Maximum likelihood estimation of natural selection and allele age from time series data of allele frequencies

Zhangyi He[a,1,*], Xiaoyang Dai[b], Mark Beaumont[b], Feng Yu[c,*]

[a] *Department of Statistics, University of Oxford, Oxford OX1 3LB, United Kingdom*
[b] *School of Biological Sciences, University of Bristol, Bristol BS8 1TQ, United Kingdom*
[c] *School of Mathematics, University of Bristol, Bristol BS8 1TW, United Kingdom*

Abstract

Thanks to advances in ancient DNA preparation and sequencing techniques, time serial samples of segregating alleles are becoming more widely available in ancestral populations. Such time series data allow for more accurate inference of population genetic parameters and hypothesis testing on the recent action of natural selection. Here we develop a likelihood-based method for co-estimating the selection coefficient and the allele age from allele frequency time series data. Our method is built on the hidden Markov model incorporating the Wright-Fisher diffusion conditioned to survive until the time of the most recent sample, which circumvents the assumption required in existing methods that the allele is created by mutation at a certain small frequency. We calculate the likelihood by numerically solving the Kolmogorov backward equation resulting from the conditioned Wright-Fisher diffusion backwards in time and re-weighting the solution by the emission probabilities of the observation at each sampling time point, which allows for a reduction of the two-dimensional numerical search for the maximum of the likelihood surface for the selection coefficient and the allele age. We show through extensive simulations that our approach can produce unbiased estimates of the selection coefficient and the allele age, even if the samples are sparsely distributed in time with small uneven sizes. We illustrate the utility of our method on real data by re-analysing the ancient DNA data associated with horse coat colouration and show that grouping samples can significantly bias the results of inference.

Keywords: Natural selection, Allele age, Conditioned Wright-Fisher diffusion, Hidden Markov model, Maximum likelihood estimation

*Corresponding author.
Email addresses: `zhangyi.he@stats.ox.ac.uk` (Zhangyi He), `feng.yu@bristol.ac.uk` (Feng Yu)
[1] Present address: MRC Toxicology Unit, University of Cambridge, Leicester LE1 7HB, United Kingdom & Email: zh320@mrc-tox.cam.ac.uk

1. Introduction

Recent advances in ancient DNA (aDNA) preparation and sequencing techniques have made available an increasing amount of high-quality time serial samples of segregating alleles in ancestral populations, *e.g.*, for humans (Sverrisdóttir et al., 2014; Mathieson et al., 2015), chickens (Flink et al., 2014; Loog et al., 2017) and horses (Ludwig et al., 2009; Pruvost et al., 2011). Such genetic time series data provide valuable information about the allele frequency trajectory through time and allow for a better understanding of the evolutionary history of populations (see Leonardi et al., 2017, for a detailed review of the most recent findings in aDNA). One of the most important applications of aDNA is to study natural selection since it enables us to directly track the change in allele frequencies over time, which is the characteristic of the action of natural selection. A number of studies over the last decade have been published capitalising on the temporal aspect of aDNA data to characterise the process of natural selection. For example, Mathieson et al. (2015) used aDNA data to identify candidate loci under natural selection in European humans.

Following the increased availability of aDNA data, several methods for statistical inference of natural selection from time series data of allele frequencies have been developed in the past ten years. Initially, Bollback et al. (2008) proposed a likelihood-based method to estimate the selection coefficient from allele frequency time series data, assuming a Wright-Fisher model introduced by Fisher (1922) and Wright (1931). The population allele frequency was modelled as a latent variable in a hidden Markov model (HMM), where the sample allele frequency drawn from the underlying population at each given time point was treated as a noisy observation of the latent population allele frequency. Since there is no tractable analytical form for the transition probabilities of the Wright-Fisher model and its numerical evaluation is computationally prohibitive for large population sizes and evolutionary timescales, the Wright-Fisher model in their likelihood calculations was approximated with its standard diffusion limit, commonly known as the Wright-Fisher diffusion. The transition probabilities of the allele frequencies were calculated by numerically solving the Kolmogorov backward equation (KBE) resulting from the Wright-Fisher diffusion. Using the approach of Bollback et al. (2008), Ludwig et al. (2009) analysed the aDNA data associated with horse coat colouration and found that natural selection acted strongly on the locus encoding the Agouti signalling peptide (*ASIP*) and the locus encoding

the melanocortin 1 receptor (*MC1R*).

Malaspinas et al. (2012) extended the HMM framework of Bollback et al. (2008) to jointly estimate the selection coefficient, the population size and the allele age from time series data of allele frequencies. Allele age is the elapsed time since the allele was created by mutation. Along with the selection coefficient, the allele age plays an important role in determining the sojourn time of a beneficial mutation (see Slatkin & Rannala, 2000, for a detailed review). The co-estimation of the allele age circumvents the assumption required in Bollback et al. (2008) that the allele frequency of the underlying population at the initial sampling time point is taken to be the observed allele frequency of the sample or is assumed to be uniformly distributed. In Malaspinas et al. (2012), the transition probabilities of the allele frequencies are calculated by approximating the Wright-Fisher diffusion with a one-step Markov process on a grid. Steinrücken et al. (2014) presented an extension of the HMM framework of Bollback et al. (2008) by capitalising on a spectral representation of the Wright-Fisher diffusion introduced by Song & Steinrücken (2012). The approach of Steinrücken et al. (2014) allows for a more general diploid model of natural selection such as the case of under- or overdominance, which, however, is unable to infer allele age as recurrent mutation is allowed in their model. Malaspinas (2016) provided an excellent review of existing approaches for studying natural selection with aDNA samples.

More recently, Schraiber et al. (2016) developed a Bayesian method under the HMM framework of Bollback et al. (2008) for the joint inference of natural selection and allele age from allele frequency time series data. Their key innovation is to apply a high-frequency path augmentation method to circumvent the difficulty inherent in calculating the transition probabilities of the allele frequencies under the Wright-Fisher diffusion. Markov chain Monte Carlo (MCMC) techniques were used to integrate over all possible allele frequency trajectories of the underlying population consistent with the observations. The computational advantage of the path augmentation approach allows for general diploid models of natural selection and changing population sizes. However, to update the sample paths of the Wright-Fisher diffusion bridge, they used Bessel bridges of order four, which is somewhat challenging from both mathematical and programming points of view.

In this work, we propose a novel likelihood-based method for co-estimating the selection coefficient and the allele age from time series data of allele frequencies. Our approach is also an

extension of Bollback et al. (2008), but unlike existing methods, we condition the Wright-Fisher diffusion to survive until the time of the most recent sample and take it to be the underlying process in our HMM framework. The conditioned Wright-Fisher diffusion avoids the somewhat arbitrary initial condition that the allele is created by mutation at a certain small frequency, *e.g.*, as used in Schraiber et al. (2016). Our likelihood calculation is carried out by numerically solving the KBE resulting from the conditioned Wright-Fisher diffusion backwards in time and re-weighting the solution by the emission probabilities of the observation at each sampling time point. The values of the re-weighted solution at frequency 0 give the likelihood of the selection coefficient and the allele age. For each fixed selection coefficient, the likelihood for all values of the allele age can be obtained by numerically solving the KBE only once. This advance enables a reduction of the two-dimensional numerical search for the maximum of the likelihood surface for the selection coefficient and the allele age, *e.g.*, as used in Malaspinas et al. (2012), to a one-dimensional numerical search over the selection coefficient.

We evaluate the performance of our method with extensive simulations and show that our approach allows for efficient and accurate estimation of natural selection and allele age from allele frequency time series data, even if the samples are sparsely distributed in time with small uneven sizes. We employ our method to re-analyse the time serial samples of segregating alleles associated with horse coat colouration from previous studies of Ludwig et al. (2009), Pruvost et al. (2011) and Wutke et al. (2016). Unlike these earlier studies, our analysis is performed on the raw samples (62 sampling time points) rather than the grouped samples (9 sampling time points) with an empirical study showing that grouping aDNA samples can significantly alter the results of the inference of natural selection and allele age. Our findings suggest that horse base coat colour variation could be associated with adaptation to the climate change caused by the transition from a glacial period to an interglacial period.

2. Materials and Methods

In this section, we begin with a brief review of the Wright-Fisher diffusion for a single locus evolving subject to natural selection and then derive the Wright-Fisher diffusion conditioned to survive until a given time point. We describe our likelihood-based method for co-estimating the selection coefficient and the allele age from allele frequency time series data, *e.g.*, how to set

up the HMM framework incorporating with the conditioned Wright-Fisher diffusion and how to compute the likelihood for the population genetic quantities of interest.

2.1. Wright-Fisher diffusion

Let us consider a population of N randomly mating diploid individuals at a single locus $\mathcal{A}$ evolving under natural selection according to the Wright-Fisher model (see, *e.g.*, Durrett, 2008, for more details). We assume discrete time, nonoverlapping generations and fixed population size. Suppose that there are two possible allele types at locus $\mathcal{A}$, labelled $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively. The symbol $\mathcal{A}_1$ is attached to the mutant allele, which is assumed to arise only once in the population and be favoured by natural selection. The symbol $\mathcal{A}_2$ is attached to the ancestral allele, which is assumed to originally exist in the population. We assume that natural selection takes the form of viability selection, where the viability is fixed from the time when the mutant allele arose. We take the relative viabilities of the three possible genotypes $\mathcal{A}_1\mathcal{A}_1$, $\mathcal{A}_1\mathcal{A}_2$ and $\mathcal{A}_2\mathcal{A}_2$ at locus $\mathcal{A}$ to be 1, $1-hs$ and $1-s$, respectively, where $s \in [0,1]$ is the selection coefficient and $h \in [0,1]$ is the dominance parameter.

2.1.1. Wright-Fisher diffusion with selection

We consider the standard diffusion limit of the Wright-Fisher model with selection, which requires the selection coefficient s to be of the order of $\frac{1}{2N}$ such that the scaled selection coefficient $\alpha = 2Ns$ is kept constant as the population size N goes to infinity. By an argument in Durrett (2008), the allele frequency trajectory through time converges to the diffusion limit of the Wright-Fisher model if we measure time in units of $2N$ generations and let the population size N approach infinity. We refer to this limit as the Wright-Fisher diffusion with selection, which has already been successfully applied as an approximation of the Wright-Fisher model in the population genetic analysis of aDNA data (*e.g.*, Bollback et al., 2008; Malaspinas et al., 2012; Steinrücken et al., 2014; Schraiber et al., 2016).

Let X denote the Wright-Fisher diffusion with selection, which models the frequency of the mutant allele evolving in the state space $[0,1]$ under natural selection after the mutant allele arose in the population at time t_0. According to Durrett (2008), the Wright-Fisher diffusion X has the infinitesimal generator

$$\mathcal{L} = a(t,x)\frac{\partial}{\partial x} + \frac{1}{2}b(t,x)\frac{\partial^2}{\partial x^2},$$

where the drift term

$$a(t,x) = \alpha x(1-x)((1-h)-(1-2h)x),$$

and the diffusion term

$$b(t,x) = x(1-x).$$

The transition probability density function of the Wright-Fisher diffusion X, defined by

$$p(t,x,t',x') = \lim_{\Delta x \to 0} \frac{1}{\Delta x} \mathbb{P}\left(X(t') \in \left(x' - \frac{\Delta x}{2}, x' + \frac{\Delta x}{2} \right) \mid X(t) = x \right)$$

for $t_0 \leq t < t'$, can then be expressed as the solution $u(t,x)$ of the corresponding KBE

$$\frac{\partial}{\partial t} u(t,x) + \mathcal{L} u(t,x) = 0 \tag{1}$$

with terminal condition $u(t',\cdot) = \delta(\cdot - x')$, where δ is a Dirac delta function.

2.1.2. Conditioned Wright-Fisher diffusion with selection

As observed in Valleriani (2016), it is desirable to account for conditioning in most aDNA analyses since a single trajectory of allele frequencies through time available in aDNA results in the very limited coverage of the fitness landscape. Valleriani (2016) conditioned both the initial and the final values of the Wright-Fisher model to be fixed, assuming a finite population with perfect sampling, which, however, is not completely realistic. We condition the Wright-Fisher diffusion X to have survived until the time of the most recent sample in this work, *i.e.*, the mutant allele frequency of the underlying population at the most recent sampling time point is strictly positive.

We let X^* be the Wright-Fisher diffusion X conditioned to survive, *i.e.*, the frequency of the mutant allele stays in the interval $(0,1]$, until at least time τ after the mutant allele arose in the population at time t_0, and define the transition probability density function of the conditioned Wright-Fisher diffusion X^* as

$$q_\tau(t,x,t',x') = \lim_{\Delta x \to 0} \frac{1}{\Delta x} \mathbb{P}\left(X^*(t') \in \left(x' - \frac{\Delta x}{2}, x' + \frac{\Delta x}{2} \right) \mid X^*(t) = x \right) \tag{2}$$

for $t_0 \leq t < t' < \tau$. To get the infinitesimal generator of the conditioned Wright-Fisher diffusion

X^*, we need to specify the drift term, denoted by $a^*(t,x)$, and the diffusion term, denoted by $b^*(t,x)$, which are the first and second infinitesimal moments of the conditioned Wright-Fisher diffusion X^*, respectively, *i.e.*,

$$a^*(t,x) = \lim_{t'\to t} \frac{1}{t'-t} \int_0^1 (x'-x) q_\tau(t,x,t',x')\, dx' \tag{3}$$

$$b^*(t,x) = \lim_{t'\to t} \frac{1}{t'-t} \int_0^1 (x'-x)^2 q_\tau(t,x,t',x')\, dx'. \tag{4}$$

From Eq. (2), we can formulate the transition probability density function of the conditioned Wright-Fisher diffusion X^* in terms of the transition probability density function of the Wright-Fisher diffusion X as

$$q_\tau(t,x,t',x') = p(t,x,t',x') \frac{P_\tau(t',x')}{P_\tau(t,x)}, \tag{5}$$

where

$$P_\tau(t,x) = \mathbb{P}\left(X(\tau) > 0 \mid X(t) = x\right)$$

is the probability that the Wright-Fisher diffusion X, starting from x at time t, survives until at least time τ. The probability of survival, $P_\tau(t,x)$, for $t_0 \le t < \tau$ can be expressed as the solution to the KBE in Eq. (1) with terminal condition $P_\tau(\tau, x_\tau) = \mathbb{1}_{\{x_\tau > 0\}}$, where $\mathbb{1}_A$ is the indicator function that equals to 1 if condition A holds and 0 otherwise. It can be calculated by

$$P_\tau(t,x) = \int_0^1 p(t,x,\tau,x_\tau) P_\tau(\tau,x_\tau)\, dx_\tau \tag{6}$$

due to linearity, where $p(t,x,\tau,x_\tau)$ is the solution to the KBE in Eq. (1) with terminal condition $p(\tau,\cdot) = \delta(\cdot - x_\tau)$.

Substituting Eq. (5) into Eqs. (3) and (4), we have

$$a^*(t,x) = \lim_{t'\to t} \frac{1}{t'-t} \int_0^1 \frac{P_\tau(t',x')}{P_\tau(t,x)} (x'-x) p(t,x,t',x')\, dx'$$

$$b^*(t,x) = \lim_{t'\to t} \frac{1}{t'-t} \int_0^1 \frac{P_\tau(t',x')}{P_\tau(t,x)} (x'-x)^2 p(t,x,t',x')\, dx'.$$

Taylor expansion yields

$$\frac{P_\tau(t',x')}{P_\tau(t,x)} = 1 + \left(\frac{\partial}{\partial x} \log P_\tau(t,x)\right)(x'-x) + \mathcal{O}(|t'-t|) + \mathcal{O}(|x'-x|^2),$$

which results in

$$a^*(t,x) = \lim_{t'\to t} \frac{1}{t'-t} \int_0^1 (x'-x)p(t,x,t',x') + \left(\frac{\partial}{\partial x} \log P_\tau(t,x)\right)(x'-x)^2 p(t,x,t',x')\,dx'$$
$$b^*(t,x) = \lim_{t'\to t} \frac{1}{t'-t} \int_0^1 (x'-x)^2 p(t,x,t',x') + \left(\frac{\partial}{\partial x} \log P_\tau(t,x)\right)(x'-x)^3 p(t,x,t',x')\,dx'.$$

As shown in Durrett (2008), the transition probability density function of the Wright-Fisher diffusion X satisfies

$$\lim_{t'\to t} \frac{1}{t'-t} \int_0^1 (x'-x)p(t,x,t',x')\,dx' = a(t,x)$$
$$\lim_{t'\to t} \frac{1}{t'-t} \int_0^1 (x'-x)^2 p(t,x,t',x')\,dx' = b(t,x)$$
$$\lim_{t'\to t} \frac{1}{t'-t} \int_0^1 (x'-x)^k p(t,x,t',x')\,dx' = 0, \quad \text{for } k \geq 3,$$

which gives rise to

$$a^*(t,x) = a(t,x) + b(t,x)\frac{\partial}{\partial x} \log P_\tau(t,x)$$
$$b^*(t,x) = b(t,x).$$

Therefore, the infinitesimal generator of the conditioned Wright-Fisher diffusion X^* can be written as

$$\mathcal{L}_\tau^* = a^*(t,x)\frac{\partial}{\partial x} + \frac{1}{2}b^*(t,x)\frac{\partial^2}{\partial x^2},$$

where the drift term

$$a^*(t,x) = \alpha x(1-x)((1-h)-(1-2h)x) + x(1-x)\frac{\partial}{\partial x} \log P_\tau(t,x),$$

and the diffusion term

$$b^*(t,x) = x(1-x).$$

The transition probability density function of the conditioned Wright-Fisher diffusion X^* can then be expressed as the solution $u(t,x)$ of the corresponding KBE

$$\frac{\partial}{\partial t}u(t,x) + \mathcal{L}_\tau^* u(t,x) = 0 \tag{7}$$

with terminal condition $u(t', \cdot) = \delta(\cdot - x')$.

2.2. Maximum likelihood estimation

Recall that the mutant allele at the locus we consider is subject to natural selection. Suppose that the available data are always sampled from the underlying population at a finite number of distinct time points, say $t_1 < t_2 < \ldots < t_K$, where the time is measured in units of $2N$ generations to be consistent with the Wright-Fisher diffusion time scale. At the sampling time point t_k, let c_k be the number of mutant alleles observed in a sample of n_k chromosomes drawn from the underlying population. In this work, the population genetic quantities of interest are the scaled selection coefficient α, the dominance parameter h and the allele age t_0.

2.2.1. Hidden Markov model

To our knowledge, Malaspinas et al. (2012) and Schraiber et al. (2016) are the only existing works that seek to jointly infer natural selection and allele age from time series data of allele frequencies. Both these methods were built on the HMM framework of Bollback et al. (2008) incorporating the Wright-Fisher diffusion with selection. In order to co-estimate the allele age, initial conditions for the Wright-Fisher diffusion X at time t_0 must be specified. Schraiber et al. (2016) took the mutant allele frequency $X(t_0)$ to be some small but arbitrary value, which was found to be feasible in their method but is slightly unsatisfying. Malaspinas et al. (2012) took the mutant allele frequency $X(t_0)$ to be $\frac{1}{2N}$, which corresponds to the case that the positively selected allele arose as a *de novo* mutation. This can be slightly problematic in that the Wright-Fisher diffusion X may hit frequency $\frac{1}{2N}$ again after the mutation arises, so the time t when $X(t) = \frac{1}{2N}$ may not be the same as the allele age.

Similar to Bollback et al. (2008), our method also relies on an HMM framework except that the underlying population evolves according to the conditioned Wright-Fisher diffusion X^* rather than the Wright-Fisher diffusion X. Given the mutant allele frequency trajectory of the underlying population, the observations are modelled as independent binomial samples drawn from the underlying population at each sampling time point. The conditioning enables us to attach an initial condition $X^*(t_0) = 0$ with 0 forming an entrance boundary (see Appendix A), *i.e.*, the conditioned Wright-Fisher diffusion X^* will reach the interval $(0, 1]$ starting from the initial mutant allele frequency $X^*(t_0) = 0$. Our setup allows us to avoid specifying any arbitrary mutant allele frequency of the underlying population at time t_0.

More specifically, our HMM framework can be fully captured by a bivariate Markov process $\{(X^*(t), C(t)) : t \in [t_1, t_K]\}$ with initial condition $X^*(t_0) = 0$, where the unobserved process $X^*(t)$ for $t \in [t_1, t_K]$ is the Wright-Fisher diffusion conditioned to survive until the most recent sampling time point t_K, and the observed process $C(t)$ for $t \in \{t_1, t_2, \ldots, t_K\}$ is a sequence of conditionally independent binomial random variables given the unobserved process $X^*(t)$ at each sampling time point. The transition probabilities for our HMM between two consecutive sampling time points t_{k-1} and t_k are

$$\mathbb{P}\left(X^*(t_k) \in dx_k \mid X^*(t_{k-1}) = x_{k-1}; \alpha, h\right) = q_\tau(t_{k-1}, x_{k-1}, t_k, x_k \mid \alpha, h)\, dx_k$$

for $k = 2, 3, \ldots, K$, where $q_\tau(t_{k-1}, x_{k-1}, t_k, x_k \mid \alpha, h)$ is the transition probability density function of the conditioned Wright-Fisher diffusion X^* that satisfies the KBE in Eq. (7) with terminal condition $q_\tau(t_k, \cdot) = \delta(\cdot - x_k)$. The emission probabilities for our HMM at the sampling time point t_k are

$$\mathbb{P}\left(C(t_k) = c_k \mid X^*(t_k) = x_k; \alpha, h\right) = b(c_k; n_k, x_k)$$

for $k = 1, 2, \ldots, K$, where

$$b(c_k; n_k, x_k) = \frac{n_k!}{c_k!(n_k - c_k)!} x_k^{c_k} (1 - x_k)^{n_k - c_k}$$

is the formula for the binomial distribution.

2.2.2. Likelihood computation

To calculate the likelihood for the population genetic parameters of interest, defined by

$$\ell(\alpha, h, t_0 \mid \boldsymbol{c}_{1:K}) = \mathbb{P}\left(C(t_1) = c_1, \ldots, C(t_K) = c_K \mid \alpha, h, t_0\right),$$

we let $\mathcal{T}_0 = (-\infty, t_1)$ and $\mathcal{T}_{k-1} = [t_{k-1}, t_k)$ for $k = 2, 3, \ldots, K$, and decompose the likelihood to a sum of terms according to which time interval $\mathcal{T}_{k-1}$ the allele age t_0 falls in,

$$\begin{aligned}\ell(\alpha, h, t_0 \mid \boldsymbol{c}_{1:K}) = {} & g_0(\alpha, h, t_0 \mid \boldsymbol{c}_{1:K}) \mathbb{1}_{\{t_0 \in \mathcal{T}_0\}} \\ & + \sum_{k=2}^{K} f_{k-1}(\alpha, h, t_0 \mid \boldsymbol{c}_{1:k-1}) g_{k-1}(\alpha, h, t_0 \mid \boldsymbol{c}_{k:K}) \mathbb{1}_{\{t_0 \in \mathcal{T}_{k-1}\}}. \quad (8)\end{aligned}$$

Note that in the decomposition of Eq. (8), only one term will be nonzero since the allele age t_0 can only be in one of the time intervals $\mathcal{T}_{k-1}$ for $k = 1, 2, \ldots, K$. In Eq. (8), if the allele age $t_0 \in \mathcal{T}_{k-1}$, then

$$f_{k-1}(\alpha, h, t_0 \mid \boldsymbol{c}_{1:k-1}) = \mathbb{P}\left(C(t_1) = c_1, \ldots, C(t_{k-1}) = c_{k-1} \mid \alpha, h, t_0\right)$$

and

$$g_{k-1}(\alpha, h, t_0 \mid \boldsymbol{c}_{k:K}) = \mathbb{P}\left(C(t_k) = c_k, \ldots, C(t_K) = c_K \mid \alpha, h, t_0\right)$$

are the probabilities for the observations $\boldsymbol{c}_{1:k-1}$ and $\boldsymbol{c}_{k:K}$ before and after the time t_0 that the mutant allele arose in the underlying population, respectively. As the conditioned Wright-Fisher diffusion $X^*(t) = 0$ for $t \in (-\infty, t_0]$, we have

$$f_{k-1}(\alpha, h, t_0 \mid \boldsymbol{c}_{1:k-1}) = \prod_{i=1}^{k-1} b(c_i; n_i, 0) = \prod_{i=1}^{k-1} \mathbb{1}_{\{c_i=0\}}.$$

Under our HMM framework, we have

$$\begin{aligned} g_{k-1}(\alpha, h, t_0 \mid \boldsymbol{c}_{k:K}) = \int_0^1 \cdots \int_0^1 & q_\tau(t_0, 0, t_k, x_k \mid \alpha, h) b(c_k; n_k, x_k) \\ & \cdot \prod_{i=k+1}^{K} q_\tau(t_{i-1}, x_{i-1}, t_i, x_i \mid \alpha, h) b(c_i; n_i, x_i)\, dx_k \cdots dx_K. \end{aligned}$$

We define

$$\beta_{k-1}(t, x \mid \alpha, h) = \mathbb{P}\left(C(t_k) = c_k, \ldots, C(t_K) = c_K \mid X(t) = x; \alpha, h\right) \tag{9}$$

to be the probability of the observations $\boldsymbol{c}_{k:K}$ given initial condition $X^*(t) = x$ for $t \in \mathcal{T}_{k-1}$, which can be calculated recursively by

$$\beta_{k-1}(t, x \mid \alpha, h) = \int_0^1 q_\tau(t, x, t_k, x_k \mid \alpha, h) b(c_k; n_k, x_k) \beta_k(t_k, x_k \mid \alpha, h)\, dx_k \tag{10}$$

for $k = K, K-1, \ldots, 1$, given terminal condition

$$\beta_K(t_K, x_K \mid \alpha, h) = \mathbb{1}_{\{x_K > 0\}}.$$

The recursive formula in Eq. (10) implies that the probability $\beta_{k-1}(t, x \mid \alpha, h)$ for $t \in \mathcal{T}_{k-1}$ can be expressed as the solution to the KBE in Eq. (7) with terminal condition

$$\beta_{k-1}(t_k, x_k \mid \alpha, h) = b(c_k; n_k, x_k)\beta_k(t_k, x_k \mid \alpha, h). \tag{11}$$

Using Eq. (9), we have

$$g_{k-1}(\alpha, h, t_0 \mid \boldsymbol{c}_{k:K}) = \beta_{k-1}(t_0, 0 \mid \alpha, h) \tag{12}$$

for $k = 1, 2, \ldots, K$.

We let

$$k' = \min\{k \in \{1, 2, \ldots, K\} : c_k > 0\}.$$

Substituting Eq. (12) into Eq. (8), we can formulate the likelihood for the population genetic parameters of interest as

$$\ell(\alpha, h, t_0 \mid \boldsymbol{c}_{1:K}) = \beta_0(t_0, 0 \mid \alpha, h)\mathbb{1}_{\{t_0 \in \mathcal{T}_0\}} + \sum_{k=2}^{k'} \beta_{k-1}(t_0, 0 \mid \alpha, h)\mathbb{1}_{\{t_0 \in \mathcal{T}_{k-1}\}}. \tag{13}$$

From the calculation leading to Eq. (13), the likelihood for all possible values of the allele age t_0 with fixed values of the scaled selection coefficient α and the dominance parameter h can be obtained by recursively calculating the probabilities $\beta_{k-1}(t, x \mid \alpha, h)$ for $t \in \mathcal{T}_{k-1}$ and $x \in [0, 1]$ for $k = K, K-1, \ldots, 1$ with Eqs. (10). At time t_K, the time of the most recent sample, we re-weight the probabilities $\beta_K(t_K, x_K \mid \alpha, h)$ according to the emission probabilities $b(c_K; n_K, x_K)$ to update the terminal condition $\beta_{K-1}(t_K, x_K \mid \alpha, h)$ with Eq. (11) and then numerically solve the KBE in Eq. (7) backwards in time until time t_{K-1} to get the probabilities $\beta_{K-1}(t, x \mid \alpha, h)$ for $t \in \mathcal{T}_{K-1}$ and $x \in [0, 1]$. At time t_{K-1}, we update the terminal condition $\beta_{K-2}(t_{K-1}, x_{K-1} \mid \alpha, h)$ and then carry on solving the KBE in Eq. (7) backwards in time until time t_{K-2} to obtain the probabilities $\beta_{K-2}(t, x \mid \alpha, h)$ for $t \in \mathcal{T}_{K-2}$ and $x \in [0, 1]$. We carry on this procedure until most of the probability mass has been absorbed at frequency 0.

In order to numerically solve the KBE in Eq. (7) for the conditioned Wright-Fisher diffusion X^*, we first need to compute the survival probability $P_\tau(t, x)$ in Eq. (6) for $t \in (-\infty, t_K]$ and $x \in [0, 1]$. For this, we numerically solve the KBE in Eq. (1) for the Wright-Fisher diffusion X. Then we use the values of the survival probability $P_\tau(t, x)$ to numerically solve the KBE

in Eq. (7). Note that we take the drift term $a^*(t,0) = \lim_{x\downarrow 0} x\frac{\partial}{\partial x}\log P_\tau(t,x)$ to be 1, which is justified in Appendix A.

For clarity, we write down the procedure we follow to obtain the likelihood $l(\alpha, h, t_0 \mid \boldsymbol{c}_{1:K})$ for all possible values of the allele age $t_0 \in (-\infty, t_K]$ with fixed values of the scaled selection coefficient α and the dominance parameter h:

Step 1: Calculate $P_\tau(t,x)$ for $t \in (-\infty, t_K)$ and $x \in [0,1]$ by numerically solving the KBE in Eq. (1) backwards in time with terminal condition $P_\tau(t_K, x_K) = \mathbb{1}_{\{x_K>0\}}$ for $x_K \in [0,1]$ and boundary conditions $P_\tau(t,0) = 0$ and $P_\tau(t,1) = 1$ for $t \in (-\infty, t_K)$.

Step 2: Calculate $\frac{\partial}{\partial x}\log P_\tau(t,x)$ for $t \in (-\infty, t_K)$ and $x \in [0,1]$ by numerically differentiating $\log P_\tau(t,x)$.

Step 3: Set $k = K$ and repeat until $k = 0$:

- Step 3a: Update the terminal condition $\beta_{k-1}(t_k, x_k \mid \alpha, h)$ with Eq. (11).
- Step 3b: Calculate $\beta_{k-1}(t, x \mid \alpha, h)$ for $t \in \mathcal{T}_{k-1}$ and $x \in [0,1]$ by numerically solving the KBE in Eq. (7) backwards in time with terminal condition $\beta_{k-1}(t_k, x_k \mid \alpha, h)$ for $x_k \in [0,1]$ and boundary condition $\beta_{k-1}(t, 0 \mid \alpha, h) = 1$ if $\beta_{k-1}(t, x \mid \alpha, h) > 0$ for $x \in (0,1]$ and stop Step 3 otherwise.
- Step 3d: Set $k = k - 1$.

Step 4: Combine $\beta_{k-1}(t_0, 0 \mid \alpha, h)$ for $k = 1, 2, \ldots, k'$ using Eq. (13) to obtain $l(\alpha, h, t_0 \mid \boldsymbol{c}_{1:K})$.

To obtain the maximum of the likelihood for the population genetic quantities of interest, we perform a section search over the scaled selection coefficient and the dominance parameter. We start with a fixed region for the possible values of the scaled selection coefficient and the dominance parameter. For each pair of the m equally spaced scaled selection coefficients, $\alpha_1, \alpha_2, \ldots, \alpha_m$, and the n equally spaced dominance parameters, $h_1, h_2, \ldots, h_n$, within the region, we perform the procedure laid out above to obtain the likelihood for all possible values of the allele age. We record the maximum value of the likelihood as well as the allele age where this maximum is attained. We assume the scaled selection coefficient $\alpha_{i'}$ and the dominance parameter $h_{j'}$ yield the highest likelihood amongst the combinations of the scaled selection coefficient α_i for $i = 1, 2, \ldots m$ and the dominance parameter h_j for $j = 1, 2, \ldots n$. If the highest likelihood is achieved at an interior point, in the next step, we narrow our search region to

$[\alpha_{i'-1}, \alpha_{i'+1}] \times [h_{j'-1}, h_{j'+1}]$. Otherwise, *i.e.*, if the highest likelihood is achieved at a boundary point, we shift the search grid to centre at the parameters values where the likelihood achieves its maximum from the previous step. We continue the procedure until the area of the search region is sufficiently small.

2.3. Data availability

Supplemental material available at FigShare. Source code for the method described in this work is available at `https://github.com/zhangyi-he/WFD-1L-MLE-KBE`.

3. Results

In this section, we first evaluate the performance of our method through a number of simulated datasets with given population genetic parameter values. We then apply our approach to analyse aDNA time series data of allele frequencies associated with horse coat colouration from previous studies of Ludwig et al. (2009), Pruvost et al. (2011) and Wutke et al. (2016), where eight coat colour loci from horse samples ranging from a pre- to a post-domestication period were sequenced.

3.1. Analysis of simulated data

To test the accuracy of our approach, we simulate sets of data with different population genetic parameter values. For each simulated dataset, we fix the population size N, the selection coefficient s, the dominance parameter h and the allele age k_0 (measured in the Wright-Fisher model time scale), and use the Wright-Fisher model with selection (see, *e.g.*, Durrett, 2008) to simulate the mutant allele frequency trajectory of the underlying population starting at frequency $\frac{1}{2N}$. Given the simulated mutant allele frequency trajectory of the underlying population, we generate the mutant allele count of the sample independently at each sampling time point according to the binomial distribution. We only keep the simulated datasets where the mutant allele survives in the underlying population until the time of the most recent sample.

Like Malaspinas et al. (2012), we consider two different sampling schemes in our empirical studies to quantify whether it is better to sample more chromosomes at fewer time points, or the opposite. In the first sampling scheme, called sampling scheme A, we draw 20 chromosomes at $K = 60$ evenly spaced sampling time points from the underlying population, whereas in

the second sampling scheme, called sampling scheme B, we draw 120 chromosomes at $K = 10$ evenly spaced sampling time points from the underlying population. For each sampling scheme, we pick a population size of $N = 16000$ and fix the selection coefficient and the dominance parameter to several potential values: $s \in \{0, 0.0025, 0.005, 0.01, 0.015, 0.02\}$ and $h \in \{0, 0.5, 1\}$.

Due to the vastly different values of the selection coefficient and the dominance parameter, it is not possible to use a single set of fixed sampling times for all simulated datasets. Instead, for each pair of given values of the selection coefficient and the dominance parameter, we perform the following steps to determine the sampling time points:

Step 1: Generate 100 realisations of the mutant allele frequency trajectory of the underlying population using the Wright-Fisher model with selection.

Step 2: Take the mean of the 100 realisations of the mutant allele frequency trajectory of the underlying population.

Step 3: Set the time when the mean mutant allele frequency trajectory of the underlying population first hits frequency 0.1 and 0.9 to be the first and last sampling time points, labelled k_1 and k_K, respectively.

Step 4: Set the rest of the sampling time points $k_2, k_3 \ldots, k_{K-1}$ to be evenly spaced between the first and last sampling time points k_1 and k_K.

Under this scheme, a significant number of the realisations of the mutant allele frequency trajectory of the underlying population will capture a significant proportion of the selective sweep. See Figure 1 for the simulated datasets for each sampling scheme as representative examples.

We illustrate the boxplot results for the resulting estimates for these simulated datasets across different parameter ranges and sampling schemes in Figures 2-4. The tips of the whiskers denote the 2.5%-quantile and the 97.5%-quantile, and the boxes represent the first and third quartile with the median in the middle. The bias and the root mean square error (RMSE) of the resulting estimates are summarised in Supplemental Material, Tables S1-S6.

As shown in Figures 2-4, we find that the estimates for both the selection coefficient and the allele age show little bias across different parameter ranges and sampling schemes, although one can discern a slight positive bias for the estimates of the allele age. Some degree of bias is not unexpected, given that there is no requirement for the maximum likelihood estimator to be unbiased in general. An additional cause of bias arises from the finite grid size when we

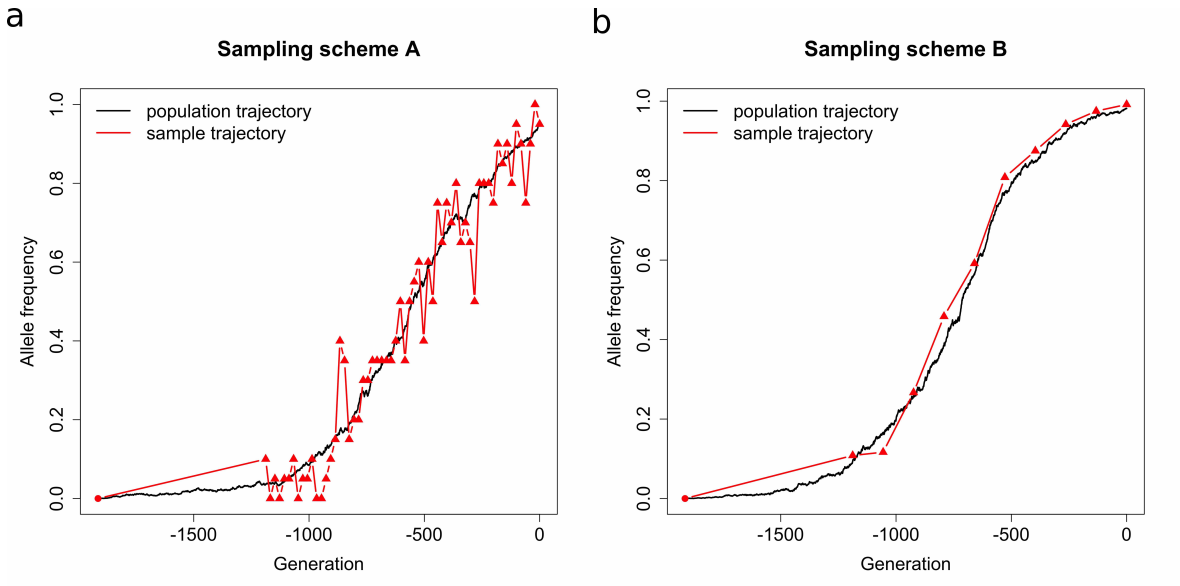

Figure 1: Representative examples of the simulated dataset for each sampling scheme with the selection coefficient $s = 0.01$, the dominance parameter $h = 0.5$ and the population size $N = 16000$. The mutant allele arises at frequency $x_0 = 3.125 \times 10^{-5}$ in the underlying population at time $k_0 = -1916$ (red filled circle) and simulate the mutant allele frequency trajectory of the underlying population is simulated using the Wright-Fisher model with selection (black line). (a) Sampling scheme A: 20 chromosomes drawn from the underlying population every 19 generations (red filled triangle) from generation -1188 to 0. (b) Sampling scheme B: 120 chromosomes drawn from the underlying population every 118 generations (red filled triangle) from generation -1188 to 0.

numerically solve the KBE in Eq. (7). Numerically solving the KBE with a finer grid results in a smaller bias, although the numerical solution takes much longer to run. With the increase of the selection coefficient, the uncertainty of the estimate for the selection coefficient increases whereas the uncertainty of the estimate for the allele age decreases. This can be explained by the relatively smaller noise with larger selection coefficients, yielding more information about the mutant allele frequency trajectory of the underlying population prior to the first sample time point. Moreover, the 2.5%-quantiles of the empirical distributions of the estimates for the selection coefficient $s > 0$ are all larger than the 97.5%-quantile of the empirical distribution of the estimates for the selection coefficient $s = 0$ (see Supplemental Material, Figure S1). This illustrates that our method has a strong power to reject neutrality.

Comparing boxplot results for different selection schemes, we find that there are many more outliers found in the case of recessive selection ($h = 1$, Figure 4) than in the cases of dominant and genic selection ($h = 0$ and $h = 0.5$, Figures 2 and 3). To understand this effect, we plot the simulated mutant allele frequency trajectories of the underlying population for all of the simulated datasets in Figure 5. In order to capture information on the selection coefficient and the allele age, the underlying trajectory should ideally grow from a lower frequency around 0 to

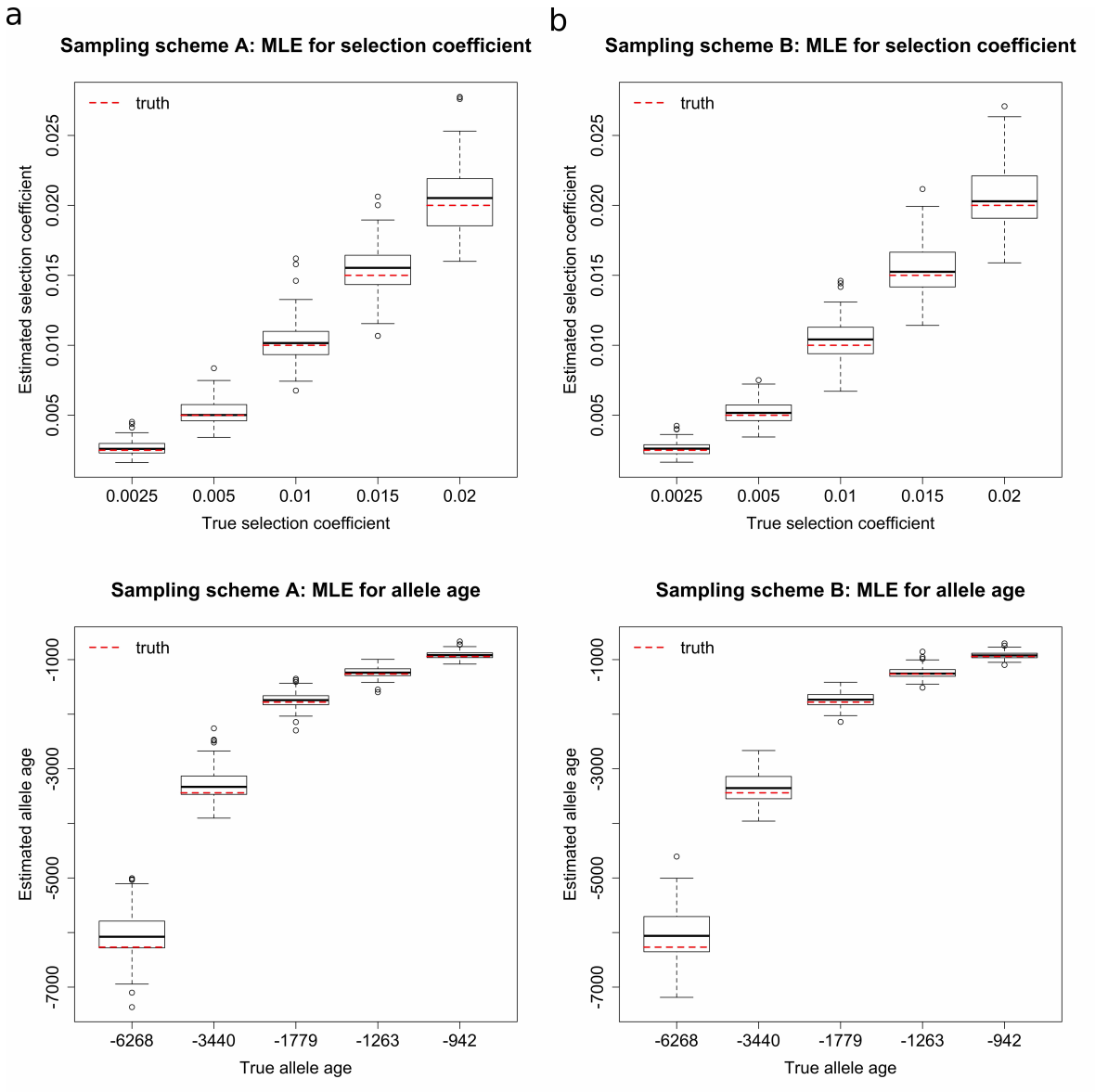

Figure 2: Empirical distributions of the estimates for 100 datasets simulated for the case of dominant selection where the dominance parameter $h = 0$. (a) Boxplots of the estimates for the sampling scheme A. (b) Boxplots of the estimates for the sampling scheme B.

a high frequency around 1. However, as can be observed from Figure 5, there are a number of simulated mutant allele frequency trajectories of the underlying population where the mutant allele frequencies at all sampling time points are all near the absorbing boundaries, 0 or 1. This effect is especially pronounced for the case of recessive selection ($h = 1$). Such mutant allele frequency trajectories of the underlying population contain little information about the underlying selection coefficient and allele age, therefore more outliers when we try to estimate these parameters.

Comparing the boxplot results for different sampling schemes, we can see similar results for

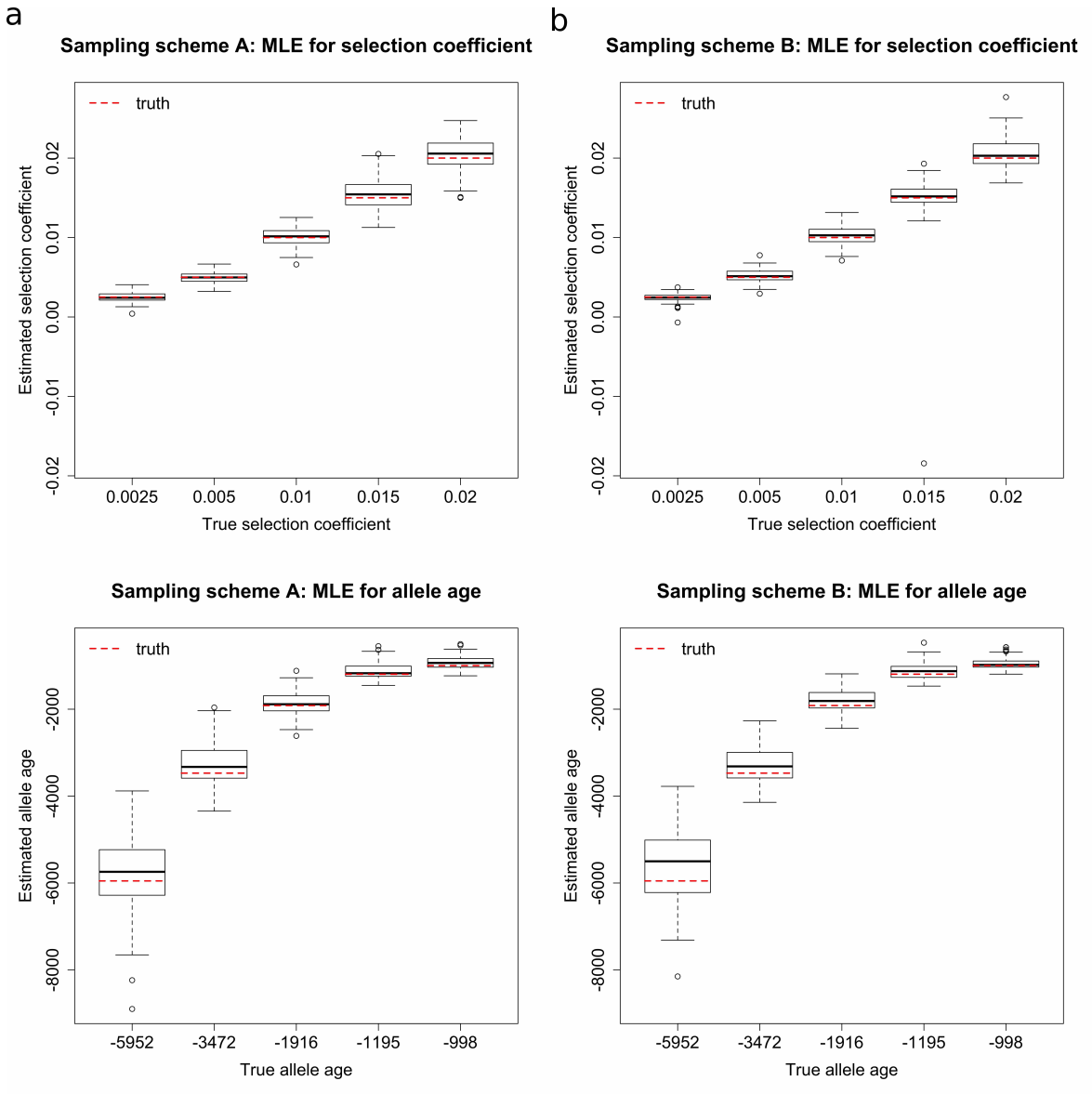

Figure 3: Empirical distributions of the estimates for 100 datasets simulated for the case of genic selection where the dominance parameter $h = 0.5$. (a) Boxplots of the estimates for the sampling scheme A. (b) Boxplots of the estimates for the sampling scheme B.

estimates of the selection coefficient, but fewer sampled chromosome and more sampling times lead to better results for the allele age. This is because increasing sampling times will reduce the time gap between the time that the mutant allele arose in the underlying population and the first sampling time point, which provide more information on the allele age.

In conclusion, our method can deliver accurate estimates of the selection coefficient and the allele age using time series data of allele frequencies across different parameter ranges and sampling schemes. Better estimates for the allele age can be achieved with larger values of the selection coefficient. Different sampling schemes have a minimal effect on the estimation of the

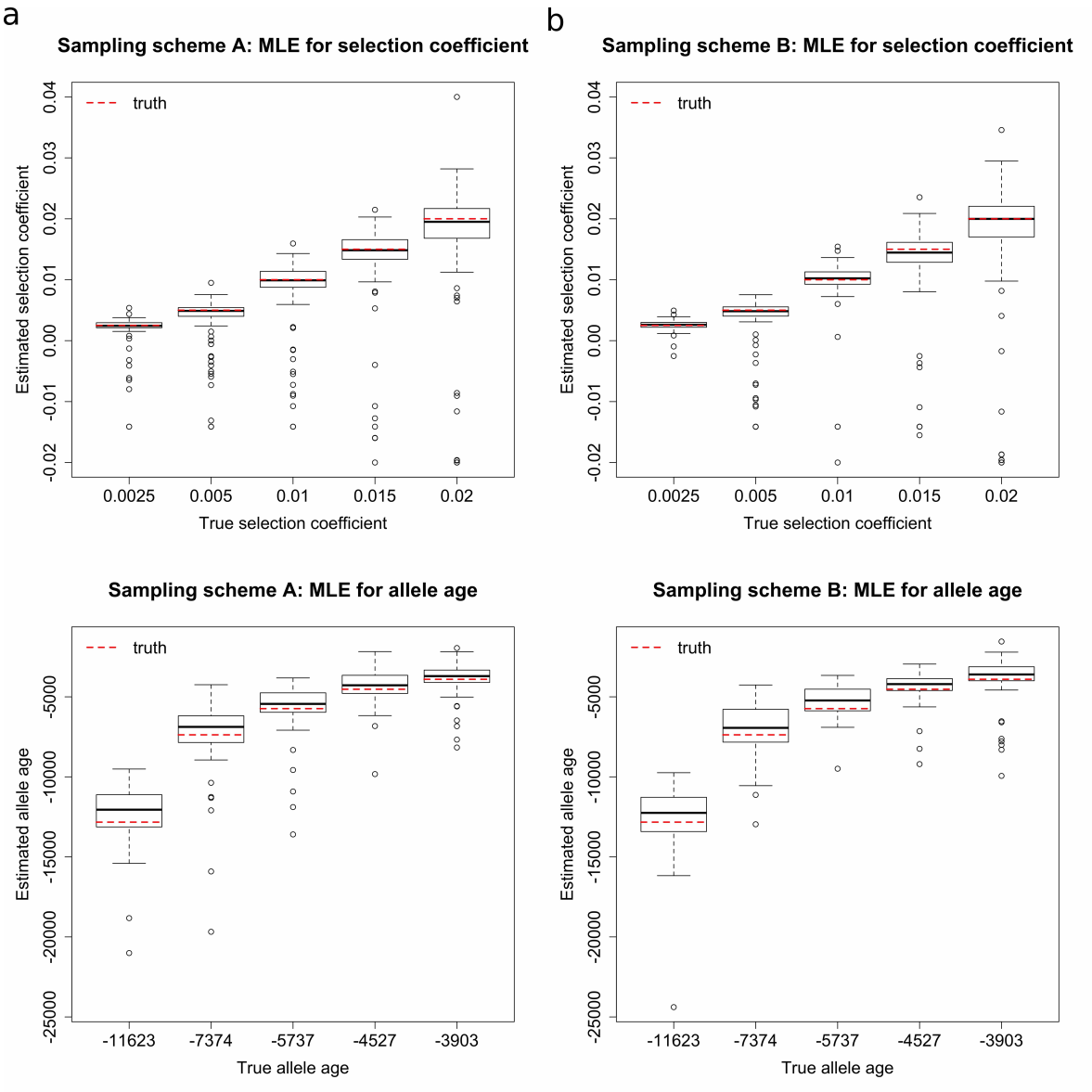

Figure 4: Empirical distributions of the estimates for 100 datasets simulated for the case of recessive selection where the dominance parameter $h = 1$. (a) Boxplots of the estimates for the sampling scheme A. (b) Boxplots of the estimates for the sampling scheme B.

selection coefficient, but sampling fewer chromosomes at more time points can lead to better results for the allele age. In these empirical studies above, the dominance parameter is treated as known. We also present the results where the dominance parameter is estimated along with the selection coefficient and the allele age (see Supplemental Material, Tables S7 and S8).

3.2. Analysis of real data

We illustrate the utility of our method on real data by re-analysing the aDNA data associated with horse coat colouration. Ludwig et al. (2009) sequenced 89 ancient horse samples for eight

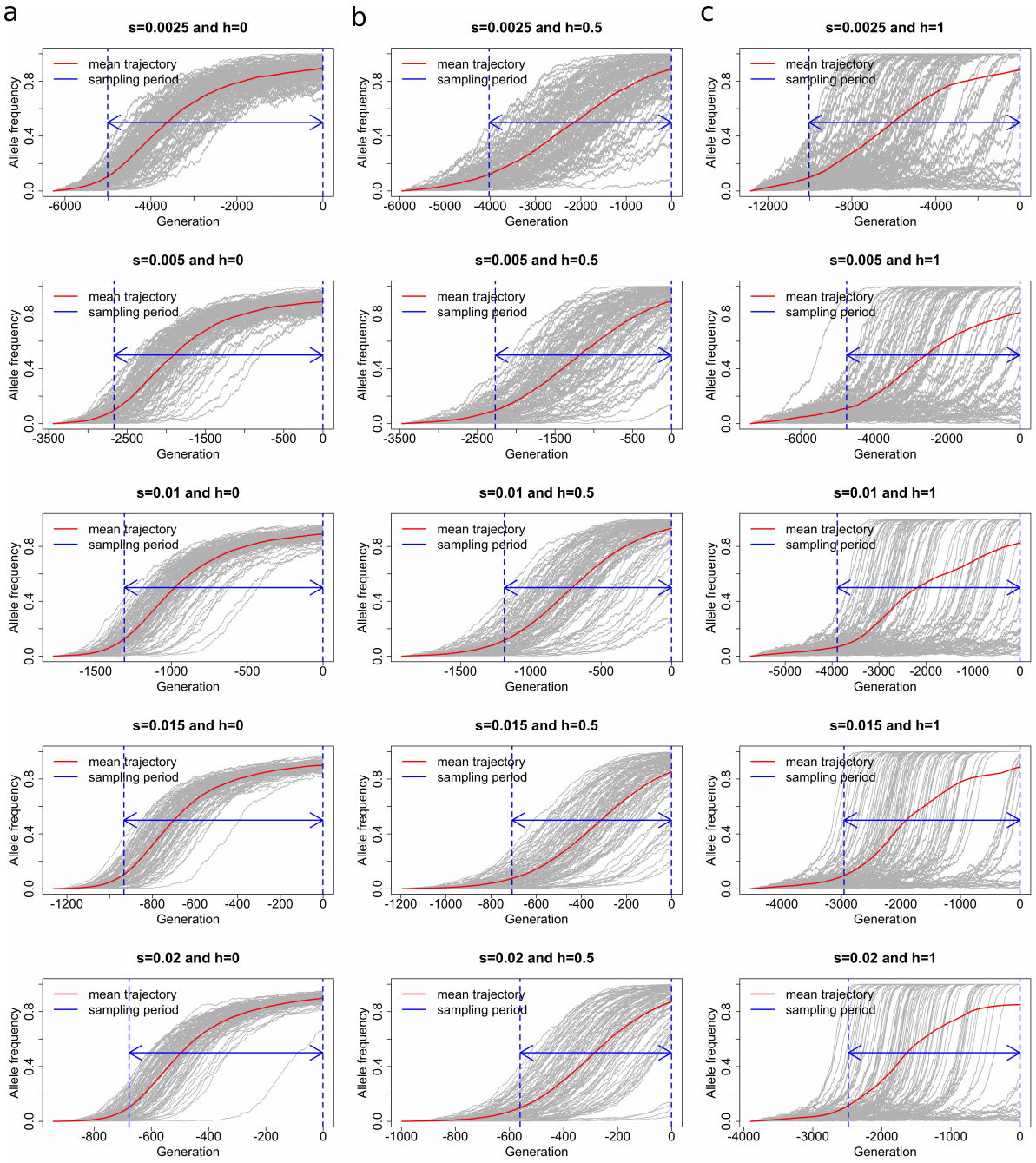

Figure 5: Simulated mutant allele frequency trajectories of the underlying population for the datasets simulated across different parameter ranges. (a) Population mutant allele frequency trajectories for the case of dominant selection. (b) Population mutant allele frequency trajectories for the case of genic selection. (c) Population mutant allele frequency trajectories for the case of recessive selection.

loci encoding coat colouration, which were obtained from Siberia, Middle and Eastern Europe, China and the Iberian Peninsula, ranging from a pre- to a post-domestication period. Ludwig et al. (2009) used the method of Bollback et al. (2008) and found two of these loci, *ASIP* and *MC1R*, which showed strong fluctuations in the mutant allele frequencies of the sample, to be

likely selectively advantageous. Malaspinas et al. (2012), Steinrücken et al. (2014) and Schraiber et al. (2016) then re-analysed the same aDNA data for *ASIP* and *MC1R* with their approaches incorporating more complex demographic scenarios.

We use the aDNA data for *ASIP* and *MC1R* from Wutke et al. (2016), which contains 107 ancient horse samples for eight loci encoding coat colouration sequenced by them along with 94 sequenced ancient horse samples from the previous studies of Ludwig et al. (2009) and Pruvost et al. (2011). Following Der Sarkissian et al. (2015), we take the average length of a generation of the horse to be 8 years. See Figure 6 for the changes in the mutant allele frequencies of the sample at the *ASIP* and *MC1R* loci. As demonstrated in Malaspinas et al. (2012), the dataset on hand is insufficient for an accurate estimate of the population size. So, instead of trying to estimate the population size from the data itself, we apply our method to estimate the selection coefficient and the allele age for the mutant allele at the *ASIP* and *MC1R* loci with a population size of $N = 16000$, the most recently reported estimate. This estimate for the population size was given in Der Sarkissian et al. (2015) and adopted in Schraiber et al. (2016). In addition, the dominance parameter is taken to be $h = 1$ in our analysis since the mutant alleles at the *ASIP* and *MC1R* loci are both recessive (Rieder et al., 2001).

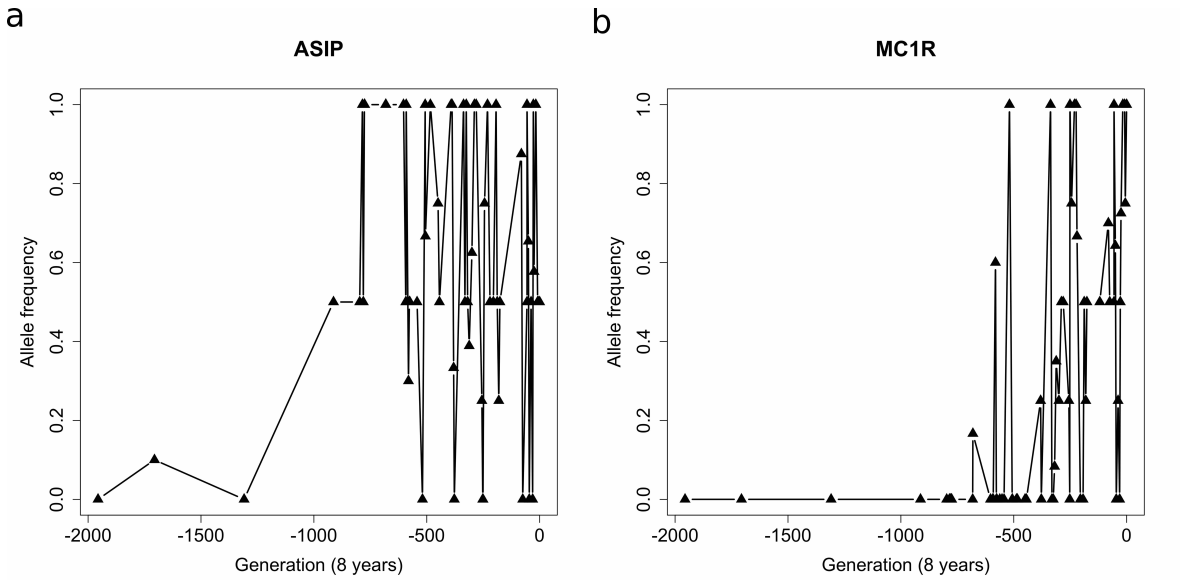

Figure 6: Changes in the mutant allele frequencies of the sample over time at the *ASIP* and *MC1R* loci. (a) Sample mutant allele frequency trajectory for *ASIP*. (b) Sample mutant allele frequency trajectory for *MC1R*.

The resulting likelihood surface for *ASIP* is shown in Figure 7a. We find that the likelihood surface attains its maximum at 0.0018 for the selection coefficient and -58706 for the allele

age, *i.e.*, 58706 years before present (BP). Our results suggest that the mutant allele at the *ASIP* locus was favoured by natural selection and was created in the late Pleistocene, which lasted from approximately 126000 to 11700 years BP (Cohen et al., 2013). In order to establish the significance of our findings, we compute the corresponding 95% confidence intervals (CI's) with a bootstrap procedure similar to that adopted in Steinrücken et al. (2014). The marginal histograms of the estimates of the selection coefficient and the allele age for the 1000 bootstrap replicates are illustrated in Figure 7b, where the 95% CI's are constructed as the interval between 2.5th and 97.5th percentiles. The 95% CI is $[0.0011, 0.0026]$ for the selection coefficient and $[-73329, -48517]$ for the allele age, respectively, indicating further strong evidence that the mutant allele at the *ASIP* locus was positively selected and was created in the middle of the late Pleistocene. We run an empirical study to demonstrate that our method works well for the aDNA data for *ASIP* with the estimated selection coefficient and allele age (see Supplemental Material, Figure S2).

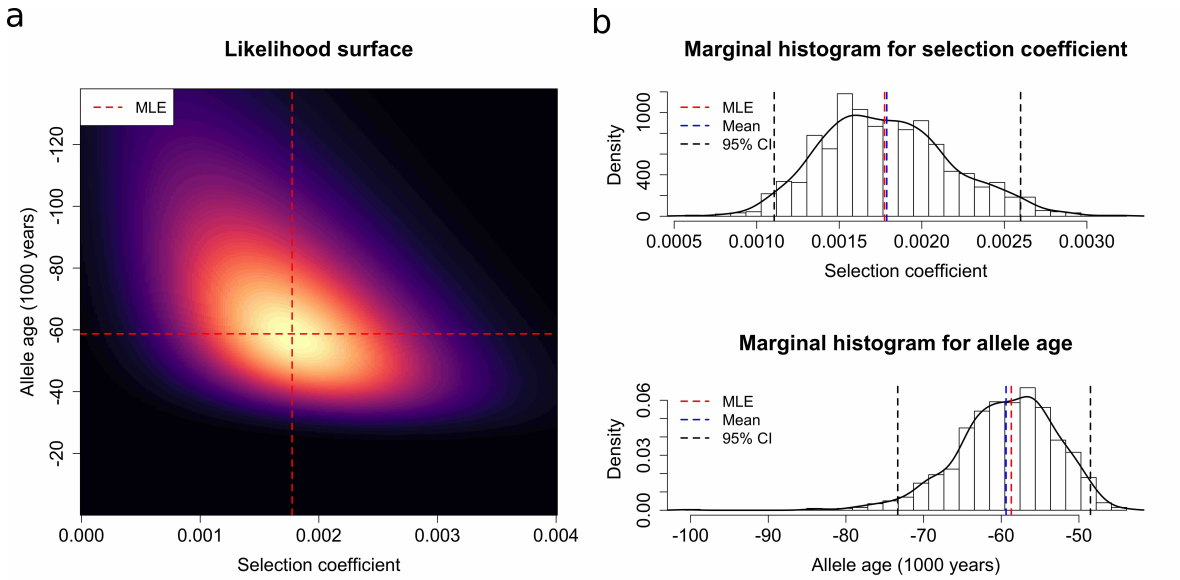

Figure 7: Analysis of the *ASIP* locus. (a) Likelihood surface. (b) Marginal histograms of the estimates for 1000 bootstrap resampled datasets of the allele frequency time series data.

The result for our analysis of *MC1R* is illustrated in Figure 8. Our estimate of the selection coefficient is 0.0129 with the 95% CI $[0.0101, 0.0162]$, which indicates that the mutant allele at the *MC1R* locus was selectively advantageous. Our estimate of the allele age is -20792 with the 95% CI $[-23378, -18770]$, which suggests that the mutant allele at the *MC1R* locus was created at the end of the late Pleistocene. An empirical study of *MC1R* with the estimated

selection coefficient and allele age can be found in Supplemental Material, Figure S3.

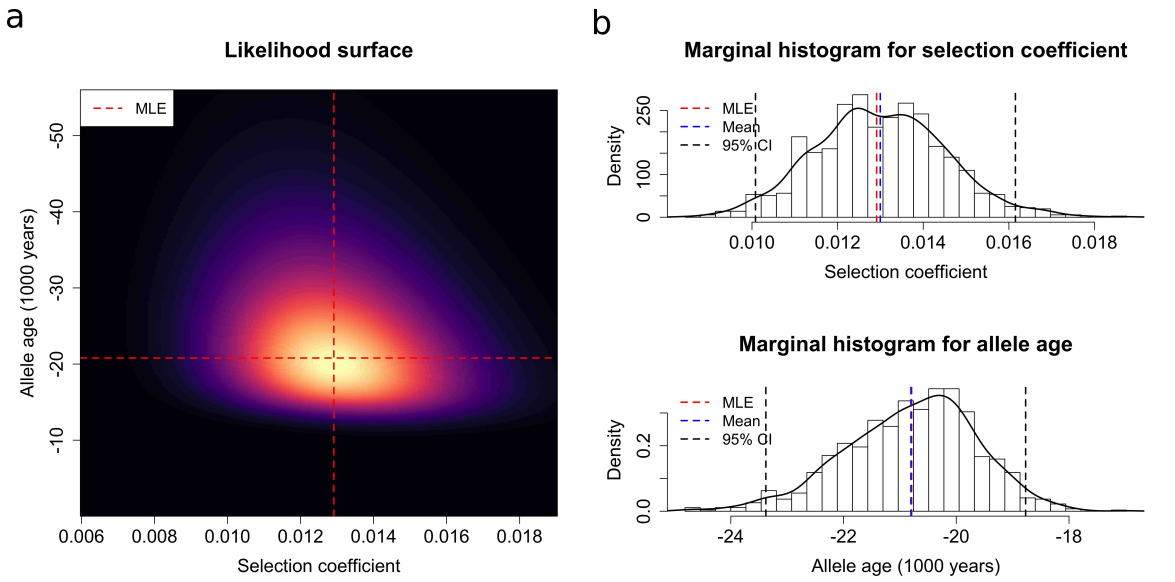

Figure 8: Analysis of the *MC1R* locus. (a) Likelihood surface. (b) Marginal histograms of the estimates for 1000 bootstrap resampled datasets of the allele frequency time series data.

In summary, our findings suggest that *ASIP* and *MC1R* mutations arose, respectively, in the middle and at the end of the late Pleistocene, and have both been favoured by natural selection. Polymorphisms at the *ASIP* and *MC1R* loci are associated with horse base coat colouration, creating a bay, black or chestnut coat (Rieder et al., 2001). *ASIP* mutations give rise to black-coloured horses, whereas *MC1R* mutations give rise to chestnut-coloured horses. It is tempting to speculate that horse base coat colour variation could be associated with adaptation to climate change. The Last Glacial Period (LGP) in the Late Pleistocene occurred between around 115000 and 11700 years BP and was followed by an interglacial period, known as the Holocene, which has been identified with the current warm period (Walker et al., 2009). The maximum extent of the LGP was the Last Glacial Maximum (LGM) and occurred between 26500 and 19000-20000 years BP (Clark et al., 2009).

Our results suggest that *ASIP* mutations, giving dark horses, were created during the LGP but before the LGM, whereas *MC1R* mutations, leading to paler horses, were created during the LGP but after the LGM. Given the findings of Finch et al. (1984) that dark-coated animals absorb heat more rapidly from solar radiation than light-coated ones, we speculate that *ASIP* mutations were favoured in horses living in cold environments during the glacial period, where dark coats would have helped them to survive, especially for the LGM. In contrast, *MC1R* mu-

tations were favoured when the climate warmed from the late Pleistocene to the early Holocene since it might be then advantageous to be light-coated.

Our findings are further confirmed with the estimates obtained with different values of the population size (*i.e.*, $N = 8000$ and $N = 32000$), which can be found in Supplemental Material, Figures S4 and S5. It is also worthwhile to note that, unlike earlier studies, where the ancient samples were grouped into 6 sampling time points (*e.g.*, Ludwig et al. (2009)) or 9 sampling time points (*e.g.*, Wutke et al. (2016)), our analysis is conducted on the raw aDNA data, *i.e.*, 201 ancient horse samples with 62 sampling time points. To investigate the effects of grouping samples on the estimates of the selection coefficient and the allele age, we perform an empirical study by random grouping the 62 sampling time points into 9 time points like Wutke et al. (2016). We apply our approach on 200 randomly grouped samples and demonstrate that the resulting estimates are roughly normally distributed around the estimates obtained from the raw samples with 62 sampling time points (see Figure 9). This shows that grouping samples can significantly alter the estimates of the selection coefficient and the allele age.

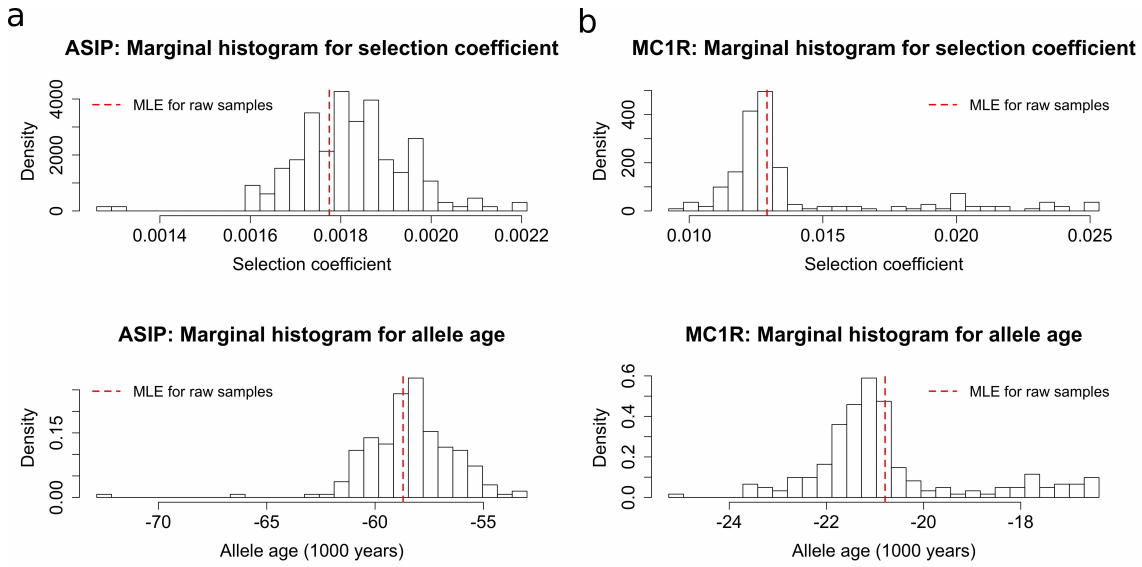

Figure 9: Effects of grouping samples on the estimates of the selection coefficient and the allele age. (a) Marginal histograms of the estimates for 200 randomly grouped datasets of the allele frequency time series data for *ASIP*. (b) Marginal histograms of the estimates for 200 randomly grouped datasets of the allele frequency time series data for *MC1R*.

3.3. Computational issues

We apply the Crank-Nicolson scheme proposed by Crank & Nicolson (1947) to numerically solve the KBE's in Eqs. (1) and (7). To achieve higher computational accuracy of the numerical

solution, it is desirable to increase the number of grid points in the Crank-Nicolson scheme, which, however, becomes burdensome computationally. We adopt a grid that is logarithmically divided close to 0 and 1, and such a partition increases the number of grid points at the two ends of the state space $[0, 1]$. More specifically, we use the grid for the state space $[0, 1]$ that is uniformly divided in $[0.03, 0.97]$ with 30 subintervals and logarithmically divided in $[0, 0.03]$ and $[0.97, 1]$ with 120 subintervals each. To find the maximum of the likelihood surface for the selection coefficient and the allele age, we start the section search on the selection coefficient with 7 equally spaced selection coefficients within a fixed interval $[-0.02, 0.04]$.

In our analysis of *ASIP* and *MC1R*, a single run for a fixed value of the selection coefficient takes about 4.4 seconds for *MC1R* on a single core of an Intel Core i7 processor at 3.5 GHz, whereas the same run takes approximately 12.4 seconds for *ASIP*, due to its much longer allele age. The entire maximum likelihood estimation procedure for *MC1R* takes 109 seconds, whereas this procedure for *ASIP* takes 324 seconds. All of our code in this work is written in R.

4. Discussion

With recent advances in next-generation sequencing technologies, aDNA data from multiple time points has become increasingly common. This additional temporal component allows for direct access to the most important dimension in evolution, time, which has the promise of providing improved power for population genetic parameter estimation, especially the selection coefficient. To take advantage of aDNA data, we developed a novel likelihood-based method for co-estimating the selection coefficient and the allele age from allele frequency time series data in this work.

To our knowledge, Malaspinas et al. (2012) and Schraiber et al. (2016) are the only existing methods that can jointly infer natural selection and allele age from time series data of allele frequencies. The key innovation of our approach is that we take the hidden Markov process in the HMM to be the Wright-Fisher diffusion conditioned to survive until at least the time of the most recent sample and calculate the likelihood by numerically solving the KBE resulting from the conditioned Wright-Fisher diffusion backwards in time. For every fixed value of the selection coefficient, we need to numerically solve the KBE only once to obtain the likelihood for all possible values of the allele age. Our backwards-in-time likelihood recursion reduces the two-dimensional numerical search for the maximum of the likelihood surface for the selection

coefficient and the allele age, *e.g.*, as used in Malaspinas et al. (2012), to a one-dimensional numerical search over the selection coefficient. Therefore, our method can achieve the estimates of the selection coefficient and the allele age relatively quickly, *e.g.*, about 324 and 109 seconds for the aDNA data for *ASIP* and *MC1R*, respectively. Furthermore, the conditioned Wright-Fisher diffusion incorporated into our HMM framework allows us to avoid the somewhat arbitrary initial condition that the allele is created by mutation at a certain small frequency, *e.g.*, as used in Schraiber et al. (2016).

We demonstrated through simulated data that our estimates for both the selection coefficient and the allele age were unbiased while the likelihood surfaces were smooth with a shape that coincided with our intuitive understanding of our method. Even though the samples are sparsely distributed in time with small uneven sizes, our approach still performs well, which is important for aDNA. We also investigated two different sampling schemes and found that it was better to increase the number of sampling times rather than the number of sampled chromosomes at each given time point, which was consistent with Malaspinas et al. (2012). This implies that it is better to infer natural selection and allele age with raw samples (*i.e.*, more sampling times but fewer sampled chromosomes each) rather than the grouped samples (*i.e.*, fewer sampling times but more sampled chromosomes each) analysed in earlier studies like Ludwig et al. (2009). Furthermore, we demonstrated that different grouping strategies for the empirical data could significantly alter the estimates of the selection coefficient and the allele age.

We re-analysed the genes that were well-studied in previous studies of Ludwig et al. (2009), Malaspinas et al. (2012), Steinrücken et al. (2014) and Schraiber et al. (2016), *ASIP* and *MC1R*, with the expanded aDNA data from Wutke et al. (2016). We found strong evidence for weak positive selection acting on the *ASIP* locus and strong positive selection acting on the *MC1R* locus, which was similar to those reported by Ludwig et al. (2009), Steinrücken et al. (2014) and Schraiber et al. (2016). In contrast, Malaspinas et al. (2012) did not have sufficient resolution to distinguish positive selection from negative selection for the *ASIP* locus with their method from the same aDNA data as analysed in Ludwig et al. (2009). In addition, Malaspinas et al. (2012) found that the allele age of *ASIP* mutations ranged from 20000 to 13100 years BP, which was significantly different from our estimate. Such a discrepancy could be caused by different demographic histories, insufficient amounts of data, or improperly grouped samples.

Steinrücken et al. (2014) and Schraiber et al. (2016) also inferred the model of natural selection, which we take to be fixed in our analyses with mutant allele homozygotes at both *ASIP* and *MC1R* loci being recessive. However, owing to the computational advantages of our method, it is ready to be applied to infer the model of general diploid natural selection, *i.e.*, running a two-dimensional numerical search over the selection coefficient and the dominance parameter to get the maximum of the likelihood. Our approach can be extended for fluctuating population sizes and selection coefficients by combining the KBE's for different population genetic parameters at the change time points suitably.

Recent findings of Sandoval-Castellanos et al. (2017) suggested that at the end of the Ice Age forests started to advance in Europe, and horses became increasingly dark in coat colour as they left the pains and adapted to live in forests, where dark-coloured coats would have helped them to hide. Our results provide evidence for an alternative possibility that horse base coat colour variation could be associated with adaptation to the transition from a glacial period to an interglacial period. *ASIP* mutations arose in the middle of the LGP but before the LGM. Its positive selection could be driven by the fact that dark-coated horses absorb more heat from solar radiation than light-coated ones so that dark-coloured coats would have helped them to survive in cold environments, especially for the LGM. *MC1R* mutations arose at the end of the LGP but after the LGM. Its positive selection could be caused by climate warming from the late Pleistocene to the early Holocene as it might be advantageous to be light-coated in warm environments for horses.

One key limitation of our method is that we infer natural selection and allele age without accounting for the interactions between the loci. Such interactions can be epistatic interaction, genetic linkage or others, *e.g.*, the *ASIP* and *MC1R* genes are found with epistatic interactions (Rieder et al., 2001), which may affect population genetic inference. With the increasing availability of aDNA data across multiple loci, performing accurate estimation on relevant population genetic quantities of interest while accounting for interactions among loci becomes more and more important. To extend our method to allele frequency time series data across multiple loci subject to epistatic interaction or genetic linkage, a significant challenge would be to find an alternative to calculate transition probabilities by numerically solving the KBE in our likelihood computation since it is computationally challenging and prohibitively expensive to numerically

solve a high-dimensional partial differential equation.

Acknowledgements

This work was funded in part by the Engineering and Physical Sciences Research Council (EPSRC) Grant EP/I028498/1 to F.Y.

References

Bollback, J. P., York, T. L., & Nielsen, R. (2008). Estimation of $2N_e s$ from temporal allele frequency data. *Genetics*, *179*, 497–502.

Clark, P. U., Dyke, A. S., Shakun, J. D., Carlson, A. E., Clark, J., Wohlfarth, B., Mitrovica, J. X., Hostetler, S. W., & McCabe, A. M. (2009). The last glacial maximum. *Science*, *325*, 710–714.

Cohen, K. M., Finney, S. C., Gibbard, P. L., & Fan, J.-X. (2013). The ICS international chronostratigraphic chart. *Episodes*, *36*, 199–204.

Crank, J., & Nicolson, P. (1947). A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Mathematical Proceedings of the Cambridge Philosophical Society*, *43*, 50–67.

Der Sarkissian, C., Ermini, L., Schubert, M., Yang, M. A., Librado, P., Fumagalli, M., Jónsson, H., Bar-Gal, G. K., Albrechtsen, A., Vieira, F. G., Petersen, B., Ginolhac, A., Seguin-Orlando, A., Magnussen, K., Fages, A., Gamba, C., Lorente-Galdos, B., Polani, S., Steiner, C., Neuditschko, M., Jagannathan, V., Feh, C., Greenblatt, C. L., Ludwig, A., Abramson, N. I., Zimmermann, W., Schafberg, R., Tikhonov, A., Sicheritz-Ponten, T., Willerslev, E., Marques-Bonet, T., Ryder, O. A., McCue, M., Rieder, S., Leeb, T., Slatkin, M., & Orlando, L. (2015). Evolutionary genomics and conservation of the endangered Przewalski's horse. *Current Biology*, *25*, 2577–2583.

Durrett, R. (2008). *Probability Models for DNA Sequence Evolution*. New York: Springer-Verlag.

Finch, V. A., Bennett, I. L., & Holmes, C. R. (1984). Coat colour in cattle: effect on thermal balance, behaviour and growth, and relationship with coat type. *The Journal of Agricultural Science*, *102*, 141–147.

Fisher, R. A. (1922). On the dominance ratio. *Proceedings of the Royal Society of Edinburgh*, *42*, 321–341.

Flink, L. G., Allen, R., Barnett, R., Malmström, H., Peters, J., Eriksson, J., Andersson, L., Dobney, K., & Larson, G. (2014). Establishing the validity of domestication genes using DNA from ancient chickens. *Proceedings of the National Academy of Sciences*, *111*, 6184–6189.

Karlin, S., & Taylor, H. E. (1981). *A Second Course in Stochastic Processes*. New York: Academic Press.

Leonardi, M., Librado, P., Der Sarkissian, C., Schubert, M., Alfarhan, A. H., Alquraishi, S. A., Al-Rasheid, K. A. S., Gamba, C., Willerslev, E., & Orlando, L. (2017). Evolutionary patterns and processes: lessons from ancient DNA. *Systematic Biology*, *66*, e1–e29.

Loog, L., Thomas, M. G., Barnett, R., Allen, R., Sykes, N., Paxinos, P. D., Lebrasseur, O., Dobney, K., Peters, J., Manica, A., Larson, G., & Eriksson, A. (2017). Inferring allele frequency trajectories from ancient DNA indicates that selection on a chicken gene coincided with changes in medieval husbandry practices. *Molecular Biology and Evolution*, *34*, 1981–1990.

Ludwig, A., Pruvost, M., Reissmann, M., Benecke, N., Brockmann, G. A., Castaños, P., Cieslak, M., Lippold, S., Llorente, L., Malaspinas, A.-S., Slatkin, M., & Hofreiter, M. (2009). Coat color variation at the beginning of horse domestication. *Science*, *324*, 485–485.

Malaspinas, A.-S. (2016). Methods to characterize selective sweeps using time serial samples: an ancient DNA perspective. *Molecular Ecology*, *25*, 24–41.

Malaspinas, A.-S., Malaspinas, O., Evans, S. N., & Slatkin, M. (2012). Estimating allele age and selection coefficient from time-serial data. *Genetics*, *192*, 599–607.

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., Sirak, K., Gamba, C., Jones, E. R., Llamas,

B., Dryomov, S., Pickrell, J., Arsuaga, J. L., de Castro, J. B., Carbonell, E., Gerritsen, F., Khokhlov, A., Kuznetsov, P., Lozano, M., Meller, H., Mochalov, O., Moiseyev, V., Guerra, M. A. R., Roodenberg, J., Vergés, J. M., Krause, J., Cooper, A., Alt, K. W., Brown, D., Anthony, D., Lalueza-Fox, C., Haak, W., Pinhasi, R., & Reich, D. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, *528*, 499–503.

Pruvost, M., Bellone, R., Benecke, N., Sandoval-Castellanos, E., Cieslak, M., Kuznetsova, T., Morales-Muñiz, A., O'Connor, T., Reissmann, M., Hofreiter, M., & Ludwig, A. (2011). Genotypes of predomestic horses match phenotypes painted in Paleolithic works of cave art. *Proceedings of the National Academy of Sciences*, *108*, 18626–18630.

Rieder, S., Taourit, S., Mariat, D., Langlois, B., & Guérin, G. (2001). Mutations in the agouti (ASIP), the extension (MC1R), and the brown (TYRP1) loci and their association to coat color phenotypes in horses (Equus caballus). *Mammalian Genome*, *12*, 450–455.

Sandoval-Castellanos, E., Wutke, S., Gonzalez-Salazar, C., & Ludwig, A. (2017). Coat colour adaptation of post-glacial horses to increasing forest vegetation. *Nature Ecology & Evolution*, *1*, 1816.

Schraiber, J. G., Evans, S. N., & Slatkin, M. (2016). Bayesian inference of natural selection from allele frequency time series. *Genetics*, *203*, 493–511.

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

Song, Y. S., & Steinrücken, M. (2012). A simple method for finding explicit analytic transition densities of diffusion processes with general diploid selection. *Genetics*, *190*, 1117–1129.

Steinrücken, M., Bhaskar, A., & Song, Y. S. (2014). A novel spectral method for inferring general diploid selection from time series genetic data. *The Annals of Applied Statistics*, *8*, 2203–2222.

Sverrisdóttir, O. O., Timpson, A., Toombs, J., Lecoeur, C., Froguel, P., Carretero, J. M., Arsuaga Ferreras, J. L., Götherström, A., & Thomas, M. G. (2014). Direct estimates of natural selection in Iberia indicate calcium absorption was not the only driver of lactase persistence in Europe. *Molecular Biology and Evolution*, *31*, 975–983.

Valleriani, A. (2016). A conditional likelihood is required to estimate the selection coefficient in ancient DNA. *Scientific Reports*, *6*, 31561.

Walker, M., Johnsen, S., Rasmussen, S. O., Popp, T., Steffensen, J.-P., Gibbard, P., Hoek, W., Lowe, J., Andrews, J., Björck, S., Cwynar, L. C., Hughen, K., Kershaw, P., Kromer, B., Litt, T., Lowe, D. J., Nakagawa, T., Newnham, R., & Schwander, J. (2009). Formal definition and dating of the GSSP (Global Stratotype Section and Point) for the base of the Holocene using the Greenland NGRIP ice core, and selected auxiliary records. *Journal of Quaternary Science*, *24*, 3–17.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, *16*, 97–159.

Wutke, S., Benecke, N., Sandoval-Castellanos, E., Döhle, H.-J., Friederich, S., Gonzalez, J., Hallsson, J. H., Hofreiter, M., Lõugas, L., Magnell, O., Morales-Muniz, A., Orlando, L., Pálsdóttir, A. H., Reissmann, M., Ruttkay, M., Trinks, A., & Ludwig, A. (2016). Spotted phenotypes in horses lost attractiveness in the Middle Ages. *Scientific Reports*, *6*, 38548.

Appendix A. Boundary condition for the conditioned Wright-Fisher diffusion

One problem that arises when numerically solving the KBE in Eq. (7) for the conditioned Wright-Fisher diffusion X^* is that $P_\tau(t,0) = 0$, which means that $\log P_\tau(t,0)$ needed for the evaluation of $a^*(t,x)$ is $-\infty$. Therefore, we can expect $\frac{\partial}{\partial x}\log P_\tau(t,x)$ to blow up to ∞ as x approaches 0. However, $b(t,x)^2 = x(1-x)$, which is roughly x when x is small. Indeed, we shall see that $\lim_{x\downarrow 0} x\frac{\partial}{\partial x}\log P_\tau(t,x) = 1$ for $t \ll \tau$. By Eq. (6.18) of Karlin & Taylor (1981), the conditioned Wright-Fisher diffusion X^* has an entrance boundary at 0, *i.e.*, the conditioned Wright-Fisher diffusion X^* starts at 0 but reaches positive values immediately after that and will never return to 0.

Under the Wright-Fisher diffusion

$$dX(t) = \alpha X(t)(1-X(t))((1-h)-(1-2h)X(t))dt + \sqrt{X(t)}dW(t), \quad t \geq t_0, \tag{A.1}$$

the fixation probability, *i.e.*, the probability of the Wright-Fisher diffusion X reaching 1 eventually, is $\mathcal{O}(1)$ for large N if $X(t_0) = \frac{1}{2N}$, *i.e.*, this probability does not go to 0 as $N \to \infty$. If X reaches levels significantly larger than $\frac{1}{2N}$, in other words, if the mutant becomes *established* in

the population, then the probability of eventual fixation becomes near certain. Thus, the key to determining whether the diffusion process reaches fixation eventually is in the early stages of the selective sweep when X is small. In this parameter regime, we can approximate the evolution of the Wright-Fisher diffusion in Eq. (A.1) by a diffusion process Y_1 in the case of $h \neq 1$:

$$dY_1(t) = \alpha(1-h)Y_1(t)dt + \sqrt{Y_1(t)}dW(t), \quad t \geq t_0, \tag{A.2}$$

and a diffusion process Y_2 in the case of $h = 1$:

$$dY_2(t) = \alpha Y_2(t)^2 dt + \sqrt{Y_2(t)}dW(t), \quad t \geq t_0. \tag{A.3}$$

The probability that X becomes established is very close to Y_1 or Y_2 becoming established. Furthermore, once Y_1 or Y_2 becomes established, since $\alpha = \mathcal{O}(N)$, *i.e.*, the effect of the drift term is much larger than that of the diffusion term in either Eq. (A.2) or (A.3), it is almost certain to go to ∞.

The probability of a diffusion process reaching a level c before hitting 0 can be calculated using the speed and scale functions (see *e.g.*, Karlin & Taylor (1981)). Let T_x denote the first hitting time of x and $\mathbb{P}^x$ denote the probability measure of the diffusion process started at x. For the diffusion process Y_1 in Eq. (A.2), we have

$$S(x) = \int_0^x s(\xi)\, d\xi = \frac{1 - e^{-\alpha(1-h)x}}{\alpha(1-h)},$$

where

$$s(\xi) = e^{-\alpha(1-h)\xi}.$$

Then the probability of Y_1 hitting c before hitting 0 is

$$\mathbb{P}^x\left(T_c < T_0\right) = \frac{S(x) - S(0)}{S(c) - S(0)} = \frac{1 - e^{-\alpha(1-h)x}}{1 - e^{-\alpha(1-h)c}}.$$

Taking $c \to \infty$ yields

$$\mathbb{P}^x\left(T_\infty < T_0\right) = 1 - e^{-\alpha(1-h)x}.$$

As discussed earlier, for $t \ll \tau$ and x close 0, $\mathbb{P}^x\left(T_\infty < T_0\right)$ is very close to the value of $P_\tau(t, x)$.

We then have

$$\begin{aligned}\lim_{x\downarrow 0} x\frac{\partial}{\partial x}\log P_\tau(t,x) &\approx \lim_{x\downarrow 0} x\frac{\partial}{\partial x}\log\left(1-e^{-\alpha(1-h)x}\right)\\ &= \lim_{x\downarrow 0} x\frac{\alpha(1-h)e^{-\alpha(1-h)x}}{1-e^{-\alpha(1-h)x}}\\ &= 1. \end{aligned} \tag{A.4}$$

Similarly, for the diffusion process Y_2 in Eq. (A.3), we have

$$S(x) = \int_0^x s(\xi)\,d\xi = \sqrt{\frac{2\pi}{\alpha}}\mathbb{P}\left(|\mathcal{N}(0,1)| < \sqrt{\alpha}x\right),$$

where

$$s(\xi) = e^{-\frac{\alpha}{2}\xi^2}$$

and $\mathcal{N}(0,1)$ denotes a standard normal random variable. Then the probability of Y_2 hitting c before hitting 0 is

$$\mathbb{P}^x(T_c < T_0) = \frac{S(x)-S(0)}{S(c)-S(0)} = \frac{\mathbb{P}\left(|\mathcal{N}(0,1)| < \sqrt{\alpha}x\right)}{\mathbb{P}\left(|\mathcal{N}(0,1)| < \sqrt{\alpha}c\right)},$$

Taking $c \to \infty$ yields

$$\mathbb{P}^x\left(T_\infty < T_0\right) = \mathbb{P}\left(|\mathcal{N}(0,1)| < \sqrt{\alpha}x\right).$$

This implies

$$\begin{aligned}\lim_{x\downarrow 0} x\frac{\partial}{\partial x}\log P_\tau(t,x) &\approx \lim_{x\downarrow 0} x\frac{\partial}{\partial x}\log\mathbb{P}\left(|\mathcal{N}(0,1)| < \sqrt{\alpha}x\right)\\ &= \lim_{x\downarrow 0} x\frac{\partial}{\partial x}\log\left(\sqrt{\frac{2\alpha}{\pi}}x + \mathcal{O}(x^2)\right)\\ &= \lim_{x\downarrow 0} x\frac{\sqrt{\frac{2\alpha}{\pi}} + \mathcal{O}(x)}{\sqrt{\frac{2\alpha}{\pi}}x + \mathcal{O}(x^2)}\\ &= 1. \end{aligned} \tag{A.5}$$

Combining Eqs. (A.4) and (A.5), we have $\lim_{x\downarrow 0} x\frac{\partial}{\partial x}\log P_\tau(t,x) = 1$ for $t \ll \tau$.