Inferring Selection Intensity and Allele Age from Multilocus Haplotype Structure

Hua Chen[1] and Montgomery Slatkin
Department of Integrative Biology, University of California, Berkeley, California 94720

ABSTRACT It is a challenging task to infer selection intensity and allele age from population genetic data. Here we present a method that can efficiently estimate selection intensity and allele age from the multilocus haplotype structure in the vicinity of a segregating mutant under positive selection. We use a structured-coalescent approach to model the effect of directional selection on the gene genealogies of neutral markers linked to the selected mutant. The frequency trajectory of the selected allele follows the Wright-Fisher model. Given the position of the selected mutant, we propose a simplified multilocus haplotype model that can efficiently model the dynamics of the ancestral haplotypes under the joint influence of selection and recombination. This model approximates the ancestral genealogies of the sample, which reduces the number of states from an exponential function of the number of single-nucleotide polymorphism loci to a quadratic function. That allows parameter inference from data covering DNA regions as large as several hundred kilo-bases. Importance sampling algorithms are adopted to evaluate the probability of a sample by exploring the space of both allele frequency trajectories of the selected mutation and gene genealogies of the linked sites. We demonstrate by simulation that the method can accurately estimate selection intensity for moderate and strong positive selection. We apply the method to a data set of the *G6PD* gene in an African population and obtain an estimate of 0.0456 (95% confidence interval 0.0144−0.0769) for the selection intensity. The proposed method is novel in jointly modeling the multilocus haplotype pattern caused by recombination and mutation, allowing the analysis of haplotype data in recombining regions. Moreover, the method is applicable to data from populations under exponential growth and a variety of other demographic histories.

KEYWORDS
selection coefficient
allele age
haplotype structure
structured coalescent
importance sampling
time-varying population size

INTRODUCTION

There is an increased interest in elucidating the role of natural selection in the evolution of human and other species using population genetic data. Evidence shows that selection has been actively shaping the genetic diversity of human populations during the process of adaptation to new environments and infectious diseases (Sabeti *et al.* 2002; Bersaglieri *et al.* 2004; Tishkoff *et al.* 2007; Simonson *et al.* 2010; Yi *et al.* 2010; Peng *et al.* 2011; Xu *et al.* 2011; Kamberov *et al.* 2013). Selection in human populations can leave "footprints" in patterns of single-nucleotide polymorphisms (SNPs) in the vicinity of the selected mutant. Numerous methods have been developed to detect natural selection based on such polymorphism patterns (Tajima 1989; Fu and Li 1993; Fay and Wu 2000; Kim and Stephan 2002; Sabeti *et al.* 2002; Nielsen *et al.* 2005; Voight *et al.* 2006; Sabeti *et al.* 2007; Tang *et al.* 2007; Chen *et al.* 2010). However, only a few methods are available for inferring quantities of the selective process, such as selection intensity and allele age. Among the existing methods, some consider single markers linked to the selected locus (*e.g.*, Slatkin 2001; Kim and Stephan 2002), whereas more sophisticated methods gain information by exploiting the haplotype structure of multiple marker loci. For example, Coop and Griffiths (2004) inferred selection intensity and allele age by analyzing mutations among different haplotypes along their genealogical history. Coop and Griffiths (2004) used an importance sampling algorithm to explore possible gene genealogies. Recombination is not allowed by their method, and thus it works only for nonrecombining regions. Rannala and Reeve (2004) extended their former likelihood approach for disease mapping (Rannala and Reeve 2001) to estimate allele age of a mutant under neutrality using multiple linked markers, and employed Markov Chain Monte Carlo to

doi: 10.1534/g3.113.006197
Manuscript received February 12, 2012; accepted for publication June 14, 2013

[1]Corresponding author: DOE Key Laboratory of Contemporary Anthropology, Fudan University, Shanghai 200433, China. E-mail: hchen007@gmail.com

generate the posterior distribution of allele age of the mutant. Slatkin (2008) presented a Bayesian method for jointly inferring selection intensity and allele age of the mutant using linkage disequilibria of multiple marker loci and generated the probability of data with an importance sampling algorithm.

The aforementioned multilocus methods all require modeling the effect of selection on the genealogical structure of neutral markers under a coalescent framework, which can be done in two ways. The first approach is to use the Krone-Neuhauser ancestral selection graph (ASG; Krone and Neuhauser 1997). In the ASG, the genealogy of the selected allele is embedded in a branching-coalescing graph so that both selection and mutation can be incorporated in the graph. The ASG approach is useful for simulating genealogies under weak selection. For moderate or strong selection, the ASG becomes so large that the computation becomes intractable. The ASG method's performance was dramatically improved by truncating the ASG (Slade 2000) to avoid generating very large ASGs. This approach, however, has not been extended to the analysis of multiple linked neutral mutations. The second approach is the structured coalescent (Kaplan *et al.* 1988; Hudson and Kaplan 1988), which generates historical frequency trajectories of the selected allele and then treats chromosomes carrying the mutant allele and nonmutant allele as subpopulations between which there exists "gene flow" caused by recombination. For alleles under balancing selection, the allele frequencies were assumed to be constant. For positively selected alleles, the allele frequency trajectories can either be generated by stochastic simulation or be approximated using deterministic equations.

The aforementioned multilocus methods all adopted the structured-coalescent model of selection to estimate the selection parameters (Coop and Griffiths 2004; Slatkin 2008), which is also the model used in our proposed method. But the approach to sampling the allele frequency trajectory from its probability distribution in our method differs from these others. Coop and Griffiths (2004) generated random trajectories of selected mutations under the Moran model, which has the property of time reversibility under mutation and additive selection (Watterson 1975) but works only for populations of constant size. Rannala and Reeve (2004) made an assumption that the historical allele frequencies of neutral markers are constant during the whole process, which may not hold for real populations, especially for markers under the hitch-hiking effect. Slatkin (2008) used a linear birth-and-death process to approximate the genealogical trees of haplotypes carrying selected mutants, which is an adequate approximation for mutants in low frequency, but may not be suitable for common mutants. We use the Wright-Fisher model instead of the Moran model to generate allele frequency trajectories, and apply the importance sampling scheme in Slatkin (2001) to weight the trajectories when estimating selection parameters. This allows us to model the selective sweeps in a population with time-varying size, and allows us to analyze both high- and low-frequency alleles under selection.

One advantage of the proposed method over the existing methods is that we model the dynamics of ancestral haplotypes under the joint effects of selection, mutation and recombination during a selective sweep process. In contrast, Coop and Griffiths (2004) assumed no recombination in their model, whereas both Slatkin (2008) and Rannala and Reeve (2004) simplified the transitions among different multi-loci haplotypes induced by recombination. In particular, some of the recombination events between different types of haplotypes within the selected haplotype group were ignored. This restriction can cause significant bias when the selected allele is in medium or high frequency in the population. Our method explicitly describes the frequencies of different selected haplotypes over time during the selective process. Therefore, our proposed method applies to both high- and low-frequency mutants. Our method also requires some approximations to improve computational efficiency.

As we will demonstrate in the section *A simplified multi-locus model for haplotype structure*, the model efficiently reduces the state space of ancestral haplotypes from an exponential function of the number of SNP loci to a quadratic function, and thus allows the inference of allele age and selection intensity from multi-SNP haplotypes spanning several hundred kilo-bases or even mega-bases affected by strong selections. We then modify the importance sampling method of Griffiths and Tavaré (1994b) to obtain the probability of a sample configuration and to estimate the selection parameters by averaging over genealogies of the linked sites. Note that an alternative choice is to adopt the existing importance sampling algorithms developed for multilocus ancestral recombination graph (ARG) under neutrality (Griffiths and Marjoram 1996; Fearnhead and Donnelly 2001) and incorporate the ARG into the structured-coalescent model. However, because of the large state space of genealogies that has to be explored by the importance sampling algorithms for a multilocus ARG, this approach is intractable on a genomic scale of hundreds of kilobases.

OVERVIEW OF THE METHOD

Suppose the data consist of n_{sample} haplotypes with known phase collected from the current population. If genotype data are collected, the phase can be estimated by available algorithms (Scheet and Stephens 2006). The haplotypes are divided into two groups: the selected haplotypes, which are the chromosomes carrying the selected allele, and the background haplotypes, which do not carry the selected allele. In this method, we view the coalescent process in the n selected haplotypes as in a structured subpopulation (see Figure 1). The n

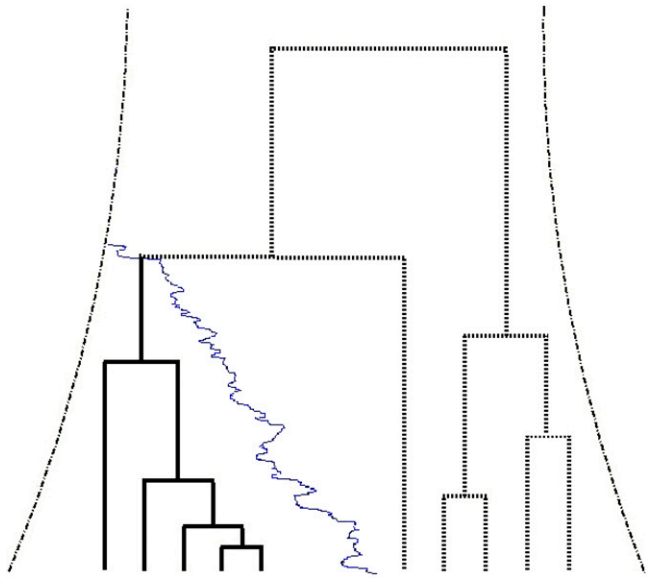

Figure 1 An illustration of the structured-coalescent approach for modeling positive selection. The historical population sizes are indicated by the distance between the two dashed lines; and the allele frequency trajectory of the selected allele is indicated with a thin solid curve. The coalescent history of the selected locus with five derived lineages (solid bold lines) and five ancestral lineages (dotted bold lines) is superimposed on the trajectory and population size curves. The present time, $t = 0$, is at the bottom. And the time at which the trajectory of the selected mutant merged to the population-size curve denotes the time when the selected mutant arose in the population, *i.e.*, the allele age T. In the model presented in the main text, only the sub-genealogies in the selected allele groups (bold solid lines) are considered.

selected haplotypes are represented by a n by m matrix with "1" or "0" for each entry, $\mathcal{D}_{i,j}$, corresponding to the allele type of the ith haplotype at the jth SNP position. The position of the selected mutant is assumed to be known, as is the genetic distance between the j_1th and j_2th SNP, $\{r_{j1j2}, 1 \leq j_1, j_2 \leq m\}$.

We model the effect of selection with the structured-coalescent scheme (see Figure 1). The likelihood function of the observed selected haplotype data, $\mathcal{D}$, can be computed from

$$\mathcal{L}(s) = \mathbb{P}(\mathcal{D}|s,\Gamma) = \iint \mathbb{P}(\mathcal{D}|\mathcal{G})\mathbb{P}(\mathcal{G}|\mathcal{H})\mathbb{P}(\mathcal{H}|s,\Gamma)d\mathcal{G}d\mathcal{H}, \quad (1)$$

where $\mathcal{G}$ denotes the genealogy and $\mathcal{H}$ the frequency trajectory of the selected mutant. Neither $\mathcal{H}$ nor $\mathcal{G}$ is observed directly. When constructing the likelihood function, they are often integrated out (Felsenstein 1988; Griffiths and Tavaré, 1994b; Kuhner *et al.* 1995).

The frequency trajectory of the selected mutant, $\mathcal{H}$, is a random process that follows the Wright-Fisher model and has a probability distribution, $\mathbb{P}(\mathcal{H}|s,\Gamma)$, that depends on the selection intensity s and the nuisance parameter set Γ, which includes all the other parameters related to the population history. Conditional on any given frequency trajectory, the sampling probability of the data is constructed by summing over all possible genealogical events: $\mathbb{P}(\mathcal{D}|\mathcal{H}) = \int \mathbb{P}(\mathcal{D}|\mathcal{G})\mathbb{P}(\mathcal{G}|\mathcal{H})d\mathcal{G}$.

For computationally efficient evaluation of the sampling probability, we propose a novel simplified multilocus model for the transition of different types of selected haplotypes (see the section *A simplified multilocus model for haplotype structure*). We compute the extent of the ancestral haplotypes in the vicinity of the selected mutant under the combined effects of recombination and selection. Together with the infinitely-many-sites model for mutations, the simplified multilocus haplotype model is used to approximate the evolutionary dynamics of the data.

Because the spaces of both gene genealogies and allele frequency trajectories are too large to explore, $\mathbb{P}(\mathcal{G}|\mathcal{H})$ and $\mathbb{P}(\mathcal{H}|s,\Gamma)$ cannot be expressed in closed forms. We use the importance sampling algorithms to sample genealogies and trajectories that are compatible with the data from the proposal distributions. Then the likelihood is estimated as the weighted average probability for the samples. The importance weights are obtained by taking the ratio of the probabilities of true distribution and the proposal distribution. The procedure of evaluating the likelihood is illustrated by the flowchart in Figure 2. Two main steps of the flowchart correspond to sampling $\mathcal{H}$ and $\mathcal{G}$ using the importance sampling algorithms. The details of the algorithms and calculation of the importance ratios are presented in the section *Importance sampling and proposal distributions*.

A SIMPLIFIED MULTILOCUS MODEL FOR HAPLOTYPE STRUCTURE

We model the transition of the selected haplotypes (haplotypes carrying the selected mutant) under the influence of evolutionary events including recombination and mutation. We start with a sample of selected haplotypes collected from the current generation. When looking backward in time, we can eventually trace these selected haplotypes to one common ancestor (the ancestral haplotype) because all copies of the selected allele are descended from a single mutation. During the selection process, recombination breaks up and mixes the fragments with the background haplotypes. Recombination combined with mutation generates the different selected haplotypes that contain some segments of the ancestral haplotype. The number of distinct selected haplotypes at different times in the history is called the ancestral process. This ancestral process, conditional on the frequency trajectory of the selected mutant, can be viewed as a structured-coalescent process, because the selected haplotypes evolve as a subpopulation of the entire haplotype pool, with the size of the subpopulation determined by the mutant allele frequency, and the transitions among different haplotypes following the simplified multilocus model.

To illustrate the state space of the ancestral process and the joint effect of selection and recombination on the transitions between the ancestral states, we will start with a simple case of only two loci, the selected locus and a partially linked SNP locus. We then extend the two-locus model to a simplified multi-locus haplotype model, after making several approximations for computational efficiency. Then in the section *Sampling probability of a multi-locus haplotype configuration*, the simplified multilocus haplotype model and the infinitely-many-sites model for mutation are used to derive the sampling probability for haplotype configuration of a sample by summing over possible ancestral states of the genealogical history, that is, the $\sum_G \mathbb{P}(\mathcal{D}|\mathcal{G})\mathbb{P}(\mathcal{G}|\mathcal{H})$ component of likelihood function conditional on a simulated allele frequency trajectory.

A two-locus model

The two-locus haplotype model involves only the selected locus and one neutral marker, the positions of which are assumed known. The selected locus has the mutant allele A and the other neutral allele a. The neutral marker locus has two alleles B and b. Let $Q(t) = (q_1, q_2; q_3, q_4)$ denote the number of haplotypes AB, Ab, aB, and ab in the sample at time t. Conditional on the ancestral allele frequency trajectory, $\{X_t, t > 0\}$, the "ancestral process" $Q(t)$, which is defined as the numbers of each haplotype, can be approximated by the inhomogeneous Markov process (Hudson and Kaplan 1988; Durrett and Schweinsberg 2004). The states that the process $Q(t)$ can jump from state $(q_1, q_2; q_3, q_4)$ to include $(q_1 - 1, q_2; q_3, q_4)$, $(q_1 - 1, q_2 + 1; q_3, q_4)$, $(q_1 - 1, q_2; q_3 + 1, q_4)$, $(q_1 - 1, q_2; q_3, q_4 + 1)$, $(q_1, q_2 - 1; q_3, q_4)$, $(q_1 + 1, q_2 - 1; q_3, q_4)$, $(q_1, q_2 - 1; q_3 + 1, q_4)$, $(q_1, q_2 - 1; q_3, q_4 + 1)$, $(q_1, q_2; q_3 - 1, q_4)$, $(q_1 + 1, q_2; q_3 - 1, q_4)$, $(q_1, q_2 + 1; q_3 - 1, q_4)$, $(q_1, q_2; q_3 - 1, q_4 + 1)$, $(q_1, q_2; q_3, q_4 - 1)$, $(q_1 + 1, q_2; q_3, q_4 - 1)$, $(q_1, q_2 + 1; q_3, q_4 - 1)$ and $(q_1, q_2; q_3 + 1, q_4 - 1)$. The transition probabilities from $(q_1, q_2; q_3, q_4)$ to the first four states are listed in Table 2, and the other transition probabilities can be constructed similarly. We assume an infinitely-many-sites model for mutations, so there are no new or recurrent mutations between the two alleles of either the selected locus or the neutral marker locus. Let $N_{AB}(t)$, $N_{Ab}(t)$, $N_{aB}(t)$, and $N_{ab}(t)$ be the population counts of the four corresponding haplotypes AB, Ab, aB and ab at time t respectively. For the transition from $(q_1, q_2; q_3, q_4)$ to $(q_1 - 1, q_2; q_3, q_4)$, no recombination has occurred and two lineages of haplotype AB are chosen to coalesce. The coalescence rate is $q_1(1-r)\frac{q_1-1}{N_{AB}(t)}$; for the transition from $(q_1, q_2; q_3, q_4)$ to $(q_1 - 1, q_2 + 1; q_3, q_4)$, one of the $N_{Ab}(t) + N_{ab}(t)$ lineages, which carry the b allele, must be chosen to recombine, and the rate is $q_1 r \frac{N_{Ab}(t)+N_{ab}(t)-q_2-q_4}{2N_t}$. And the other transition rates can be obtained by similar rationale. Note that the selected allele first entered into the population at time T, so the states of the embedded Markov chain should satisfy $Q(T) \in \{(1, 0; q_3, q_4), (0, 1; q_3, q_4)\}$ and $Q(t) = (0, 0; q_3, q_4)$, $t > T$.

A simplified multilocus model for haplotype structure

The model for two-locus haplotypes can be naturally extended to multilocus haplotypes. However, the extent of haplotype structure that is used to infer selection intensity and allele age usually spans a large region which covers several hundred kilobases, or even more than

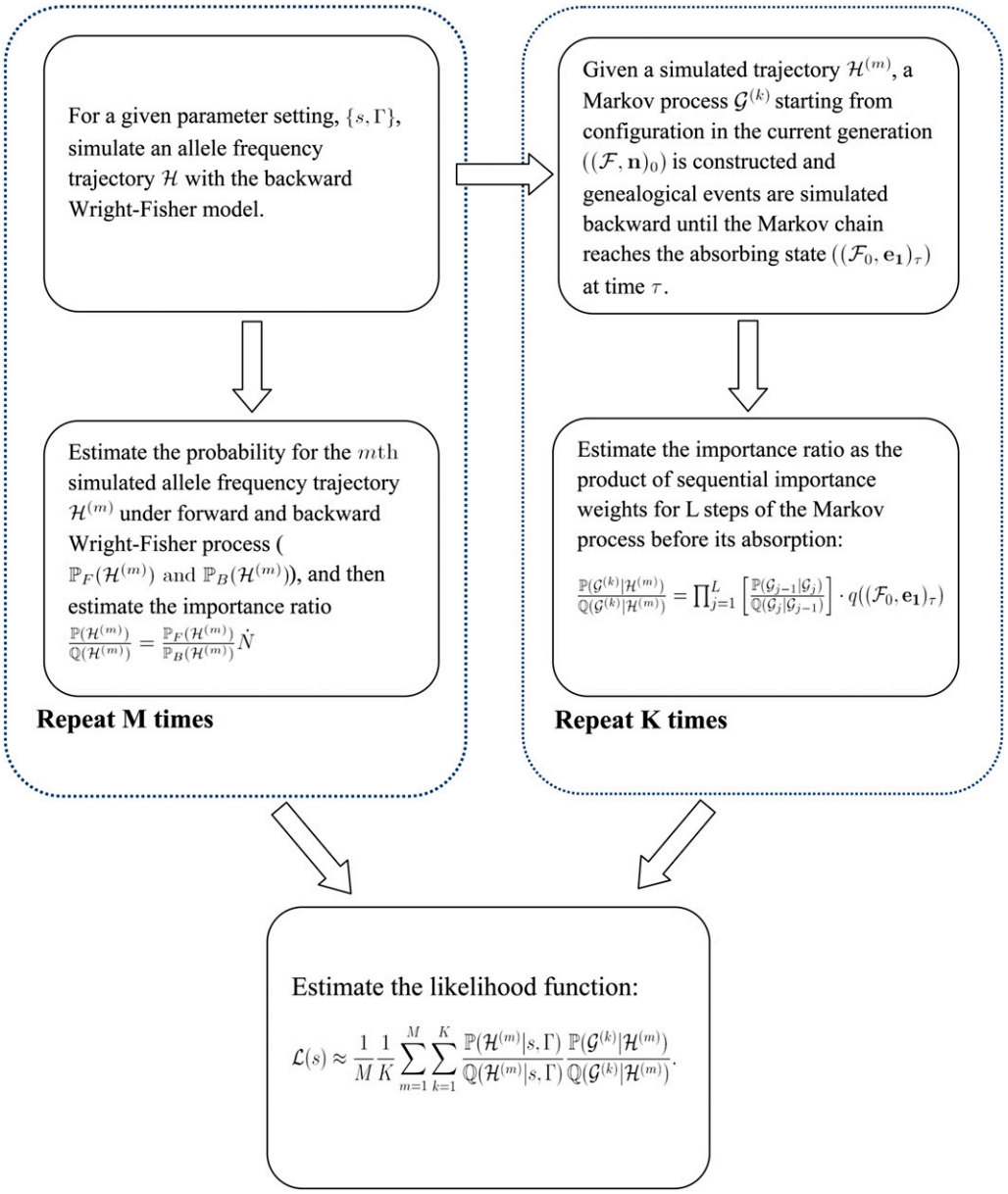

Figure 2 Flowchart of importance sampling procedures of the method.

a mega-base. In such a large region, there may be hundreds of polymorphic sites. When the number of SNP loci increases, the number of possible states of the ancestral process increases exponentially, and the transition matrix becomes so large that numerical evaluation becomes impossible. It is thus necessary to develop a parsimonious model for multilocus haplotypes that is both computationally fast and statistically efficient.

The novel multilocus model we present here exploits the extents of the "ancestral haplotypes" retained during the selection process. We use the term "ancestral haplotype" to refer to the alleles at each SNP position on the ancestral chromosome, and the term "background haplotypes" to refer to the other haplotypes. In this model, we consider the interplay between selection and recombination acting upon the ancestral haplotype. As the selected allele increases its frequency, recombination breaks up the ancestral haplotypes and mixes them with the background haplotypes, resulting in the sample we observe at present. We make several simplifications or assumptions to expedite the computation in the sections to follow.

The ancestral state of each position along the haplotypes is assumed known: For each position of a chromosome, it is assumed to be known whether the allele at that position is descended from an ancestral haplotype or one of the background haplotypes. In reality, the ancestral haplotype information cannot be observed directly from the data. The ancestral states and the break points of the ancestral haplotypes have to be inferred for each chromosome from patterns of SNP variation by other means (for example, the hidden Markov model for detecting recent positive selection, see Chen (2007)).

The haplotype structure of background haplotypes is ignored: It is reasonable to believe that the primary information for the inference of allele age and selection intensity comes from the extent of the ancestral haplotypes retained during selection. For example, states of the jump process for a two-locus haplotype model are reduced to $Q(t) = (q_1, q_2)$, and the absorbing states are now (1,0) and (0,1).

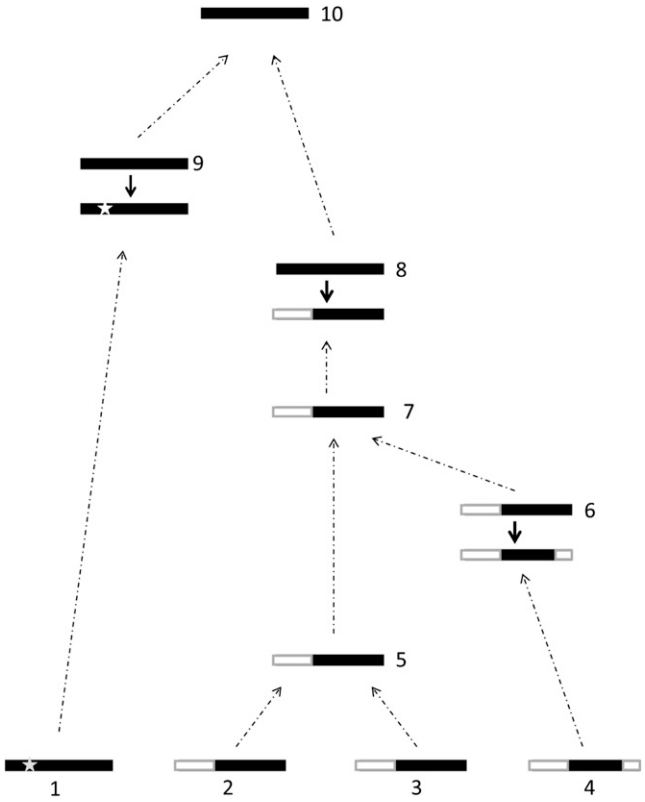

Figure 3 A realization of the genealogies for a sample of four haplotypes (lineages 1−4) to illustrate possible events in the genealogies. Black denotes the ancestral haplotype region (see the main text for the definition of "ancestral haplotypes"), and white denotes background haplotypes. A star denotes a neutral mutant arising on the ancestral haplotype. The present time, $t = 0$, is on the bottom. When going back in time, the events are coalescent (lineages 2 and 3 coalesce to the ancestral lineage 5), recombination (lineage 6 → lineage 4), coalescent (lineages 5 and 6 coalesce to the ancestral lineage 7), recombination (lineage 8 → lineage 7), mutation (on lineage 9), coalescent (lineages 8 and 9 coalesce to lineage 10) in sequence.

The population frequencies of ancestral haplotypes at time t are approximated using the expectations: To obtain the transition rates in Table 2 for a two-locus model, the population counts of haplotypes *AB*, *Ab*, *aB*, and *ab* at time *t*, $N_{AB}(t)$, $N_{Ab}(t)$, $N_{aB}(t)$, and $N_{ab}(t)$, are needed. For a multilocus haplotype model, these values correspond to the population frequencies of haplotypes at time *t*. In studies of fine-scale disease mapping, these allele frequencies were assumed to be constant over time and identical to the observed frequencies in the current population (Rannala and Reeve 2001). When there is selection, the allele frequencies of nonrecombined haplotypes change over time, and their expectations have to be derived using a deterministic model. These expectations will be used as an estimate of the true haplotype frequencies at any time *t*. The details of the equations for the haplotype frequencies at time *t* can be seen in Appendix A.

Multiple "migrations" of ancestral haplotype fragments between selected and neutral haplotypes are ignored: In the ancestral process of a two-locus haplotype model, there is a probability that a lineage of the neutral marker experiences two recombination events during the sweep process. In other words, the ancestral haplotype crosses over twice with a background haplotype at the marker position and the segment of the ancestral haplotype "migrates" to and then back from the group of neutral haplotypes. The probability for such events are small during a selective sweep, and thus are ignored (on order $\mathcal{O}\left(\frac{1}{\log(\alpha)^2}\right)$, where $\alpha = 2Ns$, see Etheridge *et al.* 2006).

Because of the assumptions (1)−(4), the state space of the ancestral process can be reduced by considering the unique pattern of ancestral haplotype lengths in combination with the occurrence of mutation since selection began. Because we have assumed that the ancestral states of SNPs along the chromosomes are known, for every haplotype, we can determine the break points of the ancestral haplotype caused by the recombination events nearest to the mutant, in addition to the locations where mutations occurred within each ancestral haplotype region. With this information known, the ancestral haplotype on each side of the mutant can be coded as follows: for every selected haplotype, we record the SNPs to the left and to the right sides that delimit the ancestral haplotype; if there are mutations within the ancestral haplotype regions, the positions of the mutants are also recorded and listed as "mutation coordinates" behind the two "recombination coordinates." In Table 3, we give an example of 10 selected haplotypes consisting of 25 SNPs, among which the ancestral haplotype regions are highlighted. The selected mutant is located at position 18 (shown in boldface type). The left end of the ancestral haplotype for the first haplotype is 7 to the left of the mutant, and the right end is 6 to the right, such that the first haplotype is recorded as (7, 6). For haplotype 3, the full code is (12, 7, 21) with a mutation occurring in position 21. By this coding rule, the configuration of the sample listed in Column 3 of Table 3 is summarized in Column 3.

For a recoded haplotype type $h = (R_1, R_2, M_1, \ldots, M_k)$, the first two entries, corresponding to the left and right break points of the ancestral haplotypes, are the "recombination coordinates" and the other entries are the "mutation coordinates." In this model, the transition among different haplotypes is caused by recombination and mutation. For the coded haplotypes consisting of only recombination coordinates, the transition among different haplotype types can occur only through recombination. If there are m_L loci to the left of the mutant and m_R loci to the right of the mutant, the total number of possible allele types is $(m_L + 1) \times (m_R + 1)$. The number of possible states is greatly reduced compared to a direct extension of the two-locus model, whose state space grows exponentially with number of SNPs.

In Table 4, we present a partial list of transition rates caused by recombination for a 4-locus haplotype model. Assume that the haplotype has 4 SNP loci, with the alleles on the ancestral haplotype being A, B, C, and D. A is the selected mutant, and the order of the four loci along the chromosome is the same as their alphabetical order. We use the notation [*ABCD*] to denote the intact segment of ancestral haplotype. And similarly [$AB - d$] indicates that the first two loci have the inherited ancestral haplotype of A and B, the allele of the third locus is arbitrary, and the fourth locus is a background haplotype. Examples in Table 4 show some of the one-step transition rates starting from state [*ABcd*]. For example, for haplotype [*ABcd*] to jump to [*ABCD*], one of the [*ABcd*] haplotypes should be chosen and recombination has to occur between allele B and c, in such a way that the chosen haplotype crosses over with haplotypes [*ABCD*], [*AbCD*] and [$a - CD$]. The one-step transition probability is: $r_{BC}\{P_{[ABCD]}(t) \cdot X_t + P_{[AbCD]}(t) \cdot X_t + P_{[a-CD]}(t) \cdot (1 - X_t)\}$, with X_t being the population allele frequency of allele *A* at time *t*, and $P_{[.]}(t)$ being the frequency of the haplotype in square parenthesis among either the selected haplotype or the background haplotype group, depending on the allele type carried by the particular haplotype at the selected mutant locus. Because of assumption (4) we made previously (also made by Durrett and Schweinsberg (2004)), the second

Table 1 Definitions of notations used in this article

Notation	Meaning
n_{sample}	Total number of haplotypes in the sample
n	Number of selected haplotypes
m	Number of SNPs of a sample
m_L and m_R	Number of SNPs on the left and right sides of the mutant
$D_{i,j} = 0$ or 1	The jth SNP of the ith haplotype
N_t	Population size at time t
T	Allele age, or the time when the mutant arose in the population
s	The selection coefficient
r	Recombination fraction of the haplotype
μ	Mutation rate of the haplotype
$\theta = 4N\mu$	The scaled mutation rate of the haplotype
$\rho = 4Nr$	The scaled recombination rate of the haplotype
β_j	The proportion of ancestral haplotype region as a fraction of the length of the jth haplotype
$\mathcal{H} = \{I_T, I_{T-1}, \ldots, I_1, I_0\}$	The allele frequency trajectory
I_t	The number of the selected allele in the whole population at time t
$X_t = I_t/(2N_t)$	The frequency of the selected allele at time t
$h_1 = (R_1, R_2, M_1, \ldots, M_k)$	A recoded haplotype which includes two recombination Coordinates and k mutation coordinates
$\mathcal{T} = \{h_1, \ldots, h_d\}$	The d different haplotypes of a sample
$\mathbf{n} = \{n_1, \ldots, n_d\}$	The number of haplotypes for each haplotype group in $\mathcal{T}$
$(\mathcal{T}, \mathbf{n})_t$	The sample configuration at time t
$q((\mathcal{T}, \mathbf{n})_t)$	Sampling probability of the sample configuration $(\mathcal{T}, \mathbf{n})$ at time t
$\mathbf{e_j} = (0, 0, \ldots, 1, \ldots, 0)$	The jth unit vector
$\gamma(v,n) = \binom{n}{2}(\lambda_v X_v)^{-1} + n\theta/2 + n\rho/2$	The total rate for events at time v
$\lambda_t = N_t/N_0$	The ratio of population size at t to that at the present
$\mathcal{S}h_k$	$\mathcal{S}$ denotes a shift operator, and $\mathcal{S}h_k$ denotes deleting the first mutation coordinate of the kth haplotype
$\mathcal{C}h_k$	$\mathcal{C}$ denotes a coordinate change operator, and $\mathcal{C}h_k$ denotes changing one of the two recombination coordinates of the kth haplotype and eliminating all mutation coordinates outside the ancestral regions delimited by the new recombination coordinates
$\mathcal{R}_k\mathcal{T}$	The deleting operator that deletes h_k haplotype from $\mathcal{T}$
$\mathcal{L}(s)$	Likelihood function of the data
$\mathcal{G}^{(m)}$	The mth genealogical history, which consists of multiple steps of events, including recombination, mutation and coalescences
$\mathbb{Q}(\mathcal{H})$	Proposal distribution for $\mathcal{H}$ in the importance sampling algorithm
$\mathbb{Q}(\mathcal{G}\|\mathcal{H})$	Proposal distribution for $\mathcal{G}$ conditional on a $\mathcal{H}$ in the importance sampling algorithm

and the third terms are small and can be ignored. With this simplification, the transition rate in Table 4 becomes $r_{BC} \cdot P_{[ABCD]}(t) \cdot X_t$. Similarly, we can simplify the other transition rates shown in Table 4.

A mutation coordinate records the SNP position at which the haplotype has an allele mutated from the ancestral haplotype. We assume an infinitely-many-sites model for mutations on the haplotypes. According to Griffiths and Tavaré (1995), the set of nonrecombining haplotypes carrying these mutations is identical to a rooted gene tree, and the sequence of mutations corresponds to the path from the haplotype to the common ancestor [the root of the gene tree, see Griffiths and Tavaré (1995) for a detailed discussion]. We use the same notation scheme for mutations as Griffiths and Tavaré (1995). Because the haplotypes we investigate are from recombining regions, an additional constraint is added to reflect the effect of recombinations: the sequence of mutations we recorded as mutation coordinates includes only those located between the two recombination

Table 2 Possible transitions from $(q_1, q_2; q_3, q_4)$ and the rates for the two-locus haplotype model

Transition	Rate
$(q_1, q_2; q_3, q_4) \to (q_1 - 1, q_2; q_3, q_4)$	$q_1(1-r)\dfrac{(q_1-1)}{N_{AB}(t)}$
$(q_1, q_2; q_3, q_4) \to (q_1 - 1, q_2 + 1; q_3, q_4)$	$q_1 r \dfrac{N_{Ab}(t) + N_{ab}(t) - q_2 - q_4}{2N_t}$
$(q_1, q_2; q_3, q_4) \to (q_1 - 1, q_2; q_3 + 1, q_4)$	$q_1 r \dfrac{N_{AB}(t) + N_{aB}(t) - q_3}{2N_t}$
$(q_1, q_2; q_3, q_4) \to (q_1 - 1, q_2; q_3, q_4 + 1)$	$q_1 r \dfrac{N_{Ab}(t) + N_{ab}(t) - q_2 - q_4}{2N_t}$

Table 3 An example of haplotype configuration to demonstrate the coding rules used to denote the haplotype structure

Haplotype Type	Count Number	Code
1111110011**0**0000000**0**0000001	5	(7,6)
1111111111111**0**0000**0**0000000	1	(4,7)
11111**0**0000000000000**0**0010000	1	(12,7,21)
0000000000000010**0**0000000	3	(17,7,15)

There are 10 haplotypes with 25 single-nucleotide polymorphisms in four distinct groups in the sample. The mutant is located in position 18 and shown in boldface type. The ancestral region for each haplotype is highlighted. The codes for the four haplotype groups are listed in the third column

coordinates, that is, the subset of mutations on the retained ancestral haplotype region. For haplotype data from nonrecombining regions, the recombination coordinates are identical for all haplotypes (the left and right ends of the whole haplotype), and the mutation coordinates define a gene tree with the rooted genealogy, meaning that the state of the common ancestor of the sample is known (Griffiths and Tavaré, 1994b). This is the type of data analyzed by the approach of Coop and Griffiths (2004). Thus their method can be viewed as a special case of our method with no recombination.

Sampling probability of a multilocus haplotype configuration

In the section *A simplified multilocus model for haplotype structure*, we described a novel simplified multilocus model that can dramatically reduce the state space of the haplotype ancestral process, and illustrated how to obtain the transition probabilities between different states. We now consider the computation of the probability of a sample of multilocus haplotypes.

A sample of selected haplotypes can be coded and summarized by the rules introduced in the section *A simplified multi-locus model for haplotype structure* and grouped into d distinct groups $\mathcal{T} = \{h_1, \ldots, h_d\}$ with the corresponding multiplicities $\mathbf{n} = \{n_1, \ldots, n_d\}$. We define the sampling probability, $q((\mathcal{T}, \mathbf{n})_t)$, to be the probability of observing the sample configuration $(\mathcal{T}, \mathbf{n})$ at t generations before the current generation. The entire history of the sample configuration $\{(\mathcal{T}, \mathbf{n})_t, t > 0\}$ can be described by a Markov process that starts at time $t = 0$ and continues until reaching the absorbing state $(\mathcal{T}, \mathbf{e_1})$ at a random time τ, where $\mathbf{e_j}$ denotes the unit vector $\mathbf{e_j} = (0,0, \ldots, 1, \ldots, 0)$ with only the jth entry being 1. When $t = 0$, $q((\mathcal{T}, \mathbf{n})_0)$ is the partial likelihood of the data which is sufficient for the inference of selection intensity and allele age. The sampling probability at time t can be obtained by recursively summing over all possible state paths in the backward Markov process. The recursive formula can be written as

$$q((\mathcal{T},\mathbf{n})_t) = \int_t^\infty \sum_{(\mathcal{T}',\mathbf{n}')} p((\mathcal{T},\mathbf{n})_t|(\mathcal{T}',\mathbf{n}')_v) q((\mathcal{T}',\mathbf{n}')_v) g(v|n,t)dv, \tag{2}$$

where $p((\mathcal{T}, \mathbf{n})_t|(\mathcal{T}', \mathbf{n}')_v)$ is the transition probability of jumping from state $(\mathcal{T}', \mathbf{n}')$ at time v to state $(\mathcal{T}, \mathbf{n})$ at time t, and $g(v|n, t)$ is the density function of the inter-arrival time to the next event given an event at time t.

As the Markovian ancestral process is restricted to the selected haplotypes, the process behaves as if in a population with temporally varying size $\{I_t, t = 0.1, \ldots, T\}$. If time is measured in a scale of $2N_0$ generations, the coalescent rate is $\binom{n}{2}(\lambda_t X_t)^{-1}$, with $X_t = I_t/2N_t$ and $\lambda_t = N_t/N_0$ being the population size ratio. With the same scaling, the mutation rate for the ith haplotype is $\theta_i = 4N_0\mu\beta_i = \theta\ \beta_i$, and the recombination rate is $\rho_i = 4N_0c\beta_i = \rho\beta_i$. $\theta = 4N_0\mu$ and $\rho = 4N_0c$ are the scaled mutation rate and recombination rate for the whole haplotype, and β_i denotes the proportion of the retained ancestral haplotype region, or the inter-region between two recombination coordinates, out of the entire length of the ith haplotype. Note that β_i changes over time with the change of recombination coordinates of the ith haplotype. The inter-arrival time to the next event, v, given that the last event happened at time t has a non-homogeneous exponential distribution, with the density function in the form of

$$g(v|n,t) = \gamma(v,n)\exp\left(-\int_t^v \gamma(u,n)du\right), \tag{3}$$

$t < v < \infty$, where at time v, $\gamma(v,n) = \binom{n}{2}(\lambda_t X_v)^{-1} + \sum_{i=1}^n \theta_i/2 + \sum_{i=1}^n \rho_i/2$ is the rate for the any events. As the allele frequency trajectory $\{X_t,\ 0 \le t \le T\}$ is a discrete-time random process following the Wright-Fisher model, we adopt the geometric distribution for discrete time instead of using the continuous approximation in Equation 3:

$$g(v|n,t) = \gamma(v,n)\times \prod_{u=t+1}^{v-1}\big(1-\gamma(u,n)\big). \tag{4}$$

Conditional on an event happening at time v, the probabilities for the event being a mutation, recombination or coalescent are respectively

$$\frac{\sum_{i=1}^n \theta_i/2}{\gamma(v,n)},\quad \frac{\sum_{i=1}^n \rho_i/2}{\gamma(v,n)} \quad \text{and} \quad \frac{\binom{n}{2}\frac{1}{\lambda_v X_v}}{\gamma(v,n)}. \tag{5}$$

If a mutation occurs, one of the lineages in the sample is chosen to mutate into other types according to the mutation model, and the mutation coordinate of that haplotype is modified correspondingly; if a coalescence event occurs within the jth haplotype group, two of the existing lineages with haplotype h_j are chosen at random to coalesce, and the number of lineages in the jth haplotype group, n_j, is decreased by 1; otherwise, a position along the haplotype is chosen for the recombination event to occur with the consequence that one of the recombination coordinates is changed to record the recombination at that position (see Figure 3 for a realization of the genealogical history for a sample of four haplotypes).

We now present the detailed recursion equation (Equation 6), expressed as a sum over the above three types of events in a way corresponding to Equation 2 for our model. Under the infinitely-many-sites mutation model (Watterson 1975) and the proposed multilocus haplotype model for the extent of ancestral haplotypes, summing over possible one-step configuration changes at time v leads to the following equation:

Table 4 The transition probabilities for some states of the multilocus haplotype model

Transition	Rate
$[ABcd] \to [ABCD]$	$r_{BC}\{P_{\lvert ABCD\rvert}(t)\cdot X_t + P_{\lvert AbCD\rvert}(t)\cdot X_t + P_{\lvert a-CD\rvert}(t)\cdot(1-X_t)\}$
$[ABcd] \to [ABCd]$	$r_{BC}\{P_{\lvert ABCd\rvert}(t)\cdot X_t + P_{\lvert AbCd\rvert}(t)\cdot X_t + P_{\lvert a-Cd\rvert}(t)\cdot(1-X_t))\}$
$[ABcd] \to [Abcd]$	$r_{AB}\{P_{\lvert Abcd\rvert}(t)\cdot X_t + P_{\lvert aBcd\rvert}(t)\cdot(1-X_t)\}$

$$\ell_q\big((\mathcal{T},\mathbf{n})_v\big) = \sum_{k:n_k\geq 2}\frac{n(n_k-1)\frac{1}{\lambda_v X_v}}{2\gamma(v,n)}q\big((\mathcal{T},\mathbf{n}-\mathbf{e_k})_v\big)$$
$$+\sum_{\substack{k:n_k=1,h_k \text{ distinct},\\ \mathcal{S}h_k\neq h_j \text{ for all } j}}\frac{\theta\beta_k}{2\gamma(v,n)}q\big((\mathcal{S}_k\mathcal{T},\mathbf{n})_v\big)$$
$$+\sum_{k:n_k=1,h_k \text{ distinct}}\sum_{j:\mathcal{S}h_k=h_j}\frac{\theta(n_j+1)\beta_j}{2\gamma(v,n)} \quad (6)$$
$$\times q\Big(\big(\mathscr{R}_k\mathcal{T},\mathscr{R}_k(\mathbf{n}+\mathbf{e_j})\big)_v\Big)$$
$$+\sum_i\sum_{j:\mathscr{C}h_i=h_j,n_j\geq 0,i\neq j}\frac{\beta_j\rho}{2\gamma(v,n)}(n_j+1)p_{h_j,h_i}$$
$$\times q\Big(\big(\mathscr{C}h_i\mathcal{T},\mathbf{n}-\mathbf{e_i}+\mathbf{e_j}\big)_v\Big).$$

The notation in Equation 6 has the following meaning: $\mathbf{e_k}$ is the kth unit vector, representing a multiplicity of the kth distinct haplotype; β_l is the length of ancestral haplotype region divided by length of the lth haplotype. We follow the notation of Griffiths and Tavaré (1995) and use several operators to denote the changes of sample configuration: $\mathcal{S}$ is the shift operator that can be operated on a specific haplotype h_k or the entire haplotype set $\mathcal{T}$. Specifically, $\mathcal{S}h_k$ represents the haplotype obtained by deleting the first mutation coordinate of haplotype h_k. Similarly, $\mathcal{S}_k\mathcal{T}$ represents the new set of distinct haplotypes obtained by deleting the first mutation coordinate of the kth distinct haplotype, h_k, in $\mathcal{T}$. Another operator that operates on the entire haplotype is $\mathcal{R}_k$ which removes the kth distinct haplotype h_k from the set $\mathcal{T}$. The third coordinate change operator $\mathcal{C}$ is defined in this manuscript to denote the coordinate changes caused by recombinations: $\mathcal{C}h_i$ changes one of the two recombination coordinates of haplotype i, $R_{h_i} = \{R_{h_i,1}, R_{h_i,2}\}$, and eliminates all mutation coordinates outside the regions delimited by $R_{h_j,1}$ and $R_{h_j,2}$.

Next we explain how we derive the recursive formula in Equation 6. Starting from time v back from the present, there are four possible paths to arrive at the sample configuration $(\mathcal{T}, \mathbf{n})$ at time v: (1) a coalescent event occurred at time v, and a possible sample configuration prior to time v was $(\mathcal{T}, \mathbf{n} - \mathbf{e_k})$; (2) a mutation occurred on a haplotype h_k that had only single multiplicity, or $n_k = 1$, at time v, and a new mutation coordinate was added to h_k; (3) a mutation occurred on a haplotype h_j that had multiplicity greater than 1, or $n_j > 1$, at time v, and a new haplotype h_k with $n_k = 1$ was generated by adding the new mutation coordinate to h_j; (4) a recombination event occurred, and altered one of the recombination coordinates of h_j by that of h_i. Note that a recombination event not only changes the recombination coordinate, it also changes the mutation coordinates: after the recombination coordinates are changed by a recombination event, all the mutation coordinates of that haplotype are checked, and only those located within the interregion between the two new recombination coordinates are kept. The four terms on the RHS of Equation 6 correspond to the above four paths respectively, and the derivation of the first three terms follows Griffiths and Tavaré (1994a, 1995). In the first path, the sample configuration at time v compatible with the occurrence of coalescence is $(\mathcal{T}, \mathbf{n} - \mathbf{n_k})$, the probability that the event occurred at time v is a coalescent event is $\frac{\binom{n}{2}}{\lambda_v X_v \gamma(v,n)}$. And whenstarting from the configuration $(\mathcal{T}, \mathbf{n} - \mathbf{n_k})$ and going forward in time, the probability that one of the $n_k - 1$ haplotype $\mathbf{e}_k$ is chosen to duplicate is $\frac{n_k-1}{n-1}$ (Griffiths and Tavaré, 1994a). The one-step transition probability of $p((\mathcal{T}, \mathbf{n})_v|(\mathcal{T}, \mathbf{n} - \mathbf{e_k})_v)$ is then $\frac{n_k-1}{n-1}\frac{\binom{n}{2}}{\lambda_v X_v \gamma(v,n)}$. Note that a restriction for haplotype group h_k is that there must be more than one lineage in group h_k at time v. Summing over all possible haplotype groups that have multiplicity $n_k \geq 2$ at time v, and are compatible for coalescent events to occur, we obtain the first term of Equation 6. The second and the third path correspond to the cases when the event occurring at time v is a mutation. Under the assumption of the infinitely many-sites model, if a mutation event occurs, it can result only in one of the single-multiplicity haplotype groups at time v ($n_k = 1$) and the mutation coordinate must be a singleton in the sample configuration at time v. In both the second and the third paths, the chance that a mutation occurred at time v is $\frac{\sum_{l=1}^n \theta_l}{2\gamma(v,n)}$. The configuration at time v compatible with the occurrence of second path is $(\mathcal{S}_k\mathcal{T}, \mathbf{n})$, and the probability for the mutation to happen to haplotype h_k is $\frac{\beta_k}{\sum_{l=1}^n \beta_l}$. Summing over all haplotypes satisfying $n_k = 1$ and $\mathcal{S}h \neq h_j$ for all j yields the second term in Equation 6. In the third path, the probability for the mutation to happen to haplotype h_k is $\sum_{j:\mathcal{S}h_k=h_j}\frac{(n_j+1)\beta_j}{\sum_{l=1}^n \beta_l}$ with the sample configuration prior to the event being $(\mathscr{R}_k\mathcal{T}, \mathscr{R}_k(\mathbf{n} + \mathbf{e_j}))$ for all k with $n_k = 1$ at time v. In the fourth path, recombination occurs with a probability of $\frac{\sum_{l=1}^n \beta_l \rho/2}{\gamma(v,n)}$. If recombination causes a haplotype h_j to become h_i, the haplotype configuration prior to the event is $(\mathcal{C}h_i\mathcal{T}, (\mathbf{n} - \mathbf{e_i} + \mathbf{e_j}))$ and the probability for a haplotype h_j changing into h_i is $\frac{\beta_j(n_j+1)}{\sum_{l=1}^n \beta_l}p_{h_j,h_i}$, with $n_j \geq 0$. The transition probability p_{h_j,h_i} between different haplotypes follows the multilocus haplotype model presented in the section *A simplified multi-locus model for haplotype structure*, where h_i and h_j correspond to one of the distinct haplotypes defined by the "recombination coordinates". Combining these possibilities and averaging over the time to the first event more ancient than t, the sampling distribution of haplotype configuration, $(\mathcal{T}, \mathbf{n})$, is analogous to Equation 2:

$$q\big((\mathcal{T},\mathbf{n})_t\big) = \int_t^\infty \ell_q\big((\mathcal{T},\mathbf{n})_v\big)\gamma\big((\mathcal{T},\mathbf{n})_v\big)\exp\Big(-\int_t^v \gamma\big((\mathcal{T},\mathbf{n})_u\big)du\Big)dv \quad (7)$$

Although we have reduced the state numbers defined by recombinations from $2^{(m_L+m_R+1)}$ to $(m_L + 1) \times (m_R + 1)$ for the simplified multilocus haplotype model, it is still difficult to numerically solve the distribution function induced by the Markov process. Therefore ,we still need to use the importance sampling algorithms to estimate the sampling probability, as will be shown in following sections.

IMPORTANCE SAMPLING AND PROPOSAL DISTRIBUTIONS

Likelihood and importance sampling

Importance sampling algorithms are used to efficiently sample from the probability spaces of frequency trajectories and intra-allelic genealogies in order to approximate the integral in the likelihood

function (see Equation 1). The likelihood function in Equation 1 can be expressed as

$$\mathcal{L}(s) = \iint \mathbb{P}(\mathcal{D}|\mathcal{G})\frac{\mathbb{P}(\mathcal{G}|\mathcal{H})}{\mathbb{Q}(\mathcal{G}|\mathcal{H})}\mathbb{Q}(\mathcal{G}|\mathcal{H})\frac{\mathbb{P}(\mathcal{H}|s,\Gamma)}{\mathbb{Q}(\mathcal{H}|s,\Gamma)}\mathbb{Q}(\mathcal{H}|s,\Gamma)d\mathcal{G}d\mathcal{H}, \tag{8}$$

where $\mathbb{Q}(\mathcal{G}|\mathcal{H})$ and $\mathbb{Q}(\mathcal{H}|s, \Gamma)$ are the proposal distributions that have non-zero weight only on genealogies and trajectories compatible with the data $\mathcal{D}$ (that is, $\mathbb{P}(\mathcal{D}|\mathcal{G}) = 1$). Suppose that M random frequency trajectories and L genealogical histories for each of the trajectories are sampled, then the approximation to Equation 8 becomes

$$\mathcal{L}(s) \approx \frac{1}{M}\frac{1}{K}\sum_{m=1}^{M}\sum_{k=1}^{K}\frac{\mathbb{P}\left(\mathcal{H}^{(m)}\middle|s,\Gamma\right)}{\mathbb{Q}\left(\mathcal{H}^{(m)}\middle|s,\Gamma\right)}\frac{\mathbb{P}\left(\mathcal{G}^{(k)}\middle|\mathcal{H}^{(m)}\right)}{\mathbb{Q}\left(\mathcal{G}^{(k)}\middle|\mathcal{H}^{(m)}\right)}, \tag{9}$$

where $\mathcal{H}^{(m)}$ and $\mathcal{G}^{(k)}$ are the mth and kth independent samples from the proposal distributions. The ratio $\frac{\mathbb{P}(\cdot)}{\mathbb{Q}(\cdot)}$ is called the importance weight. The importance sampling algorithm for the genealogies will be presented in the section *A proposal distribution for sampling genealogical histories conditional on a trajectory* and the algorithm for the allele frequency trajectories in the section *The proposal distribution for sampling allele frequency trajectories of the selected allele in a population of time-varying size*. We illustrate the proposal distributions and the calculation of importance weights in those two sections.

Allele age, T, is not explicitly expressed as a variable in the likelihood function. It is the end point of the frequency trajectory, and thus depends on s through $\mathbb{P}(\mathcal{H}|s)$. Once the maximum likelihood estimate $\hat{s}$ is found, the posterior distribution of T can be obtained from the repeated samples of $\mathcal{H}$ given $s = \hat{s}$. This method for estimating allele age has been used by Coop and Griffiths (2004), Saunders *et al.* (2005) and Wood *et al.* (2005), while it is different from the Bayesian approach of Slatkin (2008), who assumed a prior for allele age and jointly inferred both selection intensity and allele age.

A proposal distribution for sampling genealogical histories conditional on a trajectory

The recursion of the genealogical histories given in Equation 6 for the likelihood of the data cannot be computed exactly for large data sets since there are too many compatible sets of ancestral states. We adopt an importance sampling algorithm to approximate the likelihood by Monte Carlo methods. There are many ways of constructing the proposal distributions for the importance sampling algorithm (Griffiths and Tavaré, 1994b; Stephens *et al.* 2001; Paul *et al.* 2011). Here we follow the scheme developed by Griffiths and Tavaré (1994b). As described in previous sections, the infinitely-many-sites model for mutations in conjunction with the simplified multi-locus haplotype model is assumed.

In the algorithm, a Markov process starting from the configuration in the current generation $(\mathcal{T}, \mathbf{n})_0$, conditional on a randomly sampled historical frequency trajectory $\{X_t, t = 0, \ldots, T\}$, is constructed and simulated backward in time until reaching the absorbing state $(\mathcal{T}, \mathbf{e_1})_\tau$ at time $\boldsymbol{\tau}$. The algorithm is summarized as follows:

1. Generate time to the next event, v, by the density function given in Equation 4;
2. Choose one of the three possible events (recombination, mutation or coalescence) from the proposal distribution. We first define the total rate that any event occurs at time v as

$$\hbar\left((\mathcal{T},\mathbf{n})_v\right) = \sum_{k,n_k\geq 2} n(n_k-1)\frac{1}{\lambda_v X_v} + \sum_i \sum_{\substack{j:\mathscr{C}h_i=h_j \\ n_j\geq 0, i\neq j}} \beta_j\rho(n_j+1)p_{h_j,h_i} + \theta m, \tag{10}$$

where

$$m = \sum_{k,n_k=1,h_k \text{ distinct},\mathscr{S}h_k\neq h_j \text{ for all } j} \beta_k + \sum_{k:n_k=1,h_k \text{ distinct}} \sum_{j:\mathscr{T}h_k=h_j} (n_j+1)\cdot\beta_j. \tag{11}$$

The **proposal distribution** is designed in such a way that a possible event at time v is chosen with probability in proportion to the size of each term in $\hbar((\mathcal{T}, \mathbf{n})_v)$:

$$p\left((\mathcal{T}',\mathbf{n}')_v\middle|(\mathcal{T},\mathbf{n})_v\right) = \begin{cases} \dfrac{n(n_k-1)\frac{1}{\lambda_v X_v}}{\hbar\left((\mathcal{T},\mathbf{n})_v\right)}, & (\mathcal{T}',\mathbf{n}') = (\mathcal{T},\mathbf{n}-\mathbf{e_k}) \text{ and } n_k\geq 2, \\ \dfrac{\theta\beta_k}{\hbar\left((\mathcal{T},\mathbf{n})_v\right)}, & (\mathcal{T}',\mathbf{n}') = (\mathscr{S}_k\mathcal{T},\mathbf{n}), \\ \dfrac{\theta\beta_j(n_j+1)}{\hbar\left((\mathcal{T},\mathbf{n})_v\right)}, & (\mathcal{T}',\mathbf{n}') = \left(\mathscr{R}_k\mathcal{T},\mathscr{R}_k\mathbf{n}+\mathbf{e_j}\right), \\ \dfrac{\beta_j\rho(n_j+1)p_{h_j,h_i}}{\hbar\left((\mathcal{T},\mathbf{n})_v\right)}, & (\mathcal{T}',\mathbf{n}') = \left(\mathscr{C}h_i\mathcal{T},\mathbf{n}-\mathbf{e_i}+\mathbf{e_j}\right). \end{cases} \tag{12}$$

3. Update the configuration to reflect the chosen event. Let $\mathcal{G}_j$, $j \geq 0$ denote the jth event during the genealogical history, and $\mathcal{G}_0 = (\mathcal{T}, \mathbf{n})_0$. $\mathbb{Q}(\mathcal{G}_j\,|\mathcal{G}_{j-1})$ is the transition probability of the backward Markov process determined by the proposal distribution (Equation 12). Similarly, $\mathbb{P}(\mathcal{G}_{j-1}\,|\mathcal{G}_j)$ is the transition probability of the forward Markov process. The sequential importance weight for the jth step change $\frac{\mathbb{P}(\mathcal{G}_{j-1}|\mathcal{G}_j)}{\mathbb{Q}(\mathcal{G}_j|\mathcal{G}_{j-1})}$ is estimated. Here we illustrate the calculation of the importance weight for the case in which the chosen event is $\mathcal{G}_k = (\mathcal{T}, \mathbf{n} - \mathbf{e_k})$. As shown in the section *Sampling probability of a multilocus haplotype configuration*, $\mathbb{P}(\mathcal{G}_{j-1}|\mathcal{G}_j) = \frac{\binom{n}{2}(n_k-1)}{\lambda_v X_v \gamma(v,n)(n-1)}$, and from Equation 12 we have $\mathbb{Q}(\mathcal{G}_j|\mathcal{G}_{j-1}) = \frac{n(n_k-1)\frac{1}{\lambda_v X_v}}{\hbar((\mathcal{T},\mathbf{n})_v)}$. Taking the ratio of the two terms, we obtain the importance weight for the jth step: $\frac{\hbar((\mathcal{T},\mathbf{n})_v)}{2\gamma(v,n)}$. Table 5 provides more details of importance weights for other events.
4. Repeat steps 1−3 to continue generating the historical events in the genealogy backward in time.
5. Stop when the absorbing state is reached, that is, a single lineage remains in the sample configuration, $(\mathcal{T}, \mathbf{e_1})$, or if the proposed time for next event is beyond the end of the frequency trajectories;
6. Assume that there are I steps until the Markov chain reaches the absorbing states, the ratio of the forward/backward paths is the product of sequential importance weights:

$$\frac{\mathbb{P}\left(\mathcal{G}^{(k)}\middle|\mathcal{H}\right)}{\mathbb{Q}\left(\mathcal{G}^{(k)}\middle|\mathcal{H}\right)} = \prod_{j=1}^{I}\left[\frac{\mathbb{P}(\mathcal{G}_{j-1}|\mathcal{G}_j)}{\mathbb{Q}(\mathcal{G}_j|\mathcal{G}_{j-1})}\right]\cdot q\left((\mathcal{T},\mathbf{e_1})_\tau\right), \tag{13}$$

Table 5 The proposal distribution and importance weights for the importance sampling algorithm presented in the section *A proposal distribution for sampling genealogical histories conditional on a trajectory*

$\mathcal{G}_j$	$\mathbb{Q}(\mathcal{G}_j\|\mathcal{G}_{j-1})$	$\mathbb{P}(\mathcal{G}_{j-1}\|\mathcal{G}_j)$	Importance Weight
$(\mathcal{T}, \mathbf{n}-\mathbf{e_k})$	$\frac{n(n_k-1)}{\lambda_v X_v \hbar((\mathcal{T},\mathbf{n})_v)}$	$\frac{\binom{n}{2}(n_k-1)}{\lambda_v X_v \gamma(v,n)(n-1)}$	$\frac{\hbar((\mathcal{T},\mathbf{n})_v)}{2\gamma(v,n)}$
$(\mathcal{S}_k\mathcal{T}, \mathbf{n})$	$\frac{\theta\beta_k}{\hbar((\mathcal{T},\mathbf{n})_v)}$	$\frac{\sum_{l=1}^n\beta_l\theta/2}{\gamma(v,n)}\frac{\beta_k}{\sum_{l=1}^n\beta_l}$	$\frac{\hbar((\mathcal{T},\mathbf{n})_v)}{2\gamma(v,n)}$
$(\mathcal{R}_k\mathcal{T}, \mathcal{R}_k(\mathbf{n}+\mathbf{e_j}))$	$\frac{\theta(n_j+1)\beta_j}{\hbar((\mathcal{T},\mathbf{n})_v)}$	$\frac{\sum_{l=1}^n\beta_l\theta/2}{\gamma(v,n)}\frac{(n_j+1)\beta_j}{\sum_{l=1}^n\beta_l}$	$\frac{\hbar((\mathcal{T},\mathbf{n})_v)}{2\gamma(v,n)}$
$(\mathcal{C}h_i\mathcal{T}, \mathbf{n}-\mathbf{e_i}+\mathbf{e_j})$	$\frac{\beta_j\rho(n_j+1)p_{h_j,h_i}}{\hbar((\mathcal{T},\mathbf{n})_v)}$	$\frac{\beta_j\rho(n_j+1)p_{h_j,h_i}}{\gamma(v,n)}$	$\frac{\hbar((\mathcal{T},\mathbf{n})_v)}{2\gamma(v,n)}$

$\{\mathcal{G}\}$ denotes four possible events of the genealogical history. $\mathbb{P}(\mathcal{G}_{j-1}|\mathcal{G}_j)$ and $\mathbb{Q}(\mathcal{G}_j|\mathcal{G}_{j-1})$ are the one-step transition probability of the forward and backward Markov process constructed for simulating the genealogical history. The importance weight is estimated by $\frac{\mathbb{P}(\mathcal{G}_{j-1}|\mathcal{G}_j)}{\mathbb{Q}(\mathcal{G}_j|\mathcal{G}_{j-1})}$

and is used in the likelihood function. For those paths with time beyond the end of the given frequency trajectories, $\{X_t\}$, the ratio is set to zero, which means the sample is rejected.

The proposal distribution for sampling allele frequency trajectories of the selected allele in a population of time-varying size

We use the backward Wright-Fisher model under selection to sample the allele frequency trajectories. The importance sampling algorithm for sampling frequency trajectories of the selected allele is described as follows. A detailed explanation can be found in the original paper (Slatkin 2001).

1. Given a selection intensity s and parameter set Γ, a sample path is simulated from $t = 0$ (current generation) with I_0 copies of A, and then proceeds backward from generation to generation assuming the following binomial distribution:

$$\mathbb{P}(I_t|I_{t-1}) = \binom{2N_t}{I_t} Y_{t-1}'^{I_t}(1-Y_{t-1}')^{2N_t-I_t}, \tag{14}$$

where Y_{t-1}' satisfies: $Y_{t-1}'\frac{1+s_1Y_{t-1}'+s_2(1-Y_{t-1}')}{1+s_1Y_{t-1}'^2+2s_2Y_{t-1}'(1-Y_{t-1}')} = Y_{t-1}$, and $Y_{t-1} = I_{t-1}/(2N_{t-1})$. The backward process is stopped at time T when the allele is lost. The probability of the backward process is calculated as: $\mathbb{P}_B(\mathcal{H}^{(m)}) = \prod_{t=1}^{T}\mathbb{P}(I_t|I_{t-1})$.

2. For a frequency trajectory $\mathcal{H}^{(m)}$ simulated in Step 1, the probability it is generated by the forward process is computed. In the Wright-Fisher model with selection, the number of allele A from generation t to generation $t-1$ follows a binomial distribution:

$$\mathbb{P}(I_{t-1}|I_t) = \binom{2N_{t-1}}{I_{t-1}} X_t'^{I_{t-1}}\left(1-X_t'\right)^{2N_{t-1}-I_{t-1}}, \tag{15}$$

with

$$X_t' = X_t\frac{1+s_1X_t+s_2(1-X_t)}{1+s_1X_t^2+2s_2X_t(1-X_t)}, \tag{16}$$

which is the allele frequency of A after selection in generation t. In Equation 16, $X_t = I_t/(2N_t)$ is the frequency of allele A before selection in generation t. The probability of the sample path $\mathcal{H}^{(m)}$ is

$$\mathbb{P}_F\left(\mathcal{H}^{(m)}\right) = \prod_{t=T}^{1}\mathbb{P}(I_{t-1}|I_t), \tag{17}$$

where $\mathbb{P}(I_{T-1}|I_T) = 1$ if $I_{T-1} = 1$ and 0 otherwise. And the subscript F indicates that the process is "forward" in time.

3. The importance weight is calculated as (Slatkin 2001):

$$\frac{\mathbb{P}\left(\mathcal{H}^{(m)}\right)}{\mathbb{Q}\left(\mathcal{H}^{(m)}\right)} = \frac{\mathbb{P}_F\left(\mathcal{H}^{(m)}\right)}{\mathbb{P}_B\left(\mathcal{H}^{(m)}\right)}\dot{N}, \tag{18}$$

where $\dot{N}$ is the population size at the first generation after the allele is lost in the backward process. The multiplication of $\dot{N}$ is needed in Equation 18, since the rate of influx of new mutations is proportional to the population size of that generation.

APPLICATIONS

Simulation

Using the coalescent simulator SelSim (Spencer and Coop 2004), data are generated for two sets of parameters corresponding to medium and strong selection respectively: $\theta = 4N\mu = 500$, $\rho = 4Nr = 500$, $Ns = 50$, and $\theta = 4N\mu = 500$, $\rho = 4Nr = 500$, $Ns = 500$, where θ, ρ, and Ns represent the mutation rate, recombination rate and selection coefficient scaled by the effective population size. The frequencies of the selected alleles at the present are chosen to be 0.60. Since the Moran model is used in SelSim, whereas the Wright-Fisher model is used in our method, the effective population size in the simulations is scaled to match that of a Wright-Fisher model by multiplying by a factor of 2 (Watterson 1975). We estimate the log-likelihood of s for a range of selection coefficients with the other parameters in Γ known, assuming that the population has a constant size of 10,000. The curves of the log-likelihood over the grid of s values are smoothed by a local polynomial smoother. This smoother fits a linear function to a subset of data points within a local window of the target point where the log-likelihood is to be estimated. The fitting is carried out by the weighted least square regression, which gives more weight to points close to the target point and less weight to distant points. The log-likelihood is thus estimated as the fitted value at the target point. The size of the local window or the bandwidth is chosen by eye for each curve. The log-likelihood curves are plotted in Figure 4 and Figure 5.

To evaluate the performance of the importance sampling approximation, we perform eight independent simulations for every parameter combination. One million iterations in the importance sampling algorithm are required to ensure good estimates, and the likelihood curves are presented together in Figure 4 and Figure 5. For

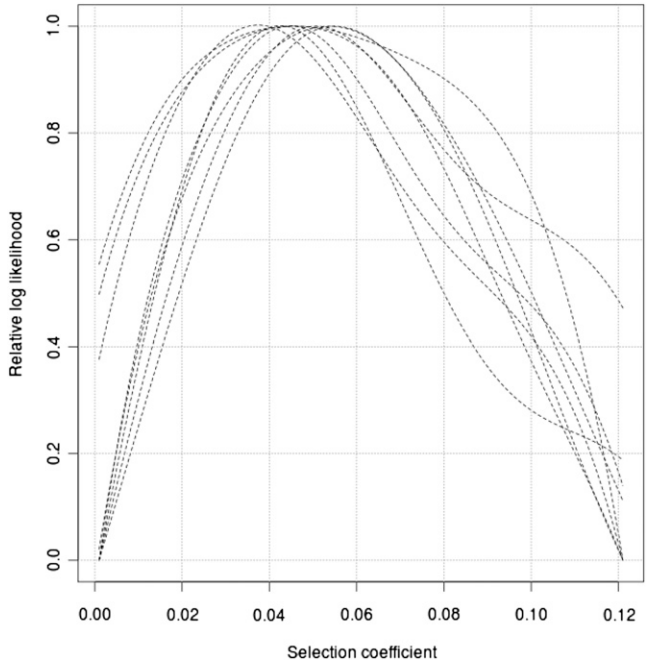

Figure 4 The relative likelihood curve for the simulated data with the selection coefficient s = 0.05 and a constant population size N = 10,000. The comparison of eight estimates of the likelihood curves is presented. Each estimate is an independent run of our method on different simulated data. The results are from 1 million iterations.

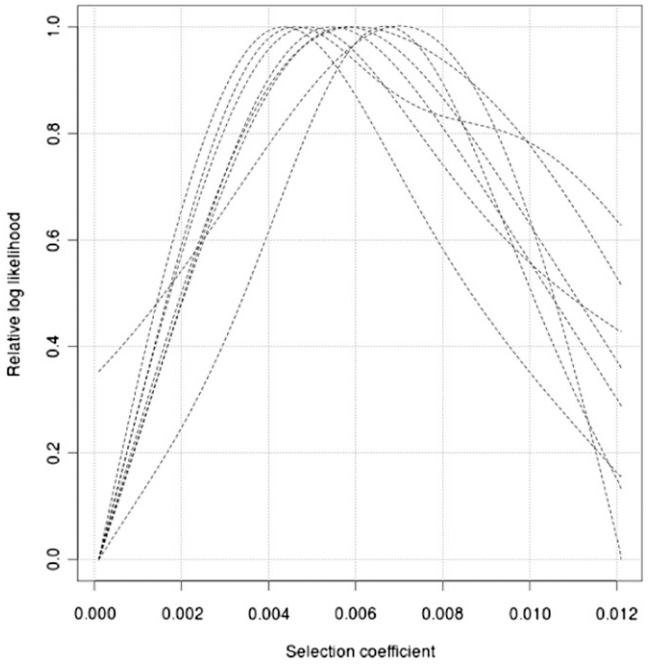

Figure 5 The relative likelihood curve for the simulated data with the selection coefficient s = 0.005 and a constant population size N = 10,000. The comparison of eight estimates of the likelihood curves is presented. Each estimate is an independent run of our method on different simulated data. The results are from 1 million iterations.

the data set simulated with s = 0.005, the MLE ranges from 0.0041 to 0.0073. For the data set with s = 0.05, the MLE ranges from 0.032 to 0.0543.

Glucose-6-phosphate dehydrogenase (*G6PD*)

The *G6PD* gene is located on the X-chromosome. Some alleles are known to confer the resistance to malaria (Ruwende *et al.* 1995). Case-control studies have demonstrated that a common variant, *G6PD-202A*, reduces the risk of malaria by approximately 50% (Ruwende *et al.* 1995). This allele is at low frequency in most populations but has an intermediate frequency in sub-Saharan Africa. Several population genetic studies have investigated the effect of a recent selective sweep in this region (Tishkoff *et al.* 2001; Sabeti *et al.* 2002). Here we use the data in Sabeti *et al.* (2002), which consists of 252 males from there African populations in a 440-Kb region covering the *G6PD* gene. We analyze only the 60 haplotypes from the Beni population. There are 10 haplotypes containing the 202-A allele in the sample. We assume that the frequency of the selected allele in the Beni population is the same as that estimated from the sample which is 0.1667. The recombination fractions among SNPs are obtained by interpolation with the Oxford fine-scale recombination map (Myers *et al.* 2005). The recombination rate in the *G6PD* gene region is heterogeneous with two recombination hot-spots, and the overall averaged recombination rate for the region is 1.4410 cM/Mb. We determine the end points of ancestral haplotypes and mutations by running the hidden Markov model (Chen 2007). The data configuration is coded by the rules presented in the section *A simplified multilocus model for haplotype structure* as shown in Table 6.

Because the hidden Markov model analysis indicates there are no mutations in the ancestral haplotype regions, we set θ to 0.0. We assume an effective population size of N = 10,000, which is constant over time. Because *G6PD* is X-linked, N_e is 3/4 of the autosomal size. We assume an additive model for selection, which means the fitness's of the three genotypes aa, Aa, and AA are 1, 1 + 1/2s, and 1 + s, respectively. The likelihood of the selection coefficient is estimated by our method from 1 million iterations of the importance sampling algorithm. The log-likelihood curve is plotted in Figure 6. The selection coefficient is estimated to be 0.0456 (95% confidence interval of 0.0144−0.0769). From the estimated selection coefficient, the age of the 202-A allele can be estimated. As shown in Figure 7, given the selection coefficients estimated, the corresponding posterior distribution of allele age is plotted.

DISCUSSION

We have developed a likelihood method for estimating selection intensity and allele age from haplotype structure of multilocus SNPs closely linked to a selected mutant. The likelihood is based on the proposed simplified multilocus haplotype model, which describes the ancestral process of haplotype extent under the joint effects of selection, recombination and mutation. In this model, the state space of the ancestral process is determined by the extent of intact ancestral haplotypes in the vicinity of the selected mutant and the new

Table 6 The sample configuration of the *G6PD* data according to coding rules in the section *A simplified multilocus model for haplotype structure*

Haplotype Type	Count Number
(11, 7)	5
(4, 7)	1
(12, 7)	1
(17, 7)	3

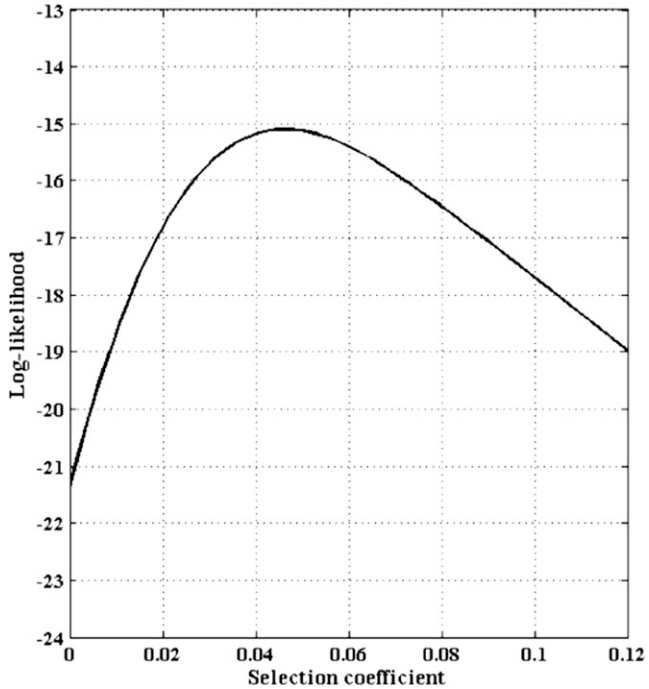

Figure 6 Likelihood curve for the *G6PD* data as a function of the selection coefficient with a constant population size of 10,000. The likelihood curve is smoothed by a local polynomial smoother. The point estimate of the selection coefficients is 0.0456 with the 95% confidence interval of (0.0144, 0.0769).

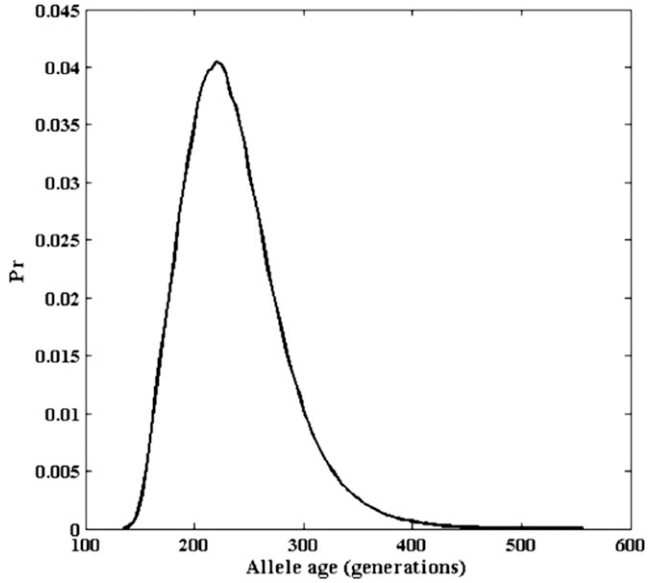

Figure 7 The posterior distribution of the allele age in generations when the selection coefficient is set to the value estimated from Figure 6.

mutations arising on the ancestral haplotypes during the selective process. Our method adopts importance sampling algorithms to efficiently explore the genealogical history of the sample for evaluation of the sampling probability. By applying the method to both simulated and real data, we demonstrate that the extent of the haplotype structure is informative for the inference of selection intensity of a recent positive selection.

Our method has two merits. First, by exploiting the extent of haplotype structure and focusing only on subprocesses of the ARG related to the retained ancestral segments on the selected haplotypes, we dramatically reduce the computational burden such that data sets from genomic regions of mega-base magnitude can be analyzed. Second, our method can allow for changes in population size. This is especially important for samples from human populations outside Africa, because population growth can affect the pattern of linkage disequilibrium and haplotype structure, and thus lead to an incorrect estimation of the selection intensity if the effect of demographic history is not explicitly modeled. In our analysis of simulated data, we found that the estimated selection coefficient is accurate but is sensitive to the recombination rates assumed. Since the variability of recombination rates is high over human genome (Myers *et al.* 2005), good estimates of local rates are necessary to obtain accurate estimates of selection coefficients.

Another factor that may affect the estimates of selection intensity and allele age is the SNP marker density in the data. Because mutants that have experienced positive selections are typically young, new mutations at nearby loci accumulate at a relatively slow rate compared to the rate of recombination that breaks down linkage disequilibrium. Therefore, we expect fewer segregating sites observed in regions under recent positive selection. In the low-density SNP data, these segregating sites are likely not typed. In the two data sets of the *G6PD* region we analyzed, Sabeti *et al.* (2002)'s and Verrelli *et al.* (2002)'s data, no new mutations at closely linked loci were detected. Because their *G6PD* data were not generated by resequencing, a proportion of mutations may not have been identified or included in the data. We expect that resequencing data from the target gene regions will be more informative for identifying the occurrence of recombinations and mutations during the selective process. The method developed in this paper is for identifying ongoing positive selection. If the selected allele has been fixed in the population, mutations accumulated since its fixation become informative and important for inferring the fixation time, for which the allele frequency spectrum after a selective sweep is a better choice (Chen 2012).

The importance sampling algorithm for the genealogies adopted in this paper was developed by Griffiths and Tavaré (1994b). In their proposal distribution, at each step any possible event that could lead to the current configuration is considered and sampled in proportion to their rate of occurrence (Felsenstein *et al.* 1999). More efficient proposal distributions have been developed (Stephens *et al.* 2001; Slatkin 2002; De Lorio and Griffiths 2004; Paul *et al.* 2011) and can be adopted to improve the computational efficiency of our method.

ACKNOWLEDGMENT

We are grateful to Drs. Kun Chen, Graham Coop, and Steve Evans for helpful discussions; to Dr. Pardis Sabeti for kindly providing us the *G6PD* data; and to Dr. Jeff Wall and the anonymous reviewers for their helpful comments. The work was supported by a National Institutes of Health grant GM-40282 to M.S.

LITERATURE CITED

Barton, N. H., 1998 The effect of hitch-hiking on neutral genealogies. Genet. Res. 72: 123–133.

Bersaglieri, T., P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner *et al.*, 2004 Genetic signatures of strong recent positive selection at the lactase gene. Am. J. Hum. Genet. 74: 1111–1120.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley, and W. Stephan, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140: 783–796.

Chen, H., 2012 The joint allele frequency spectrum of multiple populations: a coalescent theory approach. Theor. Popul. Biol. 81: 179–195.
Chen, H., 2007 Statistical methods for inference of positive selection from genetic polymorphism. Ph.D. thesis, University of California, Berkeley.
Chen, H., N. Patterson, and D. Reich, 2010 Population differentiation as a test for selective sweeps. Genome Res. 20: 393–402.
Coop, G., and R. C. Griffiths, 2004 Ancestral inference on gene trees under selection. Theor. Popul. Biol. 64: 241–251.
De Lorio, M., and R. C. Griffiths, 2004 Importance sampling in coalescent histories I. Adv. Appl. Probab. 36: 417–433.
Durrett, R., and J. Schweinsberg, 2004 Approximating selective sweeps. Theor. Popul. Biol. 66: 129–138.
Etheridge, A., P. Pfaffelhuber, and A. Wakolbinger, 2006 An approximate sampling formula under genetic hitchhiking. Ann. Appl. Probab. 16: 685–729.
Fay, J. C., and C. I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics 155: 1405–1413.
Fearnhead, P., and P. Donnelly, 2001 Estimating recombination rates from population genetic data. Genetics 159: 1299–1318.
Felsenstein, J., 1988 Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. 22: 521–565.
Felsenstein, J., M. Kuhner, J. Yamato, and P. Beerli, 1999 Likelihoods on coalescents: A Monte Carlo sampling approach to inferring parameters from population samples of molecular data, pp. 163–185 in *Statistics in Molecular Biology and Genetics*, edited by F. Seillier-Moiseiwitsch. Institute of Mathematical Statistics, Beachwood, OH.
Fu, Y. X., and W. H. Li, 1993 Statistical tests of neutrality of mutations. Genetics 133: 693–709.
Griffiths, R. C., and P. Marjoram, 1996 Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol. 3: 479–502.
Griffiths, R. C., and S. Tavaré, 1994a Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. Lond., B 344: 403–410.
Griffiths, R. C., and S. Tavaré, 1994b Simulating probability distributions in the coalescent. Theor. Popul. Biol. 46: 131–159.
Griffiths, R. C., and S. Tavaré, 1995 Unrooted genealogical tree probabilities in the infinitely-many-sites model. Math. Biosci. 127: 77–98.
Hudson, R. R., and N. L. Kaplan, 1988 The coalescent process in models with selection and recombination. Genetics 819: 831–840.
Kamberov, Y. G., S. Wang, J. Tan, P. Gerbault, A. Wark *et al.*, 2013 Modeling recent human evolution in mice by expression of a selected EDAR variant. Cell 152: 691–702.
Kaplan, N. L., T. Darden, and R. R. Hudson, 1988 The coalescent process in models with selection. Genetics 120: 819–829.
Kaplan, N. L., R. Hudson, and C. Langley, 1989 The" hitchhiking effect" revisited. Genetics 123: 887–899.
Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160: 765–777.
Krone, S. M., and C. Neuhauser, 1997 Ancestral processes with selection. Theor. Popul. Biol. 51: 210–237.
Kuhner, M. K., J. Yamato, and J. Felsenstein, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics 140: 1421–1430.
Myers, S., L. Bottolo, C. Freeman, G. McVean, and P. Donnelly, 2005 A fine-scale map of recombination rates and hotspots across the human genome. Science 310: 321–324.
Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark *et al.*, 2005 Genomic scans for selective sweeps using SNP data. Genome Res. 15: 1566–1575.
Ohta, T., and M. Kimura, 1975 The effect of selected linked locus on heterozygosity of neutral alleles (the hitch-hiking effect). Genet. Res. 25: 313–326.
Paul, J., M. SteinrÃijcken, and Y. Song, 2011 An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. Genetics 187: 1115.
Peng, Y., Z. Yang, H. Zhang, C. Cui, X. Qi *et al.*, 2011 Genetic variations in Tibetan populations and high-altitude adaptation at the Himalayas. Mol. Biol. Evol. 28: 1075–1081.
Rannala, B., and J. Reeve, 2001 High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. Am. J. Hum. Genet. 69: 159–178.
Rannala, B., and J. Reeve, 2003 Joint Bayesian estimation of mutation location and age using linkage disequilibrium. Pac Symp BioComput. 526–534.
Ruwende, C., S. C. Khoo, R. W. Snow, S. N. Yates, D. Kwiatkowski *et al.*, 1995 Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. Nature 376: 246–249.
Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.
Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter *et al.*, 2007 Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913–918.
Saunders, M., M. Slatkin, C. Garner, M. Hammer, and M. Nachman, 2005 The extent of linkage disequilibrium caused by selection on G6PD in humans. Genetics 171: 1219–1229.
Scheet, P., and M. Stephens, 2006 A fast and flexible statistical model for large-scale population genotype data: Application to inferring missing genotype and haplotypic phase. Am. J. Hum. Genet. 78: 629–644.
Simonson, T., Y. Yang, C. Huff, H. Yun, G. Qin *et al.*, 2010 Genetic evidence for high-altitude adaptation in Tibet. Science 329: 72.
Slade, P., 2000 Most recent common ancestor probability distributions in gene genealogies under selection. Theor. Popul. Biol. 58: 291–305.
Slatkin, M., 2001 Simulating genealogies of selected alleles in a population of variable size. Genet. Res. 78: 49–57.
Slatkin, M., 2002 A vectorized method of importance sampling with application to models of mutation and migration. Theor. Popul. Biol. 62: 339–348.
Slatkin, M., 2008 A Bayesian method for jointly estimating allele age and selection intensity. Genet. Res. 90(1): 129–137.
Spencer, C. A., and G. Coop, 2004 Selsim: a program to simulate population genetic data with natural selection and recombination. Bioinformatics 20(18): 3673–3675.
Stephan, W., T. Wiehe, and M. Lenz, 1992 The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor. Popul. Biol. 41: 237–254.
Stephens, M., N. Smith, and P. Donnelly, 2001 A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. 68: 978–989.
Tajima, F., 1989 Statistical methods for testing the neutral mutations hypothesis by DNA polymorphism. Genetics 123: 585–595.
Tang, K., K. R. Thornton, and M. Stoneking, 2007 A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol. 5: e171.
Tishkoff, S., R. Varkonyi, N. Cahinhinan, S. Abbes, G. Argyropoulos *et al.*, 2001 Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 293: 455.
Tishkoff, S. A., F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt *et al.*, 2007 Convergent adaptation of human lactase persistence in Africa and Europe. Nat. Genet. 39: 31–40.
Verrelli, B. C., J. H. McDonald, G. Argyropoulos, G. Destro-Bisol, A. Froment *et al.*, 2002 Evidence for balancing selection from nucleotide sequence analyses of human G6PD. Am. J. Hum. Genet. 71: 1112–1128.
Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. PLoS Biol. 4: 446–458.
Watterson, G., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7: 256–276.
Wood, E., D. Stover, M. Slatkin, M. Nachman, and M. Hammer, 2005 The β-globin recombinational hotspot reduces the effects of strong selection around HbC, a recently arisen mutation providing resistance to malaria. Am. J. Hum. Genet. 77: 637–642.
Xu, S., S. Li, Y. Yang, J. Tan, H. Lou *et al.*, 2011 A genome-wide search for signals of high-altitude adaptation in Tibetans. Mol. Biol. Evol. 28: 1003–1011.
Yi, X., Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. P. Cuo *et al.*, 2010 Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329: 75.

Communicating editor: D.-J. De Koning

APPENDIX A

The probability distribution for the extent of an intact ancestral haplotype under selection

We show how to derive the distribution for the extent of an intact ancestral haplotype at time t during a selective process. We investigate a selective sweep that starts from a single copy of the selected allele. Here we denote the time when the new selected mutant arises by 0, and look forward in time. Let X_t be frequency of the selected allele at time t. If we ignore the initial randomness of allele frequency trajectories, which is usually modeled using a supercritical branching process, X_t can be well approximated using the deterministic logistic equation (Ohta and Kimura, 1975):

$$X_t = \frac{X_0}{X_0 + (1 - X_0)e^{-st}}, \tag{19}$$

where X_0 can be set to $1/2N$ (Kaplan *et al.* 1989; Stephan *et al.* 1992). Assume the selected locus has two alleles A and a, with A being the advantageous allele. We are modeling a continuous segment between the selected mutant and the neutral marker, which has two alleles B and b. We use upper case letters to denote that the position is descended from the ancestral haplotype, and lower case letters to denote the position descended from the background haplotypes. Note here the two "alleles" are defined according to whether they are descended from the ancestral haplotype or not, instead of the true observed nucleotide types at that locus. We use $[AB]$ to denote a segment of ancestral haplotype with loci A and B being the end points. Let $P_{[AB]}(t)$ be the relative population frequency of such fragments among all haplotypes carrying the selected allele A at time t. Furthermore, we use $[A-]$ to denote an ancestral haplotype with A being from the ancestral haplotype, while the state of the other end point of the fragment is not determined.

For a random ancestral haplotype $[AB]$ or $[Ab]$, if it recombines with any $[A-]$ haplotype during the interval $[0, t]$, it does not change $P_{[AB]}(t)$. The only possible change is caused by "effective" recombinations, that is, recombination with an $[ab]$, or $[aB]$ haplotype from the neutral haplotype "sub-population". The expected number of effective recombination events for a $[AB]$ recombining with any $[a-]$ haplotype during the interval $[0, t]$ is

$$\begin{aligned} C &= r\int_{u=0}^{t}(1 - X(u))du \\ &= rt - \frac{r}{s}\ln\left(1 - X_0 + e^{st}X_0\right) \end{aligned} \tag{20}$$

It is not hard to see that the number of effective events on an $[AB]$ ancestral haplotype during the time interval $[0, t]$ follows a Poisson distribution. $P_{[AB]}(t)$ is then identical to the probability of no effective recombination between $[AB]$ for an ancestral haplotype:

$$\begin{aligned} P_{[AB]}(t) &= e^{-rt+\frac{r}{s}\ln(1-X_0+e^{st}X_0)} \\ &= e^{-rt}(1-(1-e^{st})X_0)^{r/s}. \end{aligned} \tag{21}$$

Similarly, for an ancestral haplotype $[ABC]$, the probability of being intact during the interval $(0, t)$ follows the Equation 21, except that the recombination fraction is replaced by $r_{[AC]}$.

Note that in Equation 21, when st is small, which means either the selective process is at an early stage or the selection is weak, the term $(1 - (1 - e^{st}) X_0)^{r/s} \approx 1$, and thus similar to the neutral case. However, if st gets larger, the term cannot be ignored. For example, if $r = 0.001$, $s = 0.01$, $X_0 = 0.001$ and $t = 1000$, the relative bias can be as large as ~27%.

Also note that in the aforementioned derivation for the distribution of ancestral haplotypes, we ignore the randomness of the frequency trajectory of the selected allele at the very early stage of the selective process, and approximate the trajectory with a deterministic equation. Ignoring the randomness of the allele frequency trajectory at the early stage can bias the inference of parameters related to the sweep process, but as pointed out in previous studies (Kaplan *et al.* 1989; Braverman *et al.* 1995; Barton, 1998; Durrett and Schweinsberg, 2004; Etheridge *et al.* 2006), when the selection intensity is sufficiently strong, the bias is small. For this reason, our method is more suitable to analyze genes under strong selections.