THE ROYAL SOCIETY

doi 10.1098/rstb.2000.0729

Allele age and a test for selection on rare alleles

Montgomery Slatkin

Department of Integrative Biology, University of California, Berkeley, CA 94720-3140, USA (*slatkin@socrates.berkeley.edu*)

An approximate expression for the probability distribution of the age of a neutral allele as a function of its frequency is derived for a population undergoing arbitrary changes in population size. A simple maximum-likelihood estimator of allele age based on frequency is also obtained. The distribution of allele age, combined with a model predicting the extent of intra-allelic variability generated by mutation and recombination, leads to a statistical test of whether a rare allele has experienced natural selection. The test is based on finding whether there is too little or too much intra-allelic variability to be consistent with the observed frequency. The test is applied to the locus, BRCA1, associated with early-onset breast cancer in humans and shows that two common disease-associated alleles (5382*ins*C and 185*del*AG) appear to have been subject to natural selection.

Keywords: coalescent theory; neutrality test; BRCA1; population genetics

1. INTRODUCTION

The age of an allele is the time since it arose by mutation. In this context, an allele is defined by the possession of a particular alteration in DNA sequence—a substitution, insertion or deletion. For example, the defining mutation of the ΔF508 allele at the CFTR locus associated with cystic fibrosis is the deletion of the three nucleotides for the 508th codon. With this definition of allele age, different copies of an allele do not have to be otherwise identical in sequence. Morral *et al.* (1994), for example, found that chromosomes carrying ΔF508 differed at three microsatellite loci in two introns of CFTR. Variation among different copies of the same allele, which I will call intra-allelic variability, can provide useful information about the history of the allele.

Population geneticists have been concerned with the relationship between allele age and allele frequency since Kimura & Ohta (1973) showed that the expected age in generations of an allele at frequency x in a population containing N diploid individuals is approximately

$$E(a) = \frac{-4Nx}{1-x}\ln(x), \tag{1}$$

where E denotes expectation and ln denotes the natural logarithm. Their paper led to several others. Maruyama (1974) and Li (1975) showed that most kinds of natural selection reduce average allele age from the value in equation (1) but that overdominance in fitness greatly increases average age. Watterson (1976) unified much of the theory and provided a way to find numerically the probability distribution of allele age as a function of frequency.

The conclusion from the theory of allele age as developed in the 1970s was that, although allele age in a population of constant size is to some extent indicated by allele frequency, the distribution of ages consistent with a particular frequency is sufficiently broad that frequency alone does not provide a very good estimate of age. Although the theory continued to be developed, there was little application of it and little practical interest in allele age. In the 1990s, human geneticists became interested in estimating the ages of alleles associated with human genetic diseases. The extent of intra-allelic variability, rather than allele frequency, was used to estimate allele age, and the relationship between allele age and allele frequency was largely ignored. The first such study was by Serre *et al.* (1990) who used frequencies at two polymorphic restriction sites closely linked to ΔF508 to estimate its age to be between 3000 and 6000 years. A later study by Risch *et al.* (1995) estimated the age of an allele causing idiopathic torsion dystonia (ITD) in Ashkenazi Jews to be between eight and 22 generations. The Risch *et al.* (1995) paper motivated several others who also found relatively young ages for disease-associated alleles (Slatkin & Rannala 2000).

Stephens *et al.* (1998) considered both intra-allelic variability and allele frequency in their study of the 32 base pair (bp) deletion in the CCR5 locus that confers resistance to infection by human immunodeficiency virus. The presence of a nearly conserved haplotype at two linked microsatellite loci separated by *ca.* 1 centimorgan (cM) suggested that the deletion occurred about 700 years ago. Yet this deletion is at a frequency greater than 10% in many European populations, and absent or nearly absent from other populations. Stephens *et al.* (1998) noted that equation (1) implies an average age of greater than 100 000 years, even assuming an effective population size of only 5000 individuals. They argued that the difference in the estimated ages implies that the deletion was positively selected in Europeans and they estimated the selection coefficient to be *ca.* 30%.

Stephens *et al.* (1998) did not use a statistical test of selection because the pattern of intra-allelic variability was so obviously inconsistent with the high allele frequency that a test was unnecessary. Their logic is similar to a formal test of neutrality developed by Hudson *et al.* (1994) and called the haplotype test. Hudson *et al.* (1994) sequenced 1410 bp in 41 copies of the superoxide dismutase (*Sod*) locus in *Drosophila melanogaster*. In this species, two alleles, designated slow and fast, are distinguished by a single difference in amino-acid sequence. Hudson *et al.* (1994) found no variation among slow alleles, which are found at *ca.* 18% frequency, and little variation among one subgroup of the fast alleles that differed from the slow alleles by only one nucleotide substitution. The haplotype test is based on finding the probability of observing a subset of haplotypes with such a low level of polymorphism. A low probability, as was found for these data, indicates a significant deviation from neutrality. The haplotype test is not based on considerations of allele age but the low level of intra-allelic variability in a subset indicates a much younger age for that subset than is consistent with its frequency.

The haplotype test has been refined by others. Andolfatto *et al.* (1999) provided a way to correct for multiple tests in the same data set, in order to allow for the possibility that several different subsets of haplotypes are considered. Depaulis & Veuille (1998) suggested using both haplotype number and haplotype diversity as tests for too little or too much variation at a locus.

The various versions of the haplotype test are designed to test for an overall reduction in genetic variability, as would be expected if an advantageous allele has swept through a population (Maynard Smith & Haigh 1974). A significant deviation from neutrality does not mean that the particular subset of haplotypes identified were actually subject to selection. It implies only that selection somewhere affected the pattern of variability in the region sequenced. The test of selection described later is a formalization of the argument used by Stephens *et al.* (1998) and differs from the haplotype test because it tests for a reduction in intra-allelic variability attributable to selection affecting the allele of interest. The test described in this paper is appropriate only for low-frequency alleles and intra-allelic variability at a linked biallelic marker locus. Slatkin & Bertorelle (2000) develop a general test applicable to alleles in arbitrary frequency and to other kinds of intra-allelic variability.

2. METHODS

(a) *Griffiths–Tavaré theory of allele age*

The basis for the approximate theory developed here is the analysis of allele age by Griffiths & Tavaré (1998). They considered a coalescent model of a neutral locus and assumed that, in a sample of n chromosomes, i copies of an allele are found. In the usual coalescent framework, the n chromosomes sampled are the tips of a gene genealogy representing the ancestry of the locus. When two ancestral lineages are descended from a common ancestor, those two lineages are said to coalesce and the number of ancestral lineages decreases by one. The number of lineages at any time in the past depends on the rate of coalescence, which in turn depends on the number of lineages and on the population size at that time (Hudson 1990). Griffiths & Tavaré (1998) showed that the probability distribution of the age, a, of an allele in the limit of large n is approximately

$$g(a) = CE\lfloor A(a)(A(a)-1)(1-x)^{A(a)-2}\rfloor, \tag{2}$$

where $x = i/n$ is the allele frequency, $A(a)$ is the number of ancestral lineages at time a in the past, and C is a normalization constant.

The quantity $A(a)$ is a random variable with value n at $a = 0$ (the present) and a probability distribution that can be expressed as an infinite series (Tavaré 1984). The numerical evaluation of that series is difficult, especially for small values of a, so it is impractical to use equation (2) to obtain $g(a)$. It is simpler to carry out a Monte Carlo simulation. The simulation method, which will be used to test the accuracy of the analytical approximation derived in § 2(b), is as follows. For each of a large number of replicates, the coalescent process is simulated using a method similar to that described by Hudson (1990). The probability of a coalescent event in generation t in the past is $A(t)(A(t)-1)/[4N(t)]$, where $N(t)$ is the population size at t. Then, at a set of specified times, a_1, a_2, . . . , the quantity in square brackets in equation (2) is calculated and stored. The averages of these stored quantities at the specified times provides an estimate of $g(a)$ at those times. The C program to carry out this simulation runs quickly (the source code is available from http://ib.berkeley.edu/labs/slatkin/software.html). For the results presented in figures 1 and 2, 10 000 replicates were sufficient to give very accurate results.

(b) *Approximate distribution of allele ages*

The simulation program described § 2(a) provides the distribution of allele age for a given frequency x and a given demographic history of the population, summarized by $N(t)$. It does not, however, provide us with any intuition about what factors affect the distribution of ages and how sensitive that distribution is to variation in x and $N(t)$. We can approximate $g(a)$ when x is small by using an approximate formula for $A(a)$. Slatkin & Rannala (1997) showed that the expected value of $A(a)$ is approximately

$$E[A(a)] \approx \frac{n}{1 + n\tau(a)/2}, \tag{3}$$

where

$$\tau(a) = \int_0^a \frac{\mathrm{d}t}{2N(t)} \tag{4}$$

is a scaled time.

For small a and large n, the variance of $A(a)$ about its expectation is relatively small. Although coalescent events occur at random times, there are so many of them initially that the randomness is averaged over to produce an almost deterministic decrease in $A(a)$. The randomness is apparent only later, when fewer ancestral lineages remain. Therefore, for small a, it is reasonable to replace $A(a)$ in equation (2) by its approximate expected value and obtain an analytical approximation for $g(a)$:

$$g(a) \approx C\bar{A}(a)(\bar{A}(a)-1)(1-x)^{\bar{A}(a)-2}, \tag{5}$$

where $\bar{A}(a) = E[A(a)]$ as given by equation (3).

Figure 1 tests the accuracy of equation (3) and figure 2 tests the accuracy of equation (5) by comparing the analytic results with those from the simulation program described in § 2(a). Figure 1 shows that equation (3) provides an excellent approximation to

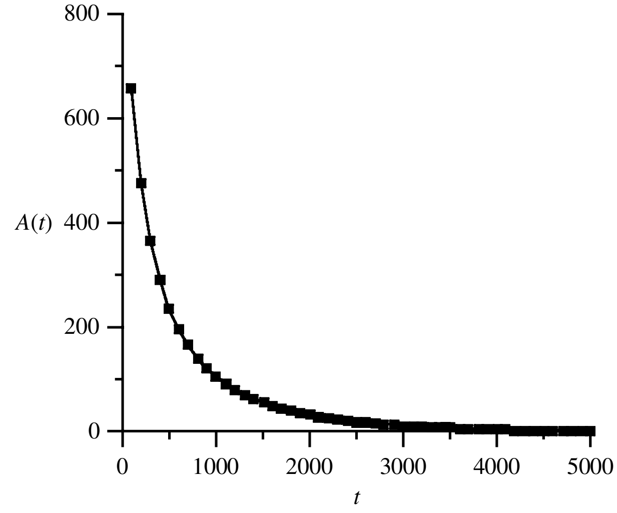

Figure 1. Comparison of the average number of ancestral lineages, $A(t)$, obtained from the simulation program described in the text and the analytical approximation given by equation (3). In this case, exponential growth at rate $r = 0.001$, a current population size $N_0 = 50\,000$, and $n = 1000$ lineages are assumed. The simulation results, indicated by the squares, are the average of 10 000 replicates. The analytical approximation is shown as the smooth line.

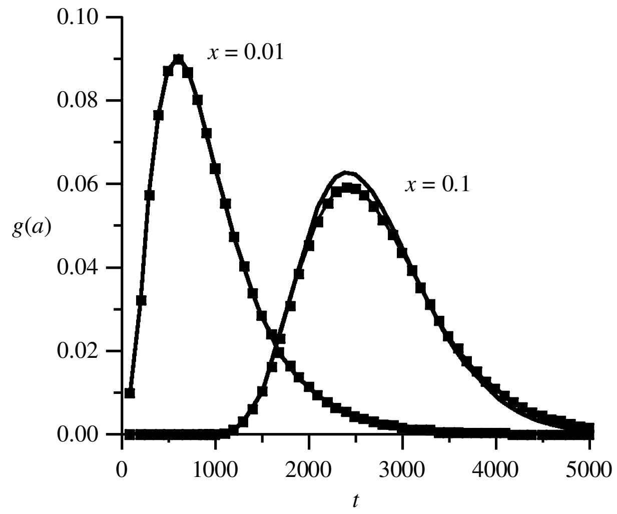

Figure 2. Comparison of the probability distribution of allele age obtained from the simulation program described in the text and the analytic approximation given by equation (5). All the results are for alleles at frequency x, and the population is assumed to have a current population size $N_0 = 50\,000$ and to have undergone exponential growth at rate $r = 0.001$. In the simulations, a sample size of $n = 10\,000$ and 10 000 replicates were used.

the average number of surviving lineages, even when that number is relatively small. Figure 2 shows that even for a relatively large allele frequency ($x = 0.1$) equation (5) is quite accurate. For $x < 0.01$, the approximate and exact distributions are indistinguishable. For $x > 0.1$, the variance of the actual distribution of ages is larger than that based on equation (5) because higher-frequency alleles are likely to have arisen by mutation when only a few ancestral lineages were present, and equation (3) does not allow for any randomness in the numbers of those few lineages.

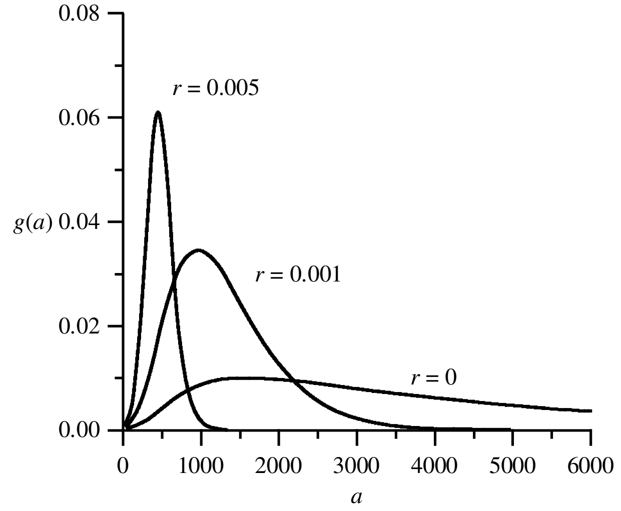

Figure 3. Approximate distributions of $g(a)$ for different population growth rates. In all cases, $N_0 = 10^6$, $n = 1000$ and $x = 0.001$. The curves were generated by plotting equation (5) in the text for different values of r.

For exponential growth at rate r, $N(t) = N_0 e^{-rt}$, and equation (5) implies that $g(a)$ depends on the combination of parameters, $R = 2N_0 r$. If $R \ll 1$, the distribution of ages is similar to that for a population of constant size and, in that case, the conclusion for the theory of allele age developed in the 1970s is correct. The distribution of ages consistent with a given frequency is broad and relatively little information about age is contained in the frequency. In a rapidly growing population, with $R \gg 1$, the results are quite different. Only a relatively narrow distribution of ages is consistent with a given frequency, so allele frequency does indicate age to within narrow bounds. Figure 3 shows $g(a)$ for $x = 0.001$ for three values of R. Many human populations are large and have undergone relatively rapid growth. For example, if we assume that the effective population size of the European population is about 10^8, even a very low growth rate implies a large value of R.

(c) *Approximate estimate of allele age*

Equation (5) provides a way to find a maximum-likelihood estimate (MLE) of age as a function of allele frequency. If we equate the derivative of $g(a)$ with respect to a to 0, the resulting equation implies that

$$\bar{A}(a) \approx \frac{2}{x}, \tag{6}$$

because $\bar{A}(a)$ is a monotonically decreasing function of a. Equation (6) implies that the MLE estimate of a, $\hat{a}$ satisfies

$$\tau(\hat{a}) = \int_0^{\hat{a}} \frac{dt}{2N(t)} \approx x. \tag{7}$$

For example, if the population had been of constant size, N, $\tau(a) = a/(2N)$ and equation (7) implies $\hat{a} = 2Nx$. If, instead, the population has grown exponentially at a rate r to the current size, N_0, $\tau(a) = (e^{ra} - 1)/2N_0 r$, and $\hat{a} = \ln(1 + 2N_0 rx)/r$. In figure 2, $2N_0 r = 100$, so $\hat{a} \approx 698$ generations if $x = 0.01$ and $\hat{a} \approx 2398$ generations if $x = 0.1$, which are in agreement with the results shown in figure 3.

It is likely that many human populations have grown at a higher rate in the recent past because of technological and medical advances. A more realistic model of growth is one in which the growth rate can change. Here I will assume that the rate changes from r_1 to r_2 T generations in the past

$$\begin{aligned} N(t) &= N_0 \mathrm{e}^{-r_1 t} \quad t < T \\ N(t) &= N_0 \mathrm{e}^{-r_1 T - r_2 (t-T)} \quad t > T. \end{aligned} \tag{8}$$

Equation (7) can be solved for the MLE of a under this model

$$\hat{a} = \frac{\ln(1 + 2N_0 r_1 x)}{r_1}, \tag{9a}$$

if this value is less than T, and

$$\hat{a} = T + \frac{1}{r_2} \ln\left[1 + 2N_0 r_2 x - \frac{r_2}{r_1}(1 - \mathrm{e}^{-r_1 T})\right], \tag{9b}$$

otherwise.

These results show that dependence of the MLE age, and hence the entire distribution of age, depends strongly on the recent growth rate, which is large for most human populations. Risch *et al.* (1995), for example, assumed a per generation growth rate for Ashkenazi Jews of *ca.* 0.4 per generation during the past 500 years, so $r_1 = 0.4$, $T = 25$ generations and a rough estimate of N_0 is 10^7 (Rannala & Slatkin 1998). If we let $r_2 = 0.001$ and consider alleles with frequencies 10^{-5}, 10^{-4}, 10^{-3} and 10^{-2}, for the sake of illustration, we find $\hat{a} = 11$, 16, 22 and 32 generations. The recent very rapid growth of such a population ensures that low-frequency alleles are very young. The value of r_2 assumed has little effect on the results.

(d) *Test for selection based on allele age*

The theory described in the preceding sections provides an estimate of allele age based on the allele frequency. An independent estimate of allele age can be obtained from the extent of intra-allelic variability (Slatkin & Rannala 1997). How this estimate is obtained depends on the kind of data available. In this paper, I will consider only a rare allele and intra-allelic variability at a closely linked biallelic marker. A data set for the test described here consists of a number of chromosomes carrying the allele of interest and the allelic state of the marker locus on each chromosome. The previous sections were concerned with a single locus but in this section a model of two loci with recombination between them is needed.

The study by Neuhausen *et al.* (1996) of intra-allelic variability of alleles at the locus BRCA1 associated with early-onset breast cancer in humans provides suitable data. In 17 chromosomes carrying the mutation 5383*ins*C at BRCA1, the genotype at a linked microsatellite locus, D17S1320, was determined. Out of those 17 chromosomes, 12 could be unambiguously typed as carrying allele 7 at the microsatellite locus, which is at a frequency 0.27 in the population as a whole. For the remaining five chromosomes, haplotype information was not available, but the genotypes at the marker locus all contained allele 7 along with allele 3, 4 or 5. The data of Neuhausen *et al.* (1996) are consistent with the hypothesis that the region of conserved ancestral haplotype extends beyond D17S1320, so it is likely that all 17 chromosomes carry allele 7 at the marker, but in the later analysis I will allow for the possibility that fewer than 17 carry that allele. Although the linked marker locus has several alleles, it will be treated as biallelic, with alleles other than allele 7 grouped.

To analyse such data, I adapt the theory described by Rannala & Slatkin (1998). For a low-frequency allele that arises by mutation at time a and is found in i copies in a sample today, the number of copies at any time can be approximated by a linear birth–death process. Using this process, it is relatively easy to find the joint distribution of intra-allelic coalescence times, denoted by $t_2, \ldots, t_i$. The time t_k is the time at which the number of ancestral lineages increased from $k-1$ to k. It is easy to draw randomly a set of coalescence times from this joint distribution.

At a biallelic marker locus with alleles M and m, one of them, say M, was on the chromosome carrying the allele of interest at t_2, the time of the most recent common ancestor of all copies in the sample. Between t_2 and the present, recombination with chromosomes not carrying the allele and mutation at the marker locus will cause some of the i chromosomes in the sample to carry m. Let j be the number of M-bearing chromosomes in the sample, and let $p_j^{(i)}$ be the probability of j ($j = 0, \ldots, i$). The superscript indicates a sample size i.

Given the set of coalescence times and a model of change in state at a linked marker locus, it is straightforward, albeit slightly complicated, to compute $p_j^{(i)}$. Assume that the mutation rate from M to m is μ, the mutation rate from m to M is ν, and that the recombination rate between the mutant allele and the M/m locus is c. In one generation, the probability that M is replaced by m because of mutation and recombination is $u = (1-q)c + \mu$ and the probability that an m is replaced by M is $v = qc + \nu$, where q is the frequency of M in the population. Therefore, the probability that a chromosome initially carrying M will carry m after t generations is approximately

$$f_{21} = \frac{u}{u+v}(1 - \mathrm{e}^{(u+v)t}), \tag{10a}$$

and the probability that a chromosome initially carrying m will carry M after t generations is

$$f_{12} = \frac{v}{u+v}(1 - \mathrm{e}^{(u+v)t}). \tag{10b}$$

These equations describe the change in the extent of intra-allelic variability on a single chromosomal lineage. To find the probability distribution of the number of lineages carrying M in a sample of i chromosomes, we have to account for both the coalescent events and the independent changes on lineages present at different times in the past. To do this, we first assume we know the set of intra-allelic coalescent times, $(t_2, \ldots, t_i)$. Between t_k and t_{k+1} there are k lineages ancestral to the sample, where we interpret t_{i+1} as 0, the present. When there are k lineages, we can represent the configuration of the population by a vector, $\boldsymbol{p}^{(k)}$ with elements $p_j^{(k)}$ representing the probabilities that there are $j = 0, 1, \ldots, k$ lineages carrying M. If we know these probabilities at $t = t_k$, then we can compute them immediately before t_{k+1} by multiplying by a $(k+1)$ by $(k+1)$ transition matrix denoted by $T^{(k)}$ whose entries we can find from equations (10) and the assumption that events on different lineages are independent. If there are j chromosomes carrying M and $k-j$ chromosomes carrying m, then the number of M to m transitions is binomially distributed with probability f_{21} and sample size j, and the number of m to M transitions is binomially distributed with probability f_{12} and sample size $k-j$. The elements of $T^{(k)}$ are found by taking the appropriate convolutions of these binomial distributions.

At t_{k+1}, there is a coalescent event, so one of the k lineages is chosen randomly to give rise to two descendent lineages. The effect of this event can be modelled by multiplying $p^{(k)}$ by a $(k+1)$ by $(k+2)$ matrix, $S^{(k)}$, whose jlth element is $1-j/k$ for $l=j$, j/k for $l=j+1$ and 0 otherwise. Multiplying the vector

$\boldsymbol{p}^{(k)}$ by $S^{(k)}$ produces $\boldsymbol{p}^{(k+1)}$ that has $k+2$ elements representing the probabilities of the $k+2$ possible configurations, $j=0, \ldots, k+1$.

Given the assumption that initially the mutant allele was on a chromosome carrying M, $\boldsymbol{p}^{(2)}=(1,0,0)$ immediately after t_2. Multiplying successively by $T^{(k)}$ and $S^{(k)}$ provides the probabilities of later configurations. In particular, the vector of probabilities that there are j As associated with the mutant in a sample today is given by

$$p^{(i)} = T^{(i)} \prod_{k=2}^{i-1} S^{(k)} T^{(k)} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}. \qquad (11)$$

To find approximately the overall probability of a configuration, the average is taken of many replicate sets of coalescence times.

For a given value of a, this theory provides the probability of each configuration of the data, that is $p_j^{(i)}$ for each j. Averaging over a using $g(a)$ as defined in the previous section gives us the overall probability of each configuration, which I will denote by p_j. The p_j provide the basis for a test for selection in a data set for which x and j are known. The probability that values of j as extreme as, or more extreme than, the observed value can be computed, and, if that probability is less than a specified threshold, neutrality can be rejected. That is, if j_o is the observed number of non-recombinants, then the tail probability is

$$P = \Pr(j \geqslant j_o) = \sum_{j=j_o}^{i} p_j, \qquad (12)$$

and the hypothesis of neutrality of the allele is rejected if P is less than a specified value, 0.05 or smaller.

To illustrate this test, consider the 5382*ins*C of BRCA1 discussed above. Rahman & Stratton (1998) estimate the frequency of alleles at BRCA1 that are associated with early-onset breast cancer to be *ca.* 0.0006. Neuhausen *et al.* (1996) studied 61 families carrying one of the six most common disease-associated alleles at BRCA1, and state that these six alleles account for roughly one-third of the disease-associated alleles at that locus. Out of these 61 families, 21 carried the 5382*ins*C allele. Therefore its allele frequency is *ca.* 0.0006 × 1/3 × 21/61 or *ca.* 7×10^{-5}. Four families were not typed for D17S1320 so $i=17$, not 21. To apply the theory developed here, something must be assumed about the demographic history of the population sampled, which in this case is a population mostly of mixed European and Ashkenazi Jewish ancestry. I assumed a current effective population size of $N_0=10^8$ individuals and past growth rates of $r=0.005$ and 0.002. The theory of Rannala & Slatkin (1998) also requires knowledge of the fraction, f, of the population represented in the sample. I used $f=10^{-4}$, 10^{-5} and 10^{-6}. The marker locus and BRCA1 are separated by *ca.* 500 kb. I assumed two recombination rates $c=0.005$ (corresponding to an average rate for humans of 1 cM = 1 mb) and $c=0.00125$ (corresponding to one-quarter of the average rate), and did not allow for mutation at the marker locus. The data presented by Neuhausen *et al.* (1996) are consistent with a relatively low mutation rate at the marker locus and imply that the main cause of intra-allelic variability is recombination. Allowing for any mutations at the marker locus would make all the P-values smaller.

For 5382*ins*C, $P=0.000\,136$ if $j_o=17$ and the most favourable combination of parameter values (the lower recombination rate and lower population growth rate) is assumed. Changing the sampling fraction, f, made almost no difference. With the other parameter values, the P-values were even lower (as low as 10^{-8}). Therefore, these data strongly reject the hypothesis of neutrality. Because of the lack of resolution of all of the haplotypes, there is some uncertainty about the value of j_o. If $j_o=16$, $P=0.001\,27$, and if $j_o=15$, $P=0.0059$, both for the most favourable cases. Therefore, even one or two recombinant chromosomes in the sample would not be consistent with neutrality.

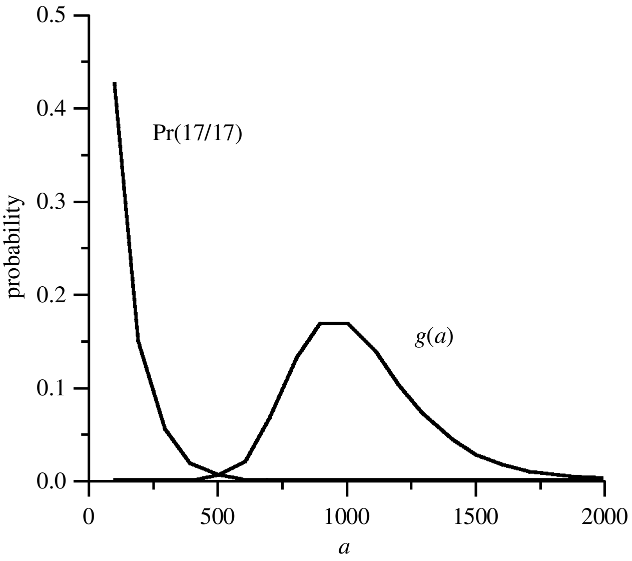

Figure 4. A plot of $g(a)$ and the probability of obtaining 17/17 non-recombinant chromosomes as a function of allele age, a, for the allele 5382*ins*C at BRCA1. The results plotted assume $r=0.002$, $N_0=10^8$, $f=0.0001$ and $c=0.00125$. Results for other combinations of parameter values are similar.

Figure 4 shows some typical results for the distribution of ages, $g(a)$, and the probabilities of observing $j_o=17$ non-recombinants under these assumptions. We can see why we reject neutrality for these parameter values. The extent of intra-allelic variability is consistent with a much younger age than is the frequency.

A similar conclusion is reached for another common allele, 185*del*AG, which was found in 19 of the families studied by Neuhausen *et al.* (1996). For this allele, 15 chromosomes were typed at D17S1320 and ten could be unambiguously typed as carrying allele 7 at the microsatellite locus. In the five others for which haplotypes were not resolved, all had allele 7 as one of the two. If $j_o=15$ for this allele, $P=0.000\,64$ for the most favourable combination of parameter values.

We can conclude that there is evidence for selection on the two most frequent disease-associated alleles at BRCA1. Their extent of intra-allelic variability as measured by the linked microsatellite locus, D17S1320, is too small to be consistent with the allele age indicated by the allele frequency.

3. DISCUSSION AND CONCLUSIONS

Understanding the evolutionary history of individual alleles is of increasing importance because of relevance to the mapping of alleles associated with genetic diseases and because of the potential for understanding the effect of natural selection on each allele. New and more efficient methods for assessing the extent of intra-allelic variability will provide additional information that will allow increasingly refined estimates of allele age and lead to more sensitive tests of past selection. The results developed here provide a general way of exploring the relationship between allele age and allele frequency under various assumptions about past population sizes. The approximations derived are quite accurate for low-frequency alleles and so can be used in place of simulations.

The theory in this paper was developed under the assumption of random mating in a single population, but it applies to a low-frequency allele in a subdivided population provided that gene flow is conservative, meaning that gene flow itself does not change the average allele frequency. The independence of population subdivision is easily seen. The number of copies of a rare allele can be described approximately by a birth–death process, and under that approximation each copy reproduces independently of how many other copies are present and independently of where each copy is. Therefore, the probability distribution of ages and the joint distribution of intra-allelic coalescence times are also approximately independent of population subdivision, so the test for selection described here is valid for low-frequency alleles in a subdivided population. My unpublished simulation results show that the distribution of allele ages is independent of migration rate in a two-deme coalescent model as well.

Any inferences about allele age and selection depend on assumptions about the demographic history of the population under study. In many cases, the demographic history is unknown or not well-enough known to allow firm conclusions to be drawn. When only one or two alleles can be studied, any conclusions about selection can only be tentative. But all autosomal loci have the same demographic history, on average, so that, as data from different regions of the human genome are accumulated, it will be possible to compare patterns of intra-allelic variability among alleles of the same frequency to determine which are aberrant. When broad patterns of intra-allelic variation are known, we can obtain an increasingly detailed understanding of the history of selection on individual alleles.

This research was supported in part by a grant from the US National Institutes of Health (RO1-GM40282) and in part by a fellowship from the John Simon Guggenheim Foundation.

REFERENCES

Andolfatto, P., Wall, J. D. & Kreitman, M. 1999 Unusual haplotype structure at the proximal breakpoint of *In*(*2L*)*t* in a natural population of *Drosophila melanogaster*. *Genetics* **153**, 1297–1311.

Depaulis, F. & Veuille, M. 1998 Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Mol. Biol. Evol.* **15**, 1788–1790.

Griffiths, R. C. & Tavaré, S. 1998 The age of a mutation in a general coalescent tree. *Stochast. Mod.* **14**, 273–295.

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

Hudson, R. R., Bailey, K., Skarecky, D., Kwiatowski, J. & Ayala, F. J. 1994 Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster*. *Genetics* **136**, 1329–1340.

Kimura, M. & Ohta, T. 1973 The age of a neutral mutant persisting in a finite population. *Genetics* **75**, 199–212.

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**, 274–286.

Maruyama, T. 1974 The age of an allele in a finite population. *Genet. Res. Camb.* **23**, 137–143.

Maynard Smith, J. & Haigh, J. 1974 The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35.

Morral, N., Bertranpetit, J., Estivill, X., Nunes, V., Casals, T., Gimenez, J., Reis, A., Varon-Mateeva, R. & Macek Jr, M. 1994 The origin of the major cystic fibrosis mutation (DELTA-F508) in European populations. *Nat. Genet.* **7**, 169–175.

Neuhausen, S. L. (and 24 others) 1996 Haplotype and phenotype analysis of six recurrent BRCA1 mutations in 61 families: results of an international study. *Am. J. Hum. Genet.* **58**, 271–280.

Rahman, N. & Stratton, M. R. 1998 The genetics of breast cancer susceptibility. *A. Rev. Genet.* **32**, 95–121.

Rannala, B. & Slatkin, M. 1998 Likelihood analysis of disequilibrium mapping, and related problems. *Am. J. Hum. Genet.* **62**, 459–473.

Risch, N., de Leon, D., Ozelius, L., Kramer, P., Almasy, L., Singer, B., Fahn, S., Breakefield, X. & Bressman, S. 1995 Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. *Nat. Genet.* **9**, 152–159.

Serre, J. L., Simon-Bouy, B., Mornet, E., Jaume-Roig, B., Balassopoulou, A., Schwartz, M., Taillandier, A., Boué, J. & Boué, A. 1990 Studies of RFLP closely linked to the cystic fibrosis locus throughout Europe lead to new considerations in population genetics. *Hum. Genet.* **84**, 449–454.

Slatkin, M. & Bertorelle, G. 2000 The use of intra-allelic variability for testing neutrality and for estimating selection intensity and population growth rate. (In preparation.)

Slatkin, M. & Rannala, B. 1997 Estimating the age of alleles by use of intraallelic variability. *Am. J. Hum. Genet.* **60**, 447–458.

Slatkin, M. & Rannala, B. 2000 Estimating allele age. *A. Rev. Genom. Hum. Genet.* **1**, 225–249.

Stephens, J. C. (and 38 others) 1998 Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. *Am. J. Hum. Genet.* **62**, 1507–1515.

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

Watterson, G. A. 1976 Reversibility and the age of an allele. I. Moran's infinitely many neutral alleles model. *Theor. Popul. Biol.* **10**, 239–253.