Genet. Res., Camb. (2008), **90**, *pp.* 129–137.
doi:10.1017/S0016672307008944 Printed in the United Kingdom

A Bayesian method for jointly estimating allele age and selection intensity†

MONTGOMERY SLATKIN*
Department of Integrative Biology, 3060 VLSB, University of California at Berkeley, Berkeley, CA 94720-3140, USA

(*Received 13 June 2007 and in revised form 23 October 2007*)

Summary

The problem of jointly estimating the intensity of past selection affecting an allele and the allele's age is formulated in a Bayesian framework. The prior distribution of allele age given its frequency is obtained from existing population genetics theory. The prior distribution of selection intensity is assumed to reflect the fact that positive selection on a new mutant is more likely to be weak than strong. The general approach is illustrated by the development of an importance sampling method applicable to low-frequency alleles. This method can be used either when the haplotypes of closely linked marker loci are known or when the lengths of linked ancestral chromosomal segments can be inferred. The method is illustrated with an application to the A− allele of G6PD in Africa. Because changes in allele frequency and recombination are both intrinsically stochastic, there are limits to the accuracy achievable with any method.

1. Introduction

Neutral loci linked to an allele of interest provide information about the history of that allele since it arose by mutation. The extent of linkage disequilibrium (LD) with linked markers allows tests of neutrality (Sabeti *et al*., 2002; Slatkin, 2000; Slatkin & Bertorelle, 2001; Toomajian *et al*., 2003; Voight *et al*., 2006), estimates of selection intensity (Charlesworth, 2006; Charlesworth *et al*., 2001; Slatkin, 2001; Wiuf, 2001*a*) and estimates of allele age (Goldstein *et al*., 1999; Guo & Xiong, 1997; Kaplan *et al*., 1994; McPeek & Strahs, 1999; Reich & Goldstein, 1999; Risch *et al*., 1995; Slatkin & Rannala, 1997; Stephens *et al*., 1998; Tishkoff *et al*., 1996). These problems are usually considered separately. When the goal is to estimate both age and selection intensity, age is estimated first and then selection intensity is inferred from the age by using a deterministic model of selection, an approach first taken by Stephens *et al*. (1998).

In this paper, I consider the problem of jointly estimating the age of an allele and the intensity of selection it has experienced since it arose by mutation. I will formulate the problem in a Bayesian framework that accounts for the stochasticity of both allele frequency change and recombination. I then illustrate the approach by presenting a method for analysing low-frequency alleles linked to one or more marker loci. The method presented here is similar to that of Rannala & Reeve (2001) for linkage disequilibrium mapping in that it assumes that the dynamics of a low-frequency allele can be modelled by a linear birth–death process, and it approximates the probability of the data using Monte Carlo summation. It differs from Rannala & Reeve's method in using importance sampling instead of the Metropolis–Hasting algorithm for approximating the probability of the data.

One of the goals of this paper is to show that even under ideal conditions, when the genetic and demographic parameters are known without error, precise estimates of both allele age and selection intensity cannot be obtained because of the intrinsic unpredictability of genetic drift and recombination. This conclusion is in contrast to the conclusion reached when allele age is estimated using a 'model-free' method that assumes that the gene genealogy is a star (meaning that there are no internal branches) and that

* Telephone: +1 (510) 6436300. Fax: +1 (510) 6436264. e-mail: slatkin@berkeley.edu
† This paper is dedicated to Deborah Charlesworth, whose experimental and theoretical studies in population genetics have illuminated so many areas.

deterministic theory can be used to estimate the selection intensity under the assumption that the estimated age is the true age. The model-free approach does not take account of all sources of uncertainty in the estimates of allele age and selection intensity and hence exaggerates the accuracy of the resulting estimates.

2. Theory

(i) *Formulation as a problem in Bayesian inference*

Throughout, I will be concerned with the history of an allele, denoted by M, that arose as a unique mutation at time t_1 in the past in a population previously fixed for the alternative allele, m, at the locus of interest. All copies of M are identical by descent with the initial copy. Although t_1 is the true allele age, the time of the most recent common ancestor of all copies of M in a sample, called t_2 below, is also referred to as the age (Slatkin & Rannala, 2000). The goal is to estimate t_1 and to infer something about selection experienced by M since t_1. The genetic information available is the frequency of M in the population (x) and data from marker loci closely linked to M.

To describe the method in as simple a framework as possible, selection on M will be additive with selection coefficient s: the relative fitnesses of MM, Mm and mm individuals are $1+2s$, $1+s$ and 1. I assume a prior distribution of s, $\Pr(s)$ that reflects what is known about selection on individual alleles. The strongest positive selection known in humans is on the S allele of the β-globin gene which causes sickle-cell anemia. Heterozygous carriers of S have roughly a 20% higher rate of surviving to adulthood than normal homozygous individuals in regions with a high incidence of malaria (Vogel & Motulsky, 1996). For analysing data from human populations, a reasonable prior distribution of s decreases with s and is small for $s>0{\cdot}2$. In the analysis of G6PD, I assume $\Pr(s)=\alpha e^{-\alpha s}$, where α is a parameter with a value in the range 10–50.

The allele frequency, along with s and assumptions about past population growth, provide a prior distribution of t_1, $\Pr(t_1|x,r,s)$ (Slatkin, 2002*b*). Genetic data from linked marker loci provide additional information about s and t_1 that leads to the posterior distribution of s and t_1:

$$\Pr(s,t_1|x,r,G)=\frac{\Pr(t_1|x,r,s)\Pr(G|t_1,x,r,s)\Pr(s)}{\Pr(G|x,r)} \quad (1)$$

where G represents the genetic data.

How the posterior distribution is used depends on the goal of a study. If the goal is to estimate selection intensity, the best estimate of s is obtained by averaging over t_1, $\Pr(s|x,r,G)=\int \Pr(s,t_1|x,r,G)dt_1$. If, instead, the goal is to estimate t_1 or to compare

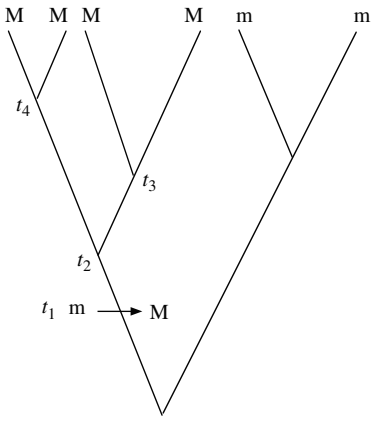

Fig. 1. The intra-allelic genealogy of a mutant M found in four copies ($i=4$), illustrating the definitions of allele age, t_1, and the intra-allelic coalescence times, t_2, t_3 and t_4.

t_1 with historical information, then the posterior distribution of t_1 is obtained by averaging over s, $\Pr(t_1|x,r,G)=\int \Pr(s,t_1|x,r,G)ds$. The resulting probability distribution can be used either to test a specific historical hypothesis, for example the probability that t_1 is larger or smaller than a given time, or to find the most probable value of t_1 and the associated confidence interval. In other situations, the goal is to find how much information about both t_1 and s is provided by the data, in which case the joint posterior distribution itself should be used.

(ii) *Linear birth–death approximation for low-frequency alleles*

To implement the above theory, I assume that the population has grown exponentially at rate r to its current size, N_0: the population size at time t in the past is $N(t)=N_0e^{-rt}$. When x is small, the evolution of the number of copies of M can be approximated by a linear birth–death process (Wiuf, 2001*b*), for which analytical expressions for $\Pr(t_1|x,r,s)$ are available (Slatkin, 2002*b*). The formulas used here are given in Appendix A.

The probability of the data given t_1, $\Pr(G|t_1,x,r,s)$, depends on the intra-allelic genealogy, meaning the gene tree of the i copies of M in a sample. As illustrated in Fig. 1, the intra-allelic genealogy is characterized by a set of $i-1$ coalescence times, $t_2,\ldots,t_i$, and the topology, B (for branching pattern). The dependence of $\Pr(G|t_1x,r,s)$ on the coalescence times and topology can be written

$$\Pr(G|t_1,x,r,s)=\sum_B \Pr(B)\int_{\{t_2,\ldots t_i\}} \Pr(G|t_2,\ldots t_i,B) \times \Pr(t_2,\ldots t_i|x,r,s,t_1)dt_2\cdots dt_i, \quad (2)$$

where the sum is over all topologies with i tips and the integral is over all sets of intra-allelic coalescence

https://doi.org/10.1017/S0016672307008944 Published online by Cambridge University Press

times such that $t_2 \geqslant t_3 \ldots \geqslant t_i > 0$ (Felsenstein, 1988). The first term in the integrand represents the effect of recombination on loci linked to M; it depends only on the intra-allelic genealogy and not on the probability attached to that genealogy. The second term represents the dependence of the joint distribution of intra-allelic coalescence times, which is easily derived for a linear birth–death process (Appendix A). Under the assumption that all copies of M are equivalent, the distribution of topologies is generated by assuming each lineage is equally likely to branch at each coalescent event, and hence depends only on i.

(iii) *Genetic data*

I will consider here two kinds of genetic data. The first is the set of multi-locus haplotypes linked to each copy of M in the sample. I assume there are K diallelic marker loci, numbered 1 to K, with alleles at each locus labelled 0 and 1. The 1 allele is assumed ancestral in the sense that it was on the chromosome carrying the first copy of M. There are 2^K possible haplotypes and their frequencies on non-M chromosomes are y_j, $j = 1 \ldots 2^K$, which are parameters of the model assumed to be known without error and not to have changed since t_1. The genetic data, G, consists of the haplotype associated with each M-bearing chromosome in the sample. I will discuss later the effects of uncertainly in y_j and the haplotype phase of the markers on the M-bearing chromosomes.

The second kind of genetic data assumes that marker loci linked to M are sufficiently dense that it is possible to determine the length of the ancestral segment still linked to each copy of M. The frequency of each marker allele is unimportant. In this case the data consist of the lengths of ancestral segments measured in terms of the total recombination rate. I will assume the same rate c between every pair of sites. In that case, the data is a list of lengths in bases, $l_1, \ldots, l_i$, of ancestral segments linked to each copy of M in the sample. This kind of data differs from the minimum length of the ancestral fragments linked to M, $l = \min[l_1, \ldots, l_i]$, which has been used as a summary statistic (McPeek & Strahs, 1999; Slatkin, 2001; Slatkin & Bertorelle, 2001).

For both kinds of data, the computational problem is to find $\Pr(G|t_1, x, r, s)$. As is usually the case when summing over tree topologies and integrating over branch lengths, exact computations are not feasible for more than three or four M-bearing chromosome. Here, I use an importance sampling method introduced previously (Slatkin, 2002*a*). The transition matrices used for the two types of data are summarized in Appendix B. The results from using the importance sampling method were checked by comparing them with results for one and two marker loci which could be obtained by other means.

3. Positive or negative selection

The Bayesian framework described in the previous sections does not require that selection favoured of the mutant of interest. In fact, unless an additional step is taken, it is not possible to determine whether the mutant experienced positive or negative selection. The problem is seen most clearly in the case of an allele with an additive effect on fitness in a population of constant size. Maruyama (1974) showed that, under these assumptions, the prior distribution of age is independent of the sign of the selection coefficient. Although Maruyama's conclusion is no longer true if there is dominance or past population growth, the implication of his result for the problem of estimating the selection coefficient from allele age is still important. A deleterious allele, if it is found in any substantial frequency, has to be relatively young and hence in LD with linked markers, just as would an advantageous allele.

Maruyama's result led Nordborg & Tavaré (2002) to conclude that extensive LD near an allele might well indicate selection against it rather than in favour of it. They suggested that which is more probable depends on the relative rates of deleterious and advantageous mutations. And they concluded that our current lack of knowledge about mutation parameters renders the problem of deciding whether selection was in favour of or against a mutant 'philosophical'. Nordborg & Tavaré's (2002) argument is not convincing, however, because it ignores the fact that an advantageous mutation is much more likely to reach a given frequency than will a mutation that is deleterious to the same extent. In Appendix C, I show how to use the birth–death approximation to find the ratio of probabilities that a mutant was advantageous instead of deleterious. If the mutant is present in any substantial frequency, that ratio is so large that the possibility of negative selection can be ignored. For example, in the case of G6PD, discussed below, it is approximately $1{\cdot}4 \times 10^{16}$ more likely that the selection coefficient affecting the A− allele of G6PD is $+0{\cdot}05$ than $-0{\cdot}05$.

4. Application to G6PD

The Bayesian method is illustrated by applying it to the data of Saunders *et al.* (2005) who sequenced 20 chromosomes carrying the A− allele at the X-linked G6PD locus. This allele is found in sub-Saharan African populations in frequencies up to 20%. It is thought to confer partial resistance to malaria (Vogel & Motulsky, 1996). Saunders *et al.* (2005) sequenced eight loci in a 2·5 Mb region surrounding G6PD in 51 unrelated males of sub-Saharan African origin. Twenty of these males carried the A− allele. LD at linked SNP markers spanned a region of 1·6 Mb on

the A− chromosomes. By applying a simplified version of the method presented here, Saunders *et al.* (2005) concluded that heterozygous carriers of A− had between a 10% and 20% fitness advantage over homozygous individuals and that A− arose by mutation less than 150 generations (3750 years) ago. Part of the purpose of this paper is to formalize and generalize the method used in that paper and in the paper of Wood *et al.* (2005).

Following Saunders *et al.* (2005), I assume no population growth ($r=0$) and a frequency, $x=0{\cdot}11$, of A−. With a population size of $N_0=100\,000$, a fraction $f=9{\cdot}1\times10^{-4}$ of the population would have to be sampled to obtain 20 copies of A− in a sample.

To analyse the linked haplotypes, I considered four SNPs (numbers 36, 41, 55 and 99), one each from the four loci BGN, IDH3G, 1CAM and G0·9MT in figure 2 of Saunders *et al.* (2005). I assumed that the nucleotide carried by the first individual listed (ALB77) was on the ancestral A− chromosome and denoted it by 1. The other nucleotide was denoted by 0. With this convention, each haplotype can be numbered from 0 to 15 by treating the haplotype as a binary number. With this convention, the data of Saunders *et al.* can be reduced to six A− chromosomes with haplotype 15, three with haplotype 9, seven with haplotype 7, one with haplotype 3 and three with haplotype 1. I used the haplotypes of the 31 non-A− chromosomes to estimate the frequencies of the 1 allele to be 0·419, 0·0323, 0·0645 and 0·433 for the four loci. There is no obvious LD on the non-A− chromosomes, so I computed the background frequencies of the 15 haplotypes by assuming linkage equilibrium.

The linkage map for these four loci was inferred by assuming that the rate per base is constant in this region and using the estimated rate between L1CAM and G6PD of 0·01675 to estimate the rate per base. The map distances from BGN were 0·0086 for IDH3G, 0·0131 for L1CAM, 0·02985 for G6PD and 0·0354 for G0·9MT. From the background frequencies and map distances, the elements of the 16×16 transition matrix were obtained.

The other way of analysing this kind of data is to infer the length of the ancestral chromosome still linked to A−. That assessment is somewhat subjective. For example, in the G6PD data set the first four of the A− chromosomes (ALB77, SHO07, VA065 and IVC17) have nearly the same haplotype as far as the BGN locus, which is 991 kb from G6PD. But the sequence of VA065 differs from the other three at SNP 20 and the sequence of IVC17 differs from the other three at SNP 36, both of which are in the BGN locus. The two apparently aberrant sites could differ from the others because of gene conversion or mutation that did not otherwise affect the ancestral fragment. Or they could indicate that recombination had taken place between BGN and IDH3G and then subsequent recombination recreated what appears to be the ancestral haplotype. I assumed that a single aberrant SNP was not sufficient evidence to indicate recombination. With that assumption, the data of Saunders *et al.* (2005) reduce to the following: four of the A− chromosomes had the conserved ancestral fragment to BGN on the centromeric side of G6PB, seven to IDH3G, one to L1CAM, four to TAZ and one had no conserved fragment. I did not analyse the telomeric side of G6PD because all 20 A− chromosomes had identical haplotypes on that side. I rounded the distances to 50 bp intervals. BGN was assumed to be 20 units from G6PD, IDH3G was 14 units, L1CAM was 11 units and TAZ was 2 units. The recombination distance per 50 bases was 0·01675/11 = 0·001528, based on the estimated recombination rate between G6PD and L1CAM.

The results from analysing the two data sets are presented in Fig. 2. Part (*a*) shows the joint posterior distribution of s and t_1 based on the analysis of ancestral fragments; (*b*) shows the same distribution obtained from the analysis of four marker loci. The two distributions have the same character. There is a ridge indicating that t_1 and s are somewhat confounded. The reason is that both the prior distribution of allele age and the distribution of intra-allelic coalescence times depend strongly, although not exclusively. on the product st_1. The two data sets nevertheless result in joint estimates of s and t_1 that are consistent with each other. The maximum for Fig. 2*a* is at $s=0{\cdot}26$ and $t_1=40$ generations and for Fig. 2*b* at $s=0{\cdot}24$ and $t_1=40$ generations.

The similarity of the most probable values of s and t_1 does not tell the whole story. Although it is somewhat difficult to see in the three-dimensional graphs in Fig. 2, there is much more variability in Fig 2*b* than in 2*a*. The variability is more obvious when the marginal distributions of s and t_1 are computed (Fig. 3). The marginal distributions of s and t_1 for the model of ancestral segments (Fig. 3*a* and *b*) and are much smoother than the corresponding distributions based on the four marker loci (Fig. 3*c* and *d*). The results for both models are based on 10^6 replicates of the importance sampling algorithm, so the difference between results from the two models reflects the greater intrinsic variability in recombination in a small set of marker loci.

5. Discussion

The Bayesian framework for jointly estimating allele age and selection intensity is quite general and can be adapted to other models of selection and to other assumptions about population history. The implementation of the method presented here is appropriate for low-frequency alleles. The application to

https://doi.org/10.1017/S0016672307008944 Published online by Cambridge University Press

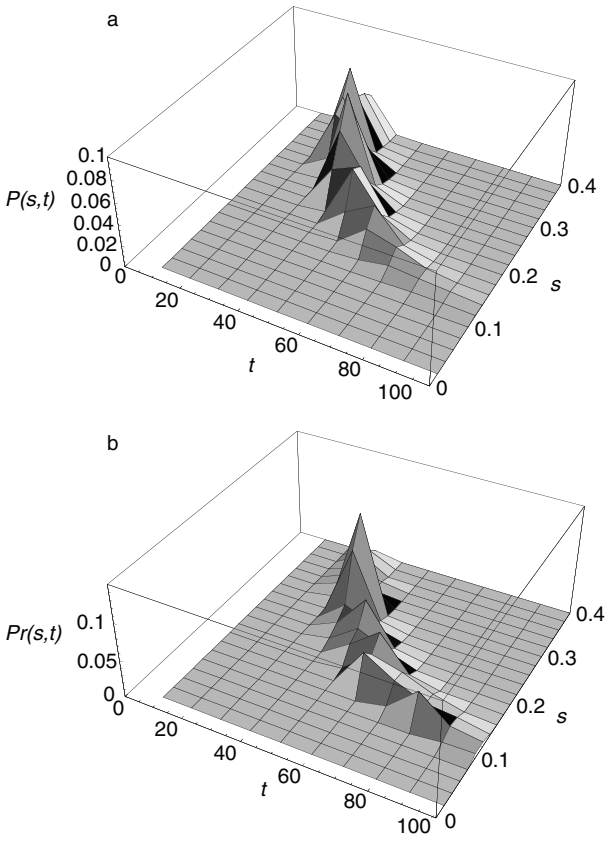

Fig. 2. The joint posterior distribution of s and t_1 computed for the A− allele of G6PD using the importance sampling method described in the text. The data and genetic map were taken from Saunders *et al.* (2005). In both cases, 10^6 replicates were used. (*a*) $\Pr(s, t_1)$ estimated from the lengths of ancestral chromosomal segments linked to A−. The identification of the ancestral segments is described in the text. (*b*) $\Pr(s, t_1)$ estimated from data for four marker loci (BGN, IDH3G, 1CAM and G0·9MT) linked to G6PD. The details of the genetic map are described in the text.

the A− allele of G6PD shows that allele age and selection intensity are partially confounded because of the underlying population genetic processes. We can see why that is true by considering a deterministic model of selection on an allele with additive effect s on fitness. The differential equation that approximates the change in allele frequency is

$$\frac{dp}{dt} = sp(1-p) \tag{3}$$

where p is the allele frequency and t is time (Hartl & Clark, 1997). This equation is unchanged if s is multiplied by a constant C, $s' = Cs$, and the units of time are reduced by the same factor, $t' = t/C$. Hence any function that depends on the product st is unchanged.

The reason that s and t_1 are not completely confounded is that, when M is at low frequency, the trajectory of allele frequencies is not deterministic; it is also affected by genetic drift. Drift and selection are both accounted for in the linear birth–death approximation. The equations in Appendix A show that both the prior distribution of age and the joint distribution of intra-allelic coalescence times depend on s and t_1 separately as well as on the product st_1.

The Bayesian method accounts for both the stochastic dynamics of a low-frequency allele and the stochastic nature of recombination. These two sources of unpredictability make it difficult to obtain precise estimates of age and selection intensity even under the idealized conditions assumed here. Any additional sources of uncertainty, for example in estimated map distances, estimated haplotype frequencies of the non-M chromosomes or inferred haplotype phases, further reduce the ability to estimate ages and selection intensities.

The importance of accounting for all sources of variability is evident in the analysis of the G6PD data set. Figs 2 and 3 show that there is more variability in the results for the four marker loci than for the ancestral segments. The reason is that, in the model of individual marker loci, there are more sources of uncertainty. In this example, the age of the A− allele is small enough ($\approx$50 generations) and the recombination distances between the markers are also small enough ($\approx$1–3 cM) that very few recombination events are expected after A− arose by mutation. The data do not allow a precise determination of the number or order of those recombination events and hence do not narrowly constrain the estimates of s and t_1. Failure to take account of all sources of variability, as in 'model-free' methods that assume that the intra-allelic genealogy is a star (Neuhausen *et al.*, 1996; Reich & Goldstein, 1999), lead to exaggerated confidence in the resulting estimates.

The estimated selection intensity on the A− allele of G6PD, 0·24, is larger than the minimum estimate of 0·05 reported by Saunders *et al.* (2005), and the estimated age, 40 generations, is smaller than their estimate. The difference is a consequence of assuming a prior distribution for s and using it in the Bayesian analysis. Saunders *et al.* (2005) found that the likelihood function depends only weakly on s for $s > 0{\cdot}05$. They used $s = 0{\cdot}05$ as a conservative estimate and then estimated t_1 from the posterior distribution of t_1, given that value of s. The result is an overestimate of t_1 and an underestimate of s. In contrast, the estimates obtained here are from a posterior distribution that has a single maximum, albeit one that lies on an evident ridge.

In the G6PD example, no population growth was assumed. If there actually had been exponential growth, then the estimated selection coefficient would estimate $r + s$ because population growth has almost the same effect on linked markers as does positive

https://doi.org/10.1017/S0016672307008944 Published online by Cambridge University Press

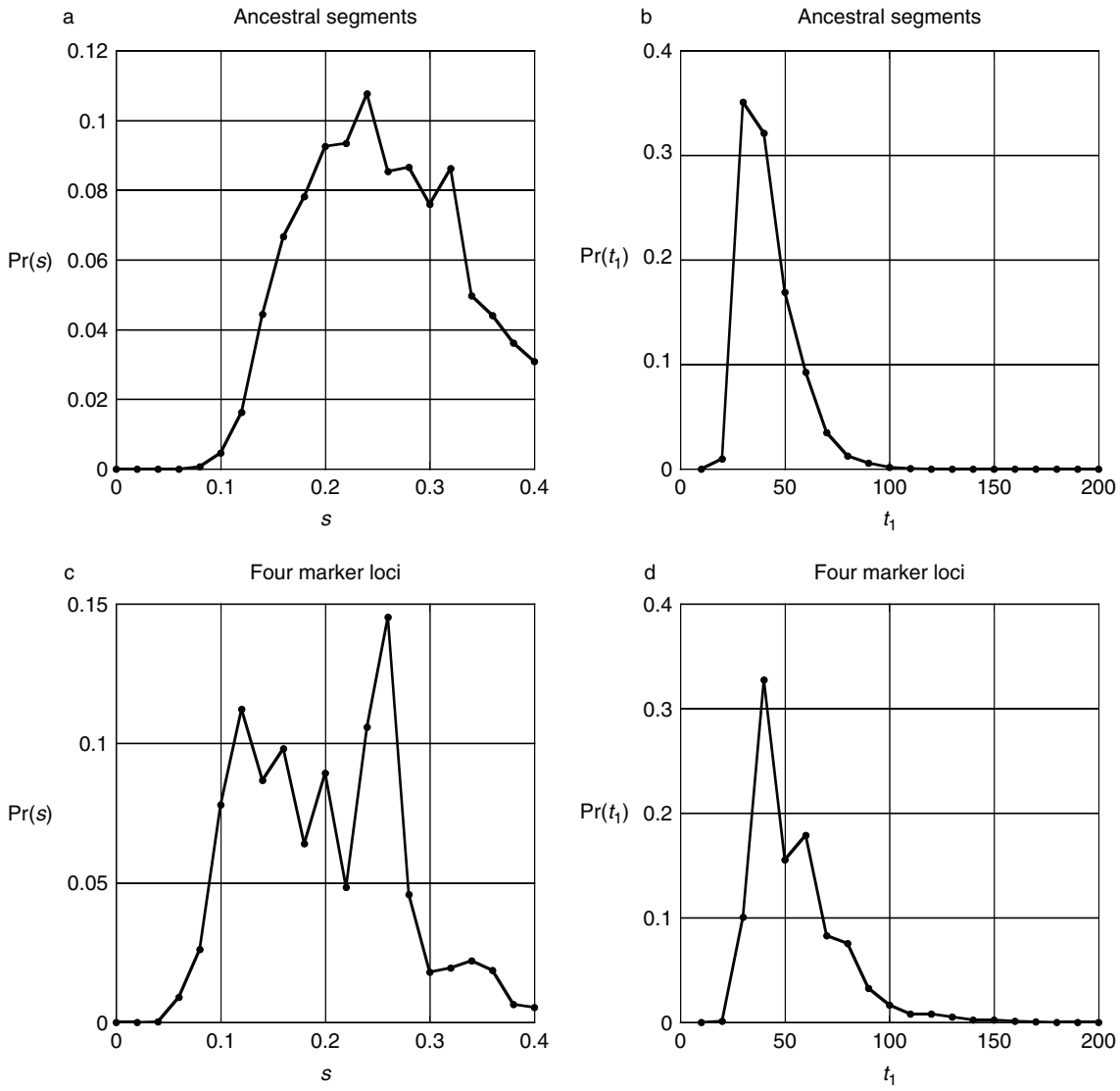

Fig. 3. Marginal distributions of s and t_1 obtained by summing over the joint distributions. (*a*) and (*b*) are based on the data presented in Fig. 2*a* and (*c*) and (*d*) are based on the data presented in Fig. 2*b*.

selection. However, exponential growth would affect all loci in the same way, resulting in more LD across the genome. The evidence for positive selection at G6PD comes from the fact that the extent of LD with the A− locus is unusually large.

In the general formulation of the birth–death approximation, only exponential growth at rate r was considered in order to order to minimize the number of parameters in the model. The method could be adapted to other models of demographic history, including population subdivision and bottlenecks in population size. As in the case of exponential growth, evidence of selection comes from LD with the allele of interest that exceeds background levels of LD created by demographic processes.

The Bayesian analysis described above assumes a prior distribution of selection intensities, but treats the observed mutant frequency (x) as a fixed parameter. It would be possible to develop a Bayesian method in which a prior distribution of x is derived from assumptions about mutation, selection, drift and other factors. Such an approach, however, would not take account of the way that mutants are chosen for detailed genetic analysis of the type done by Saunders *et al.* (2005). Until the criteria for carrying out such studies can be quantified, it seems difficult to extend the Bayesian framework to include a prior distribution of x as well.

This research was supported in part by Grant GM40282 from the US National Institutes of Health. I thank B. Rannala for helpful discussions of this topic and the reviewers for helpful comments on a previous version of this paper.

Appendix A. Approximations based on the linear birth–death process

(i) *Prior distribution of allele age*

The prior distribution is itself derived from a Bayesian formulation initially used by Kimura & Ohta (1973). The probability that an allele M of frequency x_0 at $t=0$ is at frequency x at a later time $t=T$ can be written as a transition function $\phi(x, T|x_0, 0)$ that depends on the population sizes and selection intensities between 0 and T. The prior distribution of allele age is obtained from this transition function by assuming a constant mutation rate μ. The probability that a mutation arises t_1 generations in the past is $2\mu N(-t_1)$ where $N(-t_1)$ is the population size at that time. The probability that M arises t_1 generations in the past and is found at frequency x today is

$$2\mu N(t_1)\phi(x, 0|x_0, -t_1)$$

where x_0 is the frequency of a new mutant, $1/[2N(-t_1)]$. Therefore,

$$\Pr(t_1)=\frac{N(-t_1)\phi(x, 0|x_0, -t_1)}{\int_0^\infty N(-t_1)\phi(x, 0|x_0, -t_1)dt_1}. \quad \text{(A1)}$$

For neutral alleles in a population of variable size and for arbitrary x_0 and x, ϕ can be obtained as the solution to a diffusion equation that approximates the Wright–Fisher, Moran and other models of gene frequency evolution. For selected alleles, there is no usable solution for ϕ even in a population of constant size. If x_0 and x are both small, the problem can be simplified because the number of copies of M can be modelled by a linear birth–death process that allows for both exponential population growth and additive selection. (Slatkin, 2002*b*; Wiuf, 2001*b*). In this case, analytical expressions for $\Pr(t_1)$ are available. If there are N_0 individuals in the population at the present time and n chromosomes are chosen at random and tested for M, the fraction of the population sampled is $f=n/(2N_0)$ and the number of copies of M in the sample is $i=2N_0x$. The population is assumed to have grown exponentially at rate r and s is the additive selection coefficient in favour of M. The growth rate r is non-negative, but s may take any value.

I have presented elsewhere distributions for all combinations of r and s (Slatkin, 2002*b*). Here I include only those formulas used to obtain the results presented in the text. If $r=s=0$, the prior probability that M arose t generations in the past is

$$\Pr(t)=\frac{2if^it^{i-1}}{(2+ft)^{i+1}}; \quad \text{(A2a)}$$

if $r=0$ and $s>0$,

$$\Pr(t)=\frac{2if^is^2e^{-st}(1-e^{-st})^{i-1}}{(f+(f-2s)e^{-st})^{i+1}}; \quad \text{(A2b)}$$

if $r>0$ and $s>0$,

$$\Pr(t)=\frac{\xi f^{i-1}e^{-(\xi+r)t}(1-e^{-\xi t})^{i-1}}{(f+(f-2\xi)e^{-\xi t})^{i+1}\Gamma(j)_2F_1(1+r/\xi, 1+i, 1+r/\xi+i;1-\xi/f)}, \quad \text{(A2c)}$$

where $\xi=r+s$, $\Gamma(\,)$ is the gamma function and ${}_2F_1(\,)$ is the hypergeometric function (Abramowitz & Stegun, 1965). In these equations, t is used instead of t_1 for notational convenience. Note that in equations (A2b) and (A2c), $\Pr(t)$ depends on s and t separately as well as on the product st.

(ii) *Intra-allelic coalescence times*

As shown by Slatkin & Rannala (1997), the coalescence times for the intra-allelic genealogy are easy to generate if the number of copies of the mutant follows a linear birth–death process. A random set of coalescence times $\{t_2, \ldots, t_i\}$ is obtained by generating $i-1$ random variables from a kernel distribution, $b(t|t_1)$, and then sorting them in decreasing order. If $r=s=0$,

$$b(t|t_1)=\frac{2(2+ft_1)}{t_1(2+ft)^2} \quad \text{(A3a)}$$

and if $\xi=r+s>0$,

$$b(t|t_1)=\frac{2\xi^2e^{-\xi t}}{(f-(f-2\xi)e^{-\xi t})^2}\cdot\frac{f-(f-2\xi)e^{-\xi t_1}}{1-e^{-\xi t_1}}. \quad \text{(A3b)}$$

If the intra-allelic genealogy were a star, then $t_2=\ldots=t_i=t_1$. In this model, that would be equivalent to $b(t|t_1)$ being a spike (i. e. a Dirac δ function) at $t=t_1$. The extent to which $b(t|t_1)$ is not a spike indicates how much the intra-allelic genealogy differs from a star.

Appendix B. Estimating the probability of the data by using importance sampling

The method described in the text requires the calculation of $\Pr(G|t_1, x, r, s)$, which is the probability of the genetic data (G), given the age of M (t_1), the frequency of M (x), the population growth rate (r) and the selection intensity (s). The exact calculation requires the evaluation of equation (2) in the text. Because the number of branching patterns of a genealogy increases very rapidly with the number of terminal branches, it becomes impractical to evaluate the sum exactly for a sample size larger than 3 or 4. To avoid this problem, I use an importance sampling method introduced previously (Slatkin, 2002*a*) that performs well when the intra-allelic genealogy is generated by a linear birth–death process. The idea is to sample

https://doi.org/10.1017/S0016672307008944 Published online by Cambridge University Press

branching patterns in such a way that patterns that contribute more to the overall probability of the data are sampled more frequently. By generating a large number of trees and taking the average, an estimate of the probability of the data is obtained:

$$\Pr(G|t_1, x, r, s) \approx \frac{1}{H}\sum_{h=1}^{H}\frac{\Pr_{RS}(B_h)}{\Pr_{IS}(B_h)}\Pr(G|B_h, t_2, \ldots, t_i, F), \tag{B1}$$

where H is the total number of replicates, $\Pr_{RS}(B_h)$ is the probability of branching pattern B_h under a random choice of branching patterns, $\Pr_{IS}(B_h)$ is the probability of B_h under the importance sampling method used and F is the forward transition matrix for the genetic data in one generation. This transition matrix has to be specified for each type of data. In this paper, I consider two models: one with K loci and an arbitrary recombination map and the other of the length of the ancestral chromosome associated with each copy of M.

(iii) *Multiple marker loci*

There are assumed to be K diallelic marker loci linked to M. At each locus, allele 0 is assumed to have been on the ancestral M chromosome. The 2^K haplotype frequencies on the non-M chromosomes are assumed to be known and fixed. The genetic map is specified by c, the recombination rate between M and locus 1. If $c<0$, locus 1 is to the left of M; if $c>0$, locus 1 is the to right. The recombination rate between locus k and $k+1$ is c_k ($k=1, \ldots, K-1$). Therefore, M is the leftmost locus if $c>0$ and M is the rightmost locus if $-c>\sum_{k=1}^{K-1}c_k$. Otherwise M is between one pair of marker loci. The number of M-bearing chromosomes is assumed to be so small that only recombination events between an M-bearing and a non-M-bearing chromosome are considered. There is no interference in recombination and the recombination rates are assumed to be sufficiently small that at most one crossing-over occurs each generation.

With these assumptions, it is straightforward to find the elements of F, the forward transition matrix for the 2^K states of the marker loci on an M-bearing chromosomes. Each off-diagonal element of F represents a transition from one marker haplotype to another. The transition probability is the sum of the probabilities of recombination events between each pair of marker loci. For example, if $K=4$, the probability of a transition from (1, 0, 1, 0) to (1, 0, 1, 1) can result from a recombination event between markers 2 and 3 bringing in the 1, 1 haplotype at markers 3 and 4 and a recombination event between markers 3 and 4 bringing in the 1 allele at marker 4. The diagonal elements are obtained by subtracting from 1.

(iv) *Ancestral chromosome lengths*

If the lengths of ancestral chromosomes linked to each copy of M are known, it is possible to derive an analytical expression for an arbitrary power of F, the forward transition matrix. The chromosomal segments on each side of M are considered separately and then the probabilities are multiplied to obtain the overall probability. On one side of M, assume there are L sites and the recombination rate between each pair of sites is c. The parameter L is chosen to be large enough that the results do not depend on its value. The state of a segment is k, the length of the ancestral segment linked to M ($k=0, 1, \ldots, L$).

The elements of F are probabilities of transitions from state k in generation t to state k' in generation $t+1$: $F_{kk'}=0$ if $k'>k$, $F_{kk'}=(1-c)^k$ if $k'=k$ and $F_{kk'}=c(1-c)^{k'}$ if $k'<k$. It is straightforward to verify that the elements of F^t are 0 if $k'>k$, $(1-c)^{kt}$ if $k'=k$ and $(1-c)^{k't}-(1-c)^{(k'+1)t}$ if $k'<k$.

Appendix C. Positive or negative selection

The theory of linear birth–death processes provides the probability that j alleles are found at time t, given one copy at time 0, a birth rate λ and a death rate μ:

$$\Pr(j|t)=u(1-u)^{j-1} \tag{C1}$$

($j>0$), where

$$u=\frac{\lambda\left(1-e^{-(\lambda-\mu)t}\right)}{\lambda-\mu e^{-(\lambda-\mu)t}} \tag{C2}$$

(Kendall, 1948). If a fraction f of the population is sampled with replacement, then the probability that i copies of the allele are present in the sample is obtained by multiplying (C1) by a binomial distribution with probability f and sample size j and summing over j to obtain:

$$\Pr(i|t)=u'(1-u')^{i-1} \tag{C3}$$

where $u'=fu/[1-(1-f)u]$.

Slatkin & Rannala (1997) and Wiuf (2001*b*) showed that the linear birth–death process approximates the dynamics of a rare allele in a population of size N growing exponentially at rate r and subject to additive selection of intensity s if $\lambda=1/2$ and $\mu=1/2-r-s$.

Given that i copies of a mutant are observed at time t, the relative likelihood that a mutant has a selective advantage s instead a selective disadvantage of s is the ratio $\Pr(i|s, t)/\Pr(i|-s, t)$. This ratio is very large if s is not small. In the example used in the text, in which $i=20$, $f=9{\cdot}1\times10^{-4}$ and $r=0$, the ratio is approximately $1{\cdot}4\times10^{16}$ if $s=0{\cdot}05$ and increases rapidly as s increases.

https://doi.org/10.1017/S0016672307008944 Published online by Cambridge University Press

References

Abramowitz, M. & Stegun, I. A. (eds.) (1965). *Handbook of Mathematical Functions*. New York: Dover.

Charlesworth, D. (2006). Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics* **2**, 379–384.

Charlesworth, D., Charlesworth, B. & McVean, G. A. T. (2001). Genome sequences and evolutionary biology: a two-way interaction. *Trends in Ecology & Evolution* **16**, 235–242.

Coop, G. & Griffiths, R. C. (2004). Ancestral inference on gene trees under selection. *Theoretical Population Biology* **66**, 219–232.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics* **22**, 521–565.

Goldstein, D. B., Reich, D. E., 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. *American Journal of Human Genetics* **64**, 1071–1075.

Guo, S. W. & Xiong, M. (1997). Estimating the age of mutant disease alleles based on linkage disequilibrium. *Human Heredity* **47**, 315–337.

Hartl, D. L. & Clark, A. G. (1997). *Principles of Population Genetics*. Sunderland, MA: Sinauer Associates.

Kaplan, N. L., Lewis, P. O. & Weir, B. S. (1994). Age of the DF508 cystic fibrosis mutation. *Nature Genetics* **8**, 216.

Kendall, D. G. (1948). On some modes of population growth leading to R. A. Fisher's logarithmic series distribution. *Biometrika* **35**, 6–15.

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

Maruyama, T. (1974). The age of an allele in a finite population. *Genetical Research* **23**, 137–143.

McPeek, M. S. & Strahs, A. (1999). Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *American Journal of Human Genetics* **65**, 858–875.

Neuhausen, S. L., 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. *American Journal of Human Genetics* **58**, 271–280.

Nordborg, M. & Tavaré, S. (2002). Linkage disequilibrium: what history has to tell us. *Trends in Genetics* **18**, 83–90.

Rannala, B. & Reeve, J. P. (2001). High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. *American Journal of Human Genetics* **69**, 159–178.

Reich, D. E. & Goldstein, D. B. (1999). Estimating the age of mutations using variation at linked markers. In *Microsatellites: Evolution and Applications* (ed. D. B. Goldstein & C. Schlötterer), pp. 129–138. Oxford: Oxford University Press.

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. *Nature Genetics* **9**, 152–159.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., *et al*. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837.

Saunders, M. A., Slatkin, M., Garner, C., Hammer, M. F. & Nachman, M. W. (2005). The span of linkage disequilibrium caused by selection on G6PD in humans. *Genetics* **171**, 1219–1229.

Slatkin, M. (2000). Allele age and a test for selection on rare alleles. *Philosophical Transactions of the Royal Society of London, Series B* **355**, 1663–1668.

Slatkin, M. (2001). Simulating genealogies of selected alleles in a population of variable size. *Genetical Research* **78**, 49–57.

Slatkin, M. (2002*a*). A vectorized method of importance sampling with applications to models of mutation and migration. *Theoretical Population Biology* **62**, 339–348.

Slatkin, M. (2002*b*). The age of alleles. In *Modern Developments in Theoretical Population Genetics*, pp. 233–259. Oxford: Oxford University Press.

Slatkin, M. & Bertorelle, G. (2001). The use of intra-allelic variability for testing neutrality and estimating population growth rate. *Genetics* **158**, 865–874.

Slatkin, M. & Rannala, B. (1997). Estimating the age of alleles by use of intraallelic variability. *American Journal of Human Genetics* **60**, 447–458.

Slatkin, M. & Rannala, B. (2000). Estimating allele age. *Annual Review of Genomics and Human Genetics* **1**, 225–249.

Stephens, J. C., Reich, D. E., Goldstein, D. B., Shin, H. D., Smith, M. W., *et al*. (1998). Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. *American Journal of Human Genetics* **62**, 1507–1515.

Tishkoff, S. A., Dietzsch, E., Speed, W., Pakstis, A. J., Kidd, J. R., *et al*. (1996). Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**, 1380–1387.

Toomajian, C., Ajioka, R. S., Jorde, L. B., Kushner, J. P. & Kreitman, M. (2003). A method for detecting recent selection in the human genome from allele age estimates. *Genetics* **165**, 287–297.

Vogel, F. & Motulsky, A. G. (1996). *Human Genetics: Problems and Approaches*. New York: Springer.

Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biology* **4**, 446–458.

Wiuf, C. (2001*a*). Do Delta-F508 heterozygotes have a selective advantage? *Genetical Research* **78**, 41–47.

Wiuf, C. (2001*b*). Rare alleles with selection. *Theoretical Population Biology* **59**, 287–296.

Wood, E. T., Stover, D. A., Slatkin, M., Nachman, M. W. & Hammer, M. F. (2005). The β-globin recombinational hotspot reduces the effects of strong selection around HbC, a recently arisen mutation providing resistance to malaria. *American Journal of Human Genetics* **77**, 637–642.