations (10,000 years) ago, that implies that $r = 0.0184$ and $R = 3.68 \times 10^6$.

We can illustrate the importance of population growth by considering an allele at a frequency of 0.02, roughly that of $\Delta F508$ in Europeans. The distributions of ages for two growth rates are shown in Figure 5. The MLE of age is $\sim 3.23 \times 10^{-6}$ in units of N_0 generations or ∼646 generations. If we double the value of R, to reflect the more recent rapid growth, the MLE age decreases by almost a factor of 2, to ∼342 generations. When R is large, the confidence interval tends to be small. In this example, the 95% support interval is 564–814 for the lower growth rate and 300–530 for the larger growth rate, all in units of generations.

Obviously, European and other human populations are not randomly mating, but there has been enough gene flow that there is very little genetic differentiation among Europeans (5). In general, geographically restricted dispersal results in an effective population size that is larger than the number of reproductively active individuals (30), so, assuming a current effective size of $N_0 = 10^8$ may be conservative, but there is no available theory indicating the effect of past and current population subdivision on the distribution of allele ages.

Natural Selection

The relationship between allele age and allele frequency can also be found if selection affects an allele, although the problem has received less study (23, 26, 53). If selection is additive, meaning that the fitness of the heterozygote is intermediate between the fitnesses of normal and mutant homozygotes, the distribution of allele ages is the same for advantageous and deleterious alleles (26). Of course, an

advantageous allele is much more likely than a deleterious allele to be present at all, but, given the frequency, the distribution of their ages is the same. With dominance, this result no longer holds, but it is close to being true for a low-frequency allele, if there is no overdominance in fitness, because the frequency of rare alleles depends largely on the ratio of the fitnesses of the normal homozygote and the heterozygote.

Numerical results presented by Maruyama (26) show that additive selection reduces the average age. If s is the selection coefficient for or against the mutant and N is the effective population size, the mean age depends on the product Ns. The reduction in average age is substantial if Ns is much greater than 1, but, even with relatively strong selection, the effect is not dramatic (see Table 3 in reference 26).

Li (23) considered different degrees of dominance, including overdominance. A recessive or nearly recessive allele has a higher average age than a comparable allele that has an additive effect on fitness. Even slight overdominance results in a much larger average allele age. Later work by Takahata (47) and others shows that strong overdominance, of the kind found in the MHC system in mammals and the genetic self-incompatibility systems in plants, leads to the very great persistence of alleles, thus accounting for observations of *trans*-specific polymorphisms (the sharing of alleles between species) found in such systems. To infer the ages of overdominant alleles, frequency contains little information; a very broad range of ages is consistent with almost any allele frequency.

There is as yet no theory predicting the average age or the distribution of ages of a selected allele in a growing population. For deleterious or advantageous alleles, results for neutral alleles found in the same frequency provide conservative upper bounds that are probably not too conservative unless selection is very strong.

Branching Process Approximation

For rare alleles, a good approximation of the allele frequency dynamics can be obtained using a branching process model (9). Maruyama (25) used a diffusion approximation for a branching process to study the average age of an allele that was present in a population with a given number of copies. Thompson (49) applied a branching-process model to derive an MLE of the age of a rare allele, given the number of total copies in a population. Saitou et al (40) considered the same problem when the population growth rates varied. They showed that such fluctuations tended to reduce the estimated age. Our birth-death model (45) is equivalent to that of Thompson, but with overlapping rather than discrete generations and with only a fraction of the population sampled. To obtain a joint likelihood, we calculated a likelihood of age as a function of the number of copies in a sample and combined that with the likelihood of age as a function of the intra-allelic variability.

Estimators of allele age that are derived from branching or birth-death process models have the advantages of not requiring assumptions about effective population size and of allowing for selection for or against an allele. They also may

be less sensitive to some kinds of population subdivision, because they assume independent reproduction of each copy of the allele.

COMBINING AND COMPARING ESTIMATES

Given the allele frequency and assumptions about selection and population growth, the theory in the preceding section can be used to obtain the expected age or the distribution of ages. Intra-allelic variability provides another estimate that is conditionally independent of allele frequency. Estimates of allele age and confidence intervals that are obtained from these two kinds of data can be compared informally. If the confidence intervals broadly overlap, that would tend to increase our confidence in the estimate. It should also increase our confidence that the population genetic model provides an adequate description of the dynamics of the allele since it arose.

To illustrate, Risch et al (39) used intra-allelic variability to estimate that the age of an allele causing ITD in Ashkenazi Jews is 8–22 generations. The frequency of this allele is 1/6000, and the Ashkenazi population has grown very rapidly during the past 600 years. The approximate theory described in the previous section provides a 95% confidence interval of 13.9–26.8 generations for this frequency, assuming that $r = 0.4055$ and $N_0 = 5 \times 10^6$. This confidence interval is consistent with the age that is estimated from the intra-allelic variability and supports the assumption that the allele has little effect on fitness, despite the severe effect it has on individuals afflicted with ITD.

Information from frequency and intra-allelic variability can be combined in a formal way by adopting a Bayesian perspective. Shoemaker et al (42) provided a practical discussion of Bayesian methods for geneticists. The probability distribution of allele age based on frequency can be treated as a prior distribution that is then multiplied by likelihood of age based on the intra-allelic variability to obtain a posterior distribution of age that takes account of both sources of information. This approach treats age as a random variable, rather than a parameter. In algebraic terms, the posterior distribution is obtained from

$$P(t|G) \propto P(G|t)P(t),$$

where P indicates probability, t is the age, and G represents the observed intra-allelic variability (i.e. the data). $P(G|t)$ is the probability of the data given the age (i.e. the likelihood, which can be calculated by one of the methods described previously), and $P(t)$ is the prior distribution of ages based on the observed frequency. The constant of proportionality is chosen so that $P(t|G)$ sums to one.

The formal theory of prior and posterior distributions is daunting and probably not of great interest to someone looking for the best way to analyze data, but the underlying issues are important and worth understanding to guide data analysis and the interpretation of other studies. The role of a prior distribution can be seen

in considering a data set in which no recombinants are found between the defining mutation and a linked marker. The moment and likelihood estimates are 0, which is not usually a satisfactory estimate. In practice, when such markers are found, they are ignored, and instead markers are used for which there are some recombinants. Assuming a prior distribution of ages eliminates the problem created when no recombinants are found. The frequency implies a prior distribution of age, and the observation of no recombinant chromosomes results in a posterior distribution that is shifted somewhat to the left, indicating a younger age than one based on frequency alone. More extensive intra-allelic variability could shift the distribution to the right, indicating an older age than expected from the frequency.

Some statisticians argue that treating age as a random variable and using a prior distribution of ages based on frequency are necessary because the frequency is assumed to follow a population genetic model. Hence the distribution of possible ages is constrained by that model. Wiuf & Donnelly (56), for example, say that ignoring the prior distribution and treating age as a parameter is " . . . inappropriate as a matter of statistical principle." Others do not agree with this view. Guo & Xiong (14), for example, say that treating age as a parameter "From a data-analytical viewpoint . . . seems more natural and appealing." It may be statistically correct to use a prior distribution, but it is useful only if the appropriate prior distribution is known. Treating age as a parameter requires assumptions about the history of the allele only between the time it arose and the present. Treating age as a random variable assumes something about the history of the locus even before the allele arose by mutation.

The choice of a prior distribution is especially important for estimating allele age, because the prior distribution that is based on frequency tends to dominate the resulting estimate; that is, taking account of intra-allelic variability will shift the estimated age relatively little from the estimate that is based on frequency, because a rather large range of intra-allelic genealogies is consistent with an observed low allele frequency.

EXAMPLES

Δ*F508* in Western Europe

Δ*F508* has been extensively studied because its frequency is $\sim$2% in European populations, yet it causes what was until very recently a recessive lethal condition, cystic fibrosis. Δ*F508* is in very low frequency in other populations, so its geographic restriction to Europeans suggests a recent origin and, possibly, a selective advantage in heterozygotes. Serre et al (41) estimated the age of this allele to be 3000–6000 years. They applied the moment estimator to data from two marker sites and used several values of the recombination rates with two linked marker loci to determine the confidence interval.

Morral et al (29) surveyed intra-allelic variability at three microsatellite marker loci, two in intron 17 and one in intron 8 of *CFTR*, and they obtained quite a

different estimate—a minimum age of 52,000 years. Morral et al (29) assumed that the markers were sufficiently closely linked that almost all differences from the ancestral haplotype arose by mutation. Kaplan et al (18) criticized this estimate and pointed out that Morral et al (29) did not allow for any genealogical relationship between chromosomes with identical or similar haplotypes, thereby greatly increasing the time needed for the accumulation of all of the variant chromosomes found. Kaplan et al argued that if the genealogical structure were accounted for, the estimated age would also be lower and consistent with that of Serre et al (41).

We reanalyzed the data of Morral et al (29), using a method in which the intra-allelic genealogy was generated by a birth-death process (45). We computed the joint likelihood based on frequency and intra-allelic variability and obtained an MLE of t_1 of 146 generations or $\sim$3000 years, with a 95% confidence interval of 116–178 generations. Our estimate is consistent with that of Serre et al (41) and quite different from that of Morral et al (29).

Our method allowed us to test for selection in favor of heterozygotes carrying $\Delta F508$. We did so by assuming a higher rate of population growth for heterozygous individuals. With a 1.5% growth advantage, the estimated age of $\Delta F508$ decreased substantially, to only 80 generations or $\sim$1600 years (45), which is much lower than the estimate of Serre et al (41) and seems inconsistent with the current geographic distribution of $\Delta F508$. If heterozygotes carrying $\Delta F508$ had a significant selective advantage, the intra-allelic genealogy would be compressed towards the root, thus leaving more time for mutations to accumulate at the marker loci. Our tentative conclusion was that the data of Morral et al (29) are not consistent with the hypothesis of sustained selection in favor of heterozygous carriers of $\Delta F508$.

The frequency of $\Delta F508$ also leads to an estimated age. The graphs in Figure 5, which assume neutrality, give estimated ages that are consistent with those based on intra-allelic variability. It may be that heterozygous carriers of $\Delta F508$ have a selective advantage, but the effect is not evident from considerations of allele age.

Idiopathic Torsion Dystonia in Ashkenazi Jews

As part of an effort to map *DYT1*, a locus on chromosome 9q causing ITD in Ashkenazi Jews, Risch et al (39) examined several microsatellite and RFLP markers in the candidate region. They found extensive linkage disequilibrium on disease-associated chromosomes in a region of at least 4 cmorgans. The moment estimator applied to several marker loci suggested that the allele arose 12–13 generations ago, with a confidence interval of 8–22 generations. In numerous studies of other disease alleles, haplotypes spanning relatively large chromosomal regions have been found (1, 8, 10, 21, 28, 35, 36, 46, 51, 57), also indicating young ages for alleles. Risch et al (39) argued that such a young age for the allele causing ITD in Ashkenazi Jews is not consistent with its frequency, 1/6000, a relatively high frequency for a dominant allele with 30% penetrance and causing such a serious

disorder. They argued that their observations imply that the current Ashkenazi population is derived from a relatively small number of founders, probably representing the more affluent families in the fifteenth and sixteenth centuries.

Our reanalysis (37) of the data of Risch et al (39) illustrates the possible importance of assuming a prior distribution of ages based on allele frequency when analyzing intra-allelic variability. We found the MLE of age, based on intra-allelic variability at the *ASS* locus, to be 17.8 generations, slightly larger than but comparable with the estimate obtained by Risch et al from the moment estimator, 8.4 generations. But we also found that the upper limit of the 95% confidence interval for t_1 is 132 generations (37), much larger than that estimated by Risch et al. We obtained such a large upper bound because the distribution of intra-allelic coalescence times is only weakly dependent on t_1 in a rapidly growing population (see Figure 3), so even very large ages are consistent with the relatively few recombinant haplotypes found (7 of 54 in this case). If a prior distribution based on frequency is assumed, a much smaller upper bound to the confidence interval is found. For example, applying the method based on Equation 6, the 95% support interval for the age of a neutral allele in frequency $1/6000$ in a population growing at a rate of 0.4055 per generation is $\sim$14–27 generations, and 27 generations is roughly the upper bound obtained from the posterior distribution as well.

CCR5-Δ32 AIDS-Resistance Allele

Dean et al (6) showed that a 32-bp deletion in the *CCR5* locus was associated with resistance to infection by HIV and the onset of AIDS. Individuals who are homozygous for the deletion are nearly completely resistant to infection by HIV-1, and heterozygous individuals have a delayed onset of AIDS. This deletion is in a frequency of $>$10% in Caucasian populations, with higher frequencies in the north. It is absent from east Asians and Native Americans (46). Stephens et al (46) found strong disequilibrium between two flanking microsatellite markers that were separated by a recombination distance of 1.1 centimorgans. Of the 46 chromosomes, 39 had the same two-locus haplotype. From these data, Stephens et al (46) used the moment estimator to obtain an age of 27.5 generations or $\sim$688 years.

Stephens et al (46) considered the variation in this estimate that is caused by uncertainty in the recombination rates, which were estimated from a regression analysis of a linkage map against a radiation hybrid map. The lowest estimated rate that was consistent with their data increased the estimated age to 82.5 generations. They also estimated the uncertainty arising from variation in the intra-allelic genealogy by applying the method of Reich & Goldstein (38), discussed earlier. They assumed that the intra-allelic genealogy was the same as that for a random sample of chromosomes with the same sample size. Given that assumption and a range of values of the population growth rate, they found the distribution of ages to be consistent with the number of conserved ancestral haplotypes. Their 95% confidence interval was 11–75 generations, showing that, in this study, the two sources of uncertainty are comparable in importance.

The study by Stephens et al is one of the few also to consider the allele frequency. The Kimura-Ohta formula (Equation 4) for a neutral allele in a frequency of 10% in a population of effective size 5000 yields an average allele age of 6500 generations, much larger than the estimates based on intra-allelic variability. That age would imply that the allele originated >100,000 years ago, making it very difficult to explain its current geographic range. Stephens et al (46) used the difference in the two estimates of age to argue that the deletion had been strongly selected in the recent past, and they estimated a selection coefficient, in favor of the deletion, of ~30%. Their estimate of age based on frequency did not take account of population growth, but even with past exponential growth, the estimated age in the absence of selection would be much older than implied by the intra-allelic variability.

BRCA1 and *BRCA2*

Numerous alleles at *BRCA1* and *BRCA2* are associated with early-onset breast cancer. Unlike at *CFTR*, no one allele at either locus is predominant in most populations. Neuhausen et al (32) assayed the haplotypes at nine microsatellite markers that are closely linked to *BRCA1* in 61 families that carried one of the six most common alleles associated with elevated risk of breast cancer. They derived a moment estimator of allele age that accounts for both mutation and recombination and used average mutation rates for dinucleotide and tetranucleotide repeat loci, reported by Weber & Wong (55).

Although Neuhausen et al (32) described their estimate as an MLE, it is based on computing a composite likelihood by multiplying marginal likelihoods across loci and across chromosomes. Estimates were obtained for five of the six alleles, and all have relatively small estimated ages, ranging from 9 to 170 generations.

It is notable that the five estimated ages are so small, particularly because they are estimates for the most common, although still relatively rare, alleles. In general, more common alleles tend to be the oldest (54), implying that the rarer disease-associated alleles at *BRCA1* are younger still. Relatively strong selection is required to reduce the expected age of mutations to well below the age expected under neutrality. It is impossible to know the selection affecting each of these alleles. Neuhausen et al (32) found no allele-specific effects, but selection attributable to breast cancer would be weak in any case. These alleles are only partially penetrant, and, although the breast and ovarian cancers caused by them have an early onset from a clinical perspective, the onset is late in the fecundity schedule, which implies that the effect on reproductive success is small. If selection is strong, it is probably attributable to pleiotropic effects.

For one of the alleles, *185delAG*, which has a 2-bp deletion in the twenty third codon, other data call into question the age estimated by Neuhausen et al. This allele is relatively common in Ashkenazi Jews and found in a frequency of 1%. Neuhausen et al (32) estimated that the age of this allele in Ashkenazi Jewish families is 46 generations, dating to ~1235 A.D., with an upper limit of the support interval of 80 generations. Yet Bar-Sade et al (3) surveyed a population

of Iraqi Jews and found three copies of the same allele, two of which shared a common haplotype at three linked marker loci and one of which had a haplotype that differed at only one of the three markers. Bar-Sade et al concluded that this allele was present in the Jewish population before the dispersion in 70 A.D., a date slightly older than the upper limit for the age estimated by Neuhausen et al (32). Bar-Sade and coworkers (2) surveyed other Jewish populations for this allele and found the same haplotype to be widespread. The presence in other Jewish populations suggests that this allele was in appreciable frequency well before 70 A.D.

One explanation for this discrepancy is that Ashkenazi Jews suffered a severe reduction in population size that might have reduced the number of lineages carrying this allele sufficiently that descendants of only one lineage have survived. The rapid growth of Ashkenazi Jews in the past 500 years (39) would permit a later accumulation of intra-allelic variability. The moment estimator does not estimate the time of occurrence of the mutation t_1, but rather the age of the most recent common ancestor t_2.

Neuhausen et al (31) carried out a similar study of nine alleles at *BRCA2*, by using the same statistical method. They were able to estimate ages for five alleles and found relatively young ages for all of them—even younger on average than the ages for *BRCA1*. One of the alleles, *6174delT*, is found in ~1% of individuals of Ashkenazi Jewish ancestry. Its estimated age is ~29 generations, even smaller than for *185delAG* at *BRCA1*. There are no surveys of this allele in other Jewish populations.

The two studies by Neuhausen et al (31, 32) of *BRCA1* and *BRCA2* are unusual in estimating ages of several alleles at each locus. The young ages found for alleles at both loci are surprising. The two alleles in relatively high frequency in Ashkenazi Jews could be young because of a past bottleneck in population size, but the other alleles at both loci are found in other European populations that have quite different demographic histories. Although no formal analysis has been done, it is difficult to account for these observations without assuming relatively strong selection.

Factor XI

Goldstein et al (11) estimated the ages of two alleles at the *FXI* locus that cause deficiency in coagulation factor XI. One allele, the type III mutation, is found only in Ashkenazi Jews, and the other, the type II mutation, is found in both Ashkenazi and Iraqi Jews. The type III mutation occurs at a 2.54% frequency in Ashkenazi Jews, and the type II mutation occurs at a frequency of 2.17% in Ashkenazi Jews and 1.67% in Iraqi Jews. The estimated age of the type III mutation is 31 generations, consistent with its restriction to Ashkenazi Jews, and the estimated age of the type II mutation is >120 generations, consistent with its presence in both populations. The type II mutation is similar to the *185delAG* mutation of *BRCA1* in being found in both Ashkenazi and Iraqi Jews. Goldstein et al (11) noted that, unlike the allele causing ITD, the type II mutation was not affected by the fluctuations in population size of the Ashkenazi Jewish populations. That conclusion is consistent with the

much higher frequency of the type II mutation (2.17%) than that of the mutation at ITD (0.017%).

CONCLUSIONS

Intra-allelic genetic variation and allele frequency can both be used to estimate allele age. Considering intra-allelic variability alone, the moment estimator, Equation 2, or its generalization to two or more linked marker loci leads to a reasonable estimate of age but one that is biased downwards. The confidence interval of this estimator depends on both the uncertainty in genetic parameters and the uncertainty about the intra-allelic genealogy. Assuming a star genealogy for the intra-allelic genealogy gives too narrow a confidence interval. That tendency, combined with the downward bias in the moment estimator, leads to the belief that alleles are younger than they are, possibly much younger. Methods that make use of likelihood avoid this problem, but they are more difficult to apply and, for more than one marker locus, are incompletely developed. If the goal is to show that a particular allele is relatively young, almost any method will serve, provided that large shared haplotypes are found in the data set. The actual estimate of age may not matter. But if the goal is to draw more specific conclusions and relate the age of an allele to other historical data, then it is appropriate to be cautious in the analysis and use likelihood methods that are or will become available.

Taking account of allele frequency is relatively easy and always worthwhile. In many cases, an estimate of age assuming no selection will reinforce conclusions based on the analysis of intra-allelic variability and generally narrow the confidence interval. In other cases, the discrepancy in estimated ages is large enough that selection or some other factor is seen to be important. But relying too much on allele frequency and not allowing for uncertainty in the underlying population genetics model may lead to erroneous conclusions. Estimates based on intra-allelic variability and allele frequency have a different character because intra-allelic variability reflects what has actually happened to an allele, whereas allele frequency reflects what a population geneticist thinks has happened to that allele.

More theory is needed in this area, particularly to analyze data from several marker loci and to allow for the effects of population subdivision and other demographic complications. As more studies are carried out, estimates for different alleles and different loci can together be used to draw inferences about the past history of populations.

ACKNOWLEDGMENTS

This research was supported in part by grant GM40282 from the National Institutes of Health (NIH) and by a John Simon Guggenheim Memorial Fellowship to M. S. and by grant HG01988 from NIH to B. R. We thank D. B. Goldstein and S. Tavaré for helpful discussions of this topic and D. B. Goldstein and N. Saitou for providing copies of their work.

Visit the Annual Reviews home page at www.AnnualReviews.org

LITERATURE CITED

1. Ajioka RS, Jorde LB, Gruen JR, Yu P, Dimitrova D, et al. 1997. Haplotype analysis of hemochromatosis: evaluation of different linkage-disequilibrium approaches and evolution of disease chromosomes. *Am. J. Hum. Genet.* 60:1439–47
2. Bar-Sade RB, Kruglikova A, Modan B, Gak E, Hirsh-Yechezkel G, et al. 1998. The 185delAG BRCA1 mutation originated before the dispersion of Jews in the Diaspora and is not limited to Ashkenazim. *Hum. Mol. Genet.* 7:801–5
3. Bar-Sade RB, Theodor L, Gak E, Kruglikova A, Hirsch-Yechezkel G, et al. 1997. Could the 185delAG BRCA1 mutation be an ancient Jewish mutation? *Eur. J. Hum. Genet.* 5:413–16
4. Bertorelle G, Rannala B. 1998. Using rare mutations to estimate population divergence times: a maximum likelihood approach. *Proc. Natl. Acad. Sci. USA* 95: 15452–57
5. Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. *The History and Geography of Human Genes.* Princeton, NJ: Princeton Univ. Press. 382 pp.
6. Dean M, Carrington M, Winkler C, Huttley GA, Smith MW, et al. 1996. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. *Science* 273:1856–62
7. Devlin B, Risch N, Roeder K. 1996. Disequilibrium mapping: composite likelihood for pairwise disequilibrium. *Genomics* 36:1–16
8. Díaz A, Montfort M, Cormand B, Zeng B, Pastores GM, et al. 1999. Gaucher disease: the N370S mutation in Ashkenazi Jewish and Spanish patients has a common origin and arose several thousand years ago. *Am. J. Hum. Genet.* 64:1233–38
9. Fisher RA. 1922. On the dominance ratio. *Proc. R. Soc. Edinburgh* 42:321– 41
10. Forestier L, Jean G, Attard M, Cherqui S, Lewis C, et al. 1999. Molecular characterization of *CTNS* deletions in nephropathic cystinosis: development of a PCR-based detection assay. *Am. J. Hum. Genet.* 65:353–59
11. Goldstein DB, Reich DE, Bradman N, Usher S, Seligsohn U, et al. 1999. Age estimates of two common mutations causing factor XI deficiency: recent genetic drift is not necessary for elevated disease incidence among Ashkenazi Jews. *Am. J. Hum. Genet.* 64:1071–75
12. Griffiths RC, Tavaré S. 1998. The age of a mutation in a general coalescent tree. *Stoch. Models* 14:273–95
13. Griffiths RC, Tavaré S. 1999. The ages of mutations in gene trees. *Adv. Appl. Prob.* 9:567–90
14. Guo SW, Xiong M. 1997. Estimating the age of mutant disease alleles based on linkage disequilibrium. *Hum. Hered.* 47:315–37
15. Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, et al. 1992. Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nat. Genet.* 2:204–11
16. Hill WG. 1972. Effective size of populations with overlapping generations. *Theor. Popul. Biol.* 3:278–89
17. Hudson RR. 1990. Gene genealogies and the coalescent process. *Oxford Surv. Evol. Biol.* 7:1–44
18. Kaplan NL, Lewis PO, Weir BS. 1994. Age of the ΔF508 cystic fibrosis mutation. *Nat. Genet.* 8:216–18
19. Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, et al. 1989. Identification of the cystic fibrosis gene: genetic analysis. *Science* 245:1073–80

20. Kimura M, Ohta T. 1973. The age of a neutral mutant persisting in a finite population. *Genetics* 75:199–212
21. Labuda M, Labuda D, Korab-Laskowska M, Cole DE, Zietkiewicz E, et al. 1996. Linkage disequilibrium analysis in young populations: pseudo-vitamin D-deficiency rickets and the founder effect in French Canadians. *Am. J. Hum. Genet.* 59:633–43
22. Lander ES, Botstein D. 1986. Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. *Cold Spring Harbor Symp. Quant. Biol.* 51(1):49–62
23. Li W-H. 1975. The first arrival time and mean age of a deleterious mutant gene in a finite population. *Am. J. Hum. Genet.* 27:274–86
24. Luria SE, Delbrück M. 1943. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28:491–511
25. Maruyama T. 1974. The age of a rare mutant gene in a large population. *Am. J. Hum. Genet.* 26:669–73
26. Maruyama T. 1974. The age of an allele in a finite population. *Genet. Res., Camb.* 23:137–43
27. McPeek MS, Strahs A. 1999. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am. J. Hum. Genet.* 65:858–75
28. Moisio AL, Sistonen P, Weissenbach J, de la Chapelle A, Peltomäki P. 1996. Age and origin of two common MLH1 mutations predisposing to hereditary colon cancer. *Am. J. Hum. Genet.* 59:1243–51
29. Morral N, Bertranpetit J, Estivill X, Nunes V, Casals T, et al. 1994. The origin of the major cystic fibrosis mutation (ΔF508) in European populations. *Nat. Genet.* 7:169–75
30. Nei M, Takahata N. 1993. Effective population size, genetic diversity, and coalescence time in subdivided populations. *J. Mol. Evol.* 37:240–44
31. Neuhausen SL, Godwin AK, Gershoni-Baruch R, Schubert E, Garber J, et al. 1998. Haplotype and phenotype analysis of nine recurrent *BRCA2* mutations in 111 families: results of an international study. *Am. J. Hum. Genet.* 62:1381–88
32. Neuhausen SL, Mazoyer S, Friedman L, Stratton M, Offit K, et al. 1996. Haplotype and phenotype analysis of six recurrent BRCA1 mutations in 61 families: results of an international study. *Am. J. Hum. Genet.* 58:271–80
33. Nielsen R, Weinreich DM. 1999. The age of nonsynonymous and synonymous mutations in animal mtDNA and implications for the mildly deleterious theory. *Genetics* 153:497–506
34. Pagnier J, Mears JG, Dunda-Belkhodja O, Schaefer-Rego KE, Beldjord C, et al. 1984. Evidence for the multicentric origin of the sickle cell hemoglobin gene in Africa. *Proc. Natl. Acad. Sci. USA* 81:1771–73
35. Piccolo F, Jeanpierre M, Leturcq F, Dodé C, Azibi K, et al. 1996. A founder mutation in the γ-sarcoglycan gene of Gypsies possibly predating their migration out of India. *Hum. Mol. Genet.* 5:2019–22
36. Pras E, Pras E, Kreiss Y, Frishberg Y, Prosen L, et al. 1999. Refined mapping of the CSNU3 gene to a 1.8-Mb region on chromosome 19q13.1 using historical recombinants in Libyan Jewish cystinuria patients. *Genomics* 60:248–50
37. Rannala B, Slatkin M. 1998. Likelihood analysis of disequilibrium mapping, and related problems. *Am. J. Hum. Genet.* 62:459–73
38. Reich DE, Goldstein DB. 1999. Estimating the age of mutations using variation at linked markers. In *Microsatellies: Evolution and Applications*, ed. DB Goldstein, C Schlötterer, pp. 129–38. Oxford, UK: Oxford Univ. Press
39. Risch N, de Leon D, Ozelius L, Kramer P, Almasy L, et al. 1995. Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small

founder population. *Nat. Genet.* 9:152–59
40. Saitou N, Shimizu H, Omoto K. 1988. On the effect of the fluctuating population size on the age of a mutant allele. *J. Anthropol. Soc. Nippon* 96:449–58
41. Serre JL, Simon-Bouy B, Mornet E, Jaume-Roig B, Balassopoulou A, et al. 1990. Studies of RFLP closely linked to the cystic fibrosis locus throughout Europe lead to new considerations in populations genetics. *Hum. Genet.* 84:449–54
42. Shoemaker JS, Painter IS, Weir BS. 1999. Bayesian statistics in genetics: a guide for the uninitiated. *Trends Genet.* 15:354–58
43. Slatkin M. 1996. Gene genealogies within mutant allelic classes. *Genetics* 143:579–87
44. Slatkin M, Hudson RR. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129:555–62
45. Slatkin M, Rannala B. 1997. Estimating the age of alleles by use of intraallelic variability. *Am. J. Hum. Genet.* 60:447–58
46. Stephens JC, Reich DE, Goldstein DB, Shin HD, Smith MW, et al. 1998. Dating the origin of the CCR5-Δ32 AIDS-resistance allele by the coalescence of haplotypes. *Am. J. Hum. Genet.* 62:1507–15
47. Takahata N. 1990. A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. *Proc. Natl. Acad. Sci. USA* 87: 2419–23
48. Terwilliger JD. 1995. A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am. J. Hum. Genet.* 56:777–87
49. Thompson EA. 1976. Estimation of age and rate of increase of rare variants. *Am. J. Hum. Genet.* 28:442–52
50. Thomson R. 1998. Ages of mutations on a coalescent tree. *Math. Biosci.* 153:41–61
51. Varilo T, Savukoski M, Norio R, Santavuori P, Peltonen L, et al. 1996. The age of human mutation: genealogical and linkage disequilibrium analysis of the CLN5 mutation in the Finnish population. *Am. J. Hum. Genet.* 58:506–12
52. Watterson GA. 1976. Reversibility and the age of an allele. I. Moran's infinitely many neutral alleles model. *Theor. Popul. Biol.* 10:239–53
53. Watterson GA. 1977. Reversibility and the age of an allele. II. Two-allele models, with selection and mutation. *Theor. Popul. Biol.* 12:179–96
54. Watterson GA, Guess HA. 1977. Is the most frequent allele the oldest? *Theor. Popul. Biol.* 11:141–60
55. Weber JL, Wong C. 1993. Mutation of human short tandem repeats. *Hum. Mol. Genet.* 2:1123–28
56. Wiuf C, Donnelly P. 2000. Conditional genealogies and the age of a neutral mutant. *Theor. Popul. Biol.* 56:183–201
57. Zühlke C, Gehlken U, Purmann S, Kunisch M, Müller-Myhsok B, et al. 1999. Linkage disequilibrium and haplotype analysis in German Friedreich ataxia families. *Hum. Hered.* 49:90–96