Detecting and quantifying natural selection at two linked loci from time series data of allele frequencies

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

[a] *School of Mathematics, University of Bristol, Bristol BS8 1TW, United Kingdom*
[b] *School of Biological Sciences, University of Bristol, Bristol BS8 1TQ, United Kingdom*

Abstract

The recent rapid improvement of DNA sequencing technology has made it possible to monitor genomes in great detail over time, which provides us with an opportunity for investigating natural selection based on time serial samples of genomes while accounting for genetic recombination effect and local linkage information. Such time series data allow for more accurate inference of population genetic parameters and hypothesis testing on the recent action of natural selection. In this work, we develop a novel Bayesian statistical framework for inferring natural selection at a pair of linked loci by capitalising on the temporal aspect of DNA data with the additional flexibility of modelling the sampled chromosomes that contain unknown alleles. Our approach is built on a hidden Markov model incorporating the two-locus Wright-Fisher diffusion with selection, which allows us to explicitly model genetic recombination effect and local linkage information. The posterior probability distribution for selection coefficients is obtained with the particle marginal Metropolis-Hastings, which enables us to efficiently calculate the likelihood. We evaluate the performance of our Bayesian inference procedure through extensive simulations, showing that our estimates of selection coefficients are unbiased, and the addition of genetic recombination and local linkage brings about significant improvement in the inference of natural selection. We illustrate the utility of our approach on real data with an application to ancient DNA data associated with white spotting pattern in horses.

Keywords: Natural selection, Linked loci, Wright-Fisher diffusion, Hidden Markov model, Particle marginal Metropolis-Hastings

*Corresponding author.
Email addresses: zh0131@bristol.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

Natural selection is a fundamental evolutionary process that maintains function and drives adaptation, thereby altering patterns of diversity at the genetic level. Methods for detecting and quantifying natural selection have important applications such as identifying the genetic basis of diseases and understanding the molecular basis of adaptation. There has been a long line of theoretical and experimental research devoted to modelling and inferring natural selection, and the vast majority of earlier analyses are based on allele frequency data obtained at a single time point (see Bank et al., 2014, for a short review). With the advances in DNA sequencing technologies, an ever-increasing amount of allele frequency data sampled at multiple time points are becoming available. For example, such time series data arise from experimental evolution (*e.g.*, Burke et al., 2010; Orozco-terWengel et al., 2012; Lang et al., 2013; Wiser et al., 2013), viral/phage populations (*e.g.*, Wichman et al., 1999, 2005; Holder & Bull, 2001; Bollback & Huelsenbeck, 2007), or ancient DNA (aDNA) (*e.g.*, Hummel et al., 2005; Ludwig et al., 2009; Orlando et al., 2013; Mathieson et al., 2015). Time serial samples of segregating alleles provide much more valuable information regarding natural selection since expected changes in allele frequencies over time are closely related to the strength of natural selection acting on the population. One can therefore expect allele frequency time series data to improve our ability to estimate selection coefficients and test hypotheses regarding natural selection.

There has been a growing literature on the statistical inference of natural selection from time series data of allele frequencies over the past decade (*e.g.*, Bollback et al., 2008; Malaspinas et al., 2012; Mathieson & McVean, 2013; Steinrücken et al., 2014; Lacerda & Seoighe, 2014; Feder et al., 2014; Foll et al., 2014, 2015; Terhorst et al., 2015; Schraiber et al., 2016; Shim et al., 2016; Ferrer-Admetlla et al., 2016; He et al., 2019b), reviewed in Bank et al. (2014) and Malaspinas (2016). A common approach to analysing allele frequency time series data is built on the hidden Markov model (HMM) framework of Williamson & Slatkin (1999), where the underlying population is assumed to evolve according to the Wright-Fisher model introduced by Fisher (1922) and Wright (1931), and the observations are modelled through independent binomial sampling from the underlying population at each given time point (see Tataru et al., 2017, for a detailed review of the statistical inference in the Wright-Fisher model using allele frequency data). However, such methods can become computationally infeasible for large pop-

ulations because it requires a prohibitively large amount of computation and storage in the calculation of the likelihood. Most existing HMM-based methods are therefore built on either diffusion approximation of the Wright-Fisher model (*e.g.*, Bollback et al., 2008; Malaspinas et al., 2012; Steinrücken et al., 2014; Schraiber et al., 2016; He et al., 2019b) or moment-based approximation of the Wright-Fisher model (*e.g.*, Lacerda & Seoighe, 2014; Feder et al., 2014; Terhorst et al., 2015). Such approximations enable efficient integration over all possible allele frequency trajectories of the underlying population, thereby allowing the likelihood computation to be completed in a reasonable amount of time.

The recent advent of high-throughput sequencing technologies has made it possible to monitor genomes in great detail over time. This provides an opportunity for detecting and estimating natural selection at multiple linked loci from time series data of allele frequencies while taking the process of genetic recombination and the information of local linkage into account. Properly accounting for genetic recombination and local linkage can be expected to provide more precise estimates for the selection coefficient and more accurate hypothesis testing on the recent action of natural selection since genetic recombination may either reinforce or oppose changes in allele frequencies caused by natural selection according to the levels of linkage disequilibrium (He et al., 2019a). However, with the exception of Terhorst et al. (2015), all existing Wright-Fisher model based methods for inferring natural selection from allele frequency time series data are limited to either a single locus (*e.g.*, Bollback et al., 2008; Malaspinas et al., 2012; Steinrücken et al., 2014; Schraiber et al., 2016; He et al., 2019b) or multiple independent loci (*e.g.*, Foll et al., 2014, 2015; Shim et al., 2016; Ferrer-Admetlla et al., 2016). These methods ignore genetic recombination effect and local linkage information. The exception amongst these methods, Terhorst et al. (2015), extended a moment-based approximation of the Wright-Fisher model introduced by Feder et al. (2014) to multiple linked loci, where the allele frequency transition between two given time points is modelled deterministically, with added Gaussian noise. The method of Terhorst et al. (2015) is tailored toward pooled sequencing (Pool-Seq) data from evolve-and-resequence (E&R) experiments, typically on groups of up to three linked loci.

In the present work, we propose a novel HMM-based method for Bayesian inference of natural selection at a pair of linked loci from time series data of allele frequencies while accounting for the process of genetic recombination, thereby incorporating information on local linkage. The

key innovation of our method is that the Wright-Fisher diffusion of the stochastic evolutionary dynamics under natural selection at a pair of linked loci is used as the hidden Markov process to characterise the changes in the haplotype frequencies of the underlying population over time, which allows us to explicitly model genetic recombination and local linkage. The diffusion approximation we use in our method enables us to avoid the restriction that the allele frequency trajectory of the underlying population remains far away from allele fixation or loss, which was imposed by the Gaussian approximation applied in Terhorst et al. (2015). Our posterior computation is carried out with the particle marginal Metropolis-Hastings (PMMH) algorithm developed by Andrieu et al. (2010), which allows for efficient likelihood calculations. In addition, our method is able to handle sampled chromosomes with unknown alleles, which is common in aDNA data due to postmortem DNA damage. Finally, our method can be readily extended to model changes in population size and selection coefficients like Schraiber et al. (2016), which benefits from our method that is built on simulating the Wright-Fisher diffusion.

We evaluate the performance of our Bayesian inference procedure through extensive simulations. We illustrate that our approach allows for efficient and accurate estimation of selection coefficients from allele frequency time series data, regardless of whether sampled chromosomes contain unknown alleles or not. We present two scenarios under which the existing single-locus approach fails to deliver precise estimates for selection coefficients whereas our approach still works well, especially when the loci are tightly linked. This shows the efficacy of our method in modelling genetic recombination and local linkage. Moreover, we employ our method to analyse the aDNA data associated with white spotting patterns in horses from Wutke et al. (2016). We find that in horses there is no evidence showing that the tobiano pattern is positively selected but strong evidence of the sabino pattern being negatively selected.

2. Materials and Methods

In this section, we begin with a short review of the Wright-Fisher diffusion for a pair of linked loci evolving subject to natural selection over time, and then describe our Bayesian procedure for the inference of natural selection at these two loci using allele frequency time series data, *e.g.*, how to set up the HMM framework and how to obtain the posterior probability distribution for the population genetic quantities of interest with Markov chain Monte Carlo (MCMC) techniques.

2.1. Wright-Fisher diffusion

Consider a diploid population of randomly mating individuals at a pair of linked loci $\mathcal{A}$ and $\mathcal{B}$ evolving under natural selection according to the two-locus Wright-Fisher model with selection (see, *e.g.*, He et al., 2017), for which we assume discrete time and nonoverlapping generations. At each locus, there are two possible allele types, labelled $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{B}_1$, $\mathcal{B}_2$, respectively, resulting in four possible haplotypes $\mathcal{A}_1\mathcal{B}_1$, $\mathcal{A}_1\mathcal{B}_2$, $\mathcal{A}_2\mathcal{B}_1$ and $\mathcal{A}_2\mathcal{B}_2$, labelled haplotypes 1, 2, 3 and 4, respectively. We attach symbols $\mathcal{A}_1$ and $\mathcal{B}_1$ to the mutant alleles, which are assumed to arise only once in the population and be selectively advantageous, and we attach symbols $\mathcal{A}_2$ and $\mathcal{B}_2$ to the ancestral alleles, which are assumed to originally exist in the population.

We incorporate viability selection into the population dynamics and assume that the viability is fixed from the time that the mutant allele arises and is only determined by the genotype at a single locus. More specifically, we assume that the relative viabilities of the sixteen possible (ordered) genotypes at the two loci are determined multiplicatively from the relative viabilities at individual loci, and the relative viabilities of the three possible genotypes at each locus, *e.g.*, genotypes $\mathcal{A}_1\mathcal{A}_1$, $\mathcal{A}_1\mathcal{A}_2$ and $\mathcal{A}_2\mathcal{A}_2$ at a given locus $\mathcal{A}$, are taken to be 1, $1 - h_{\mathcal{A}}s_{\mathcal{A}}$ and $1-s_{\mathcal{A}}$, respectively, where $s_{\mathcal{A}} \in [0, 1]$ is the selection coefficient and $h_{\mathcal{A}} \in [0, 1]$ is the dominance parameter. For example, the relative viability of the $\mathcal{A}_1\mathcal{B}_2/\mathcal{A}_2\mathcal{B}_2$ genotype is $(1-h_{\mathcal{A}}s_{\mathcal{A}})(1-s_{\mathcal{B}})$. We let $r \in [0, 0.5]$ be the recombination rate of the two loci on the same chromosome (*i.e.*, the fraction of recombinant offspring showing a crossover between the two loci). We assume that the population size is fixed to be N individuals for all time.

We consider a scaling limit of the Wright-Fisher model, where the unit of time is rescaled by $2N$. The scaled selection coefficients $\alpha_{\mathcal{A}} = 2Ns_{\mathcal{A}}$ and $\alpha_{\mathcal{B}} = 2Ns_{\mathcal{B}}$, and the scaled recombination rate $\rho = 4Nr$ are kept constant while the population size N is taken to infinity. As the population size approaches infinity, the haplotype frequency trajectories follow a standard diffusion limit of the two-locus Wright-Fisher model with selection (see, *e.g.*, He et al., 2019a), called the two-locus Wright-Fisher diffusion with selection. The Wright-Fisher diffusion has already been successfully applied in the inference of natural selection from allele frequency time series data. The partial differential equation (PDE) satisfied by the transition probability density function of the Wright-Fisher diffusion was used in *e.g.*, Bollback et al. (2008); Gutenkunst et al. (2009); Steinrücken et al. (2014); He et al. (2019b). In this work, as used in *e.g.*, Schraiber et al. (2016),

we characterise the Wright-Fisher diffusion as the solution of the stochastic differential equation (SDE) instead.

We let $X_i(t)$ be the frequency of haplotype i in the population at time t for $i = 1, 2, 3, 4$, and denote the haplotype frequencies of the four possible types in the population by $\boldsymbol{X}(t)$, which evolves in the state space (*i.e.*, a 3-simplex)

$$\Omega_{\boldsymbol{X}} = \left\{ \boldsymbol{x} \in [0,1]^4 : \sum_{i=1}^{4} x_i = 1 \right\},$$

and satisfies the SDE in the form of He et al. (2019a)

$$d\boldsymbol{X}(t) = \boldsymbol{\mu}(\boldsymbol{X}(t))dt + \boldsymbol{\sigma}(\boldsymbol{X}(t))d\boldsymbol{W}(t), \quad t \geq t_0 \tag{1}$$

with initial condition $\boldsymbol{X}(t_0) = \boldsymbol{x}_0$. In Eq. (1), the drift term $\boldsymbol{\mu}(\boldsymbol{x})$ is

$$\begin{aligned}
\mu_1(\boldsymbol{x}) &= \alpha_{\mathcal{A}} x_1(x_3 + x_4)\left[(x_1 + x_2)h_{\mathcal{A}} + (x_3 + x_4)(1 - h_{\mathcal{A}})\right] \\
&\quad + \alpha_{\mathcal{B}} x_1(x_2 + x_4)\left[(x_1 + x_3)h_{\mathcal{B}} + (x_2 + x_4)(1 - h_{\mathcal{B}})\right] - \frac{\rho}{2}(x_1x_4 - x_2x_3) \\
\mu_2(\boldsymbol{x}) &= \alpha_{\mathcal{A}} x_2(x_3 + x_4)\left[(x_1 + x_2)h_{\mathcal{A}} + (x_3 + x_4)(1 - h_{\mathcal{A}})\right] \\
&\quad - \alpha_{\mathcal{B}} x_2(x_1 + x_3)\left[(x_1 + x_3)h_{\mathcal{B}} + (x_2 + x_4)(1 - h_{\mathcal{B}})\right] + \frac{\rho}{2}(x_1x_4 - x_2x_3) \\
\mu_3(\boldsymbol{x}) &= -\alpha_{\mathcal{A}} x_3(x_1 + x_2)\left[(x_1 + x_2)h_{\mathcal{A}} + (x_3 + x_4)(1 - h_{\mathcal{A}})\right] \\
&\quad + \alpha_{\mathcal{B}} x_3(x_2 + x_4)\left[(x_1 + x_3)h_{\mathcal{B}} + (x_2 + x_4)(1 - h_{\mathcal{B}})\right] + \frac{\rho}{2}(x_1x_4 - x_2x_3) \\
\mu_4(\boldsymbol{x}) &= -\alpha_{\mathcal{A}} x_4(x_1 + x_2)\left[(x_1 + x_2)h_{\mathcal{A}} + (x_3 + x_4)(1 - h_{\mathcal{A}})\right] \\
&\quad - \alpha_{\mathcal{B}} x_4(x_1 + x_3)\left[(x_1 + x_3)h_{\mathcal{B}} + (x_2 + x_4)(1 - h_{\mathcal{B}})\right] - \frac{\rho}{2}(x_1x_4 - x_2x_3),
\end{aligned} \tag{2}$$

the diffusion term $\boldsymbol{\sigma}(\boldsymbol{x})$ is

$$\boldsymbol{\sigma}(\boldsymbol{x}) = \begin{pmatrix}
\sqrt{x_1x_2} & \sqrt{x_1x_3} & \sqrt{x_1x_4} & 0 & 0 & 0 \\
-\sqrt{x_2x_1} & 0 & 0 & \sqrt{x_2x_3} & \sqrt{x_2x_4} & 0 \\
0 & -\sqrt{x_3x_1} & 0 & -\sqrt{x_3x_2} & 0 & \sqrt{x_3x_4} \\
0 & 0 & -\sqrt{x_4x_1} & 0 & -\sqrt{x_4x_2} & -\sqrt{x_4x_3}
\end{pmatrix}, \tag{3}$$

and $\boldsymbol{W}(t)$ is a six-dimensional standard Brownian motion. The term $x_1x_4 - x_2x_3$ in Eq. (2) is a

measure of the linkage disequilibrium between loci $\mathcal{A}$ and $\mathcal{B}$, which quantifies the non-random association of the alleles at these two loci.

2.2. Bayesian inference of natural selection

We consider a pair of linked loci on the same chromosome, labelled $\mathcal{A}$ and $\mathcal{B}$, and recall that the mutant alleles at the $\mathcal{A}$ and $\mathcal{B}$ loci are both 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 k-th sampling time point, we let $\boldsymbol{u}_k = (u_k^{\mathcal{A}}, u_k^{\mathcal{B}})$ and $\boldsymbol{v}_k = (v_k^{\mathcal{A}}, v_k^{\mathcal{B}})$ denote the counts of mutant alleles and ancestral alleles at loci $\mathcal{A}$ and $\mathcal{B}$ in the sample of n_k chromosomes drawn from the underlying population, respectively. The population genetic quantities of interest here are the scaled selection coefficients $\alpha_{\mathcal{A}}$ and $\alpha_{\mathcal{B}}$, the dominance parameters $h_{\mathcal{A}}$ and $h_{\mathcal{B}}$, and the scaled recombination rate ρ, which are denoted by $\boldsymbol{\vartheta} = (\alpha_{\mathcal{A}}, h_{\mathcal{A}}, \alpha_{\mathcal{B}}, h_{\mathcal{B}}, \rho)$.

2.2.1. Hidden Markov model

In our HMM framework, the underlying population is assumed to evolve according to the Wright-Fisher diffusion in Eqs. (1)-(3), and the observations are modelled as independent samples drawn from the underlying population at each sampling time point. In order to compute the posterior probability distribution $p(\boldsymbol{\vartheta} \mid \boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K})$, we condition and integrate over all possible haplotype frequency trajectories of the underlying population at each sampling time point. More specifically, let $\boldsymbol{x}_{1:K} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K)$ denote the haplotype frequency trajectories of the underlying population at the sampling time points $\boldsymbol{t}_{1:K}$. The posterior probability distribution for the population genetic quantities of interest can then be written as

$$p(\boldsymbol{\vartheta} \mid \boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K}) = \int_{\Omega_{\boldsymbol{X}}} \cdots \int_{\Omega_{\boldsymbol{X}}} p(\boldsymbol{\vartheta}, \boldsymbol{x}_{1:K} \mid \boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K})\, d\boldsymbol{x}_{1:K}, \tag{4}$$

where

$$p(\boldsymbol{\vartheta}, \boldsymbol{x}_{1:K} \mid \boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K}) \propto p(\boldsymbol{\vartheta}) p(\boldsymbol{x}_{1:K} \mid \boldsymbol{\vartheta}) p(\boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K} \mid \boldsymbol{x}_{1:K}). \tag{5}$$

In Eq. (5), $p(\boldsymbol{\vartheta})$ is the prior probability distribution for the population genetic quantities of interest and can be taken to be a uniform prior over the parameter space if prior knowledge

is poor, $p(\boldsymbol{x}_{1:K} \mid \boldsymbol{\vartheta})$ is the probability distribution for the haplotype frequency trajectories of the underlying population at the sampling time points $\boldsymbol{t}_{1:K}$, and $p(\boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K} \mid \boldsymbol{x}_{1:K})$ is the conditional probability for the observations at the sampling time points $\boldsymbol{t}_{1:K}$ given the haplotype frequency trajectories of the underlying population.

Since the Wright-Fisher diffusion is a Markov process, the probability distribution for the haplotype frequency trajectories of the underlying population at the sampling time points $\boldsymbol{t}_{1:K}$ can be decomposed as

$$p(\boldsymbol{x}_{1:K} \mid \boldsymbol{\vartheta}) = p(\boldsymbol{x}_1 \mid \boldsymbol{\vartheta}) \prod_{k=1}^{K-1} p(\boldsymbol{x}_{k+1} \mid \boldsymbol{x}_k; \boldsymbol{\vartheta}),$$

where $p(\boldsymbol{x}_1 \mid \boldsymbol{\vartheta})$ is the prior probability distribution for the haplotype frequencies of the underlying population at the initial sampling time point and can be taken to be a uniform prior over the state space $\Omega_{\boldsymbol{X}}$ if prior knowledge is poor. This is known as the flat Dirichlet distribution. The term in the product above, $p(\boldsymbol{x}_{k+1} \mid \boldsymbol{x}_k; \boldsymbol{\vartheta})$, is the transition probability density of the Wright-Fisher diffusion between two consecutive sampling time points for $k = 1, 2, \ldots, K-1$. It can be obtained by numerically solving the Kolmogorov backward equation (or its adjoint) associated with the Wright-Fisher diffusion. However, this requires a fine enough discretisation of the state space $\Omega_{\boldsymbol{X}}$ and strongly depends on the underlying population genetic parameters. In addition, numerically solving such a PDE in three dimensions for our posterior computation is computationally challenging and prohibitively expensive. We resort to an 'exact-approximate' Monte Carlo procedure (Andrieu & Vihola, 2016) that involves simulating the Wright-Fisher SDE in Eqs. (1)-(3), as a tractable alternative.

Given the haplotype frequency trajectories of the underlying population, the observations at each sampling time point are independent of one another, which means that

$$p(\boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K} \mid \boldsymbol{x}_{1:K}) = \prod_{k=1}^{K} p(\boldsymbol{u}_k, \boldsymbol{v}_k \mid \boldsymbol{x}_k),$$

where $p(\boldsymbol{u}_k, \boldsymbol{v}_k \mid \boldsymbol{x}_k)$ is the conditional probability for the observations at the k-th sampling time point given the haplotype frequency trajectories of the underlying population for $k = 1, 2, \ldots, K$. To calculate the emission probability $p(\boldsymbol{u}_k, \boldsymbol{v}_k \mid \boldsymbol{x}_k)$, we let $\boldsymbol{z}_k = (z_{1,k}, z_{2,k}, z_{3,k}, z_{4,k})$ denote the counts of the $\mathcal{A}_1\mathcal{B}_1$, $\mathcal{A}_1\mathcal{B}_2$, $\mathcal{A}_2\mathcal{B}_1$ and $\mathcal{A}_2\mathcal{B}_2$ haplotypes in the sample at the k-th sampling time

point, which are usually unobserved. We then have

$$p(\boldsymbol{u}_k, \boldsymbol{v}_k \mid \boldsymbol{x}_k) = \sum_{\boldsymbol{z}_k \in \Omega_{\boldsymbol{Z}_k}} p(\boldsymbol{z}_k \mid \boldsymbol{x}_k) p(\boldsymbol{u}_k, \boldsymbol{v}_k \mid \boldsymbol{z}_k), \tag{6}$$

where

$$\Omega_{\boldsymbol{Z}_k} = \left\{ \boldsymbol{z}_k \in \mathbb{N}^4 : \sum_{i=1}^{4} z_{i,k} = n_k, u_k^{\mathcal{A}} \leq z_{1,k} + z_{2,k} \leq n_k - v_k^{\mathcal{A}}, u_k^{\mathcal{B}} \leq z_{1,k} + z_{3,k} \leq n_k - v_k^{\mathcal{B}} \right\}.$$

Conditional on the haplotype frequency trajectories of the underlying population at the k-th sampling time point, the haplotype counts of the sample can be modelled through multinomial sampling from the underlying population with sample size n_k. We can then formulate the first term in the summation of Eq. (6) as

$$p(\boldsymbol{z}_k \mid \boldsymbol{x}_k) = \frac{n_k!}{\prod_{i=1}^{4} z_{i,k}!} \prod_{i=1}^{4} x_{i,k}^{z_{i,k}}. \tag{7}$$

The second term in the summation of Eq. (6) can be decomposed as

$$p(\boldsymbol{u}_k, \boldsymbol{v}_k \mid \boldsymbol{z}_k) = p(u_k^{\mathcal{A}}, v_k^{\mathcal{A}} \mid \boldsymbol{z}_k) p(u_k^{\mathcal{B}}, v_k^{\mathcal{B}} \mid \boldsymbol{z}_k). \tag{8}$$

Let ϕ denote the probability that a sampled chromosome at a single locus is of unknown type, which we assume to be identical for all loci. We therefore have

$$p(u_k^{\mathcal{A}}, v_k^{\mathcal{A}} \mid \boldsymbol{z}_k) = b(u_k^{\mathcal{A}}; z_{1,k} + z_{2,k}, 1 - \phi) b(v_k^{\mathcal{A}}; z_{3,k} + z_{4,k}, 1 - \phi) \tag{9}$$

$$p(u_k^{\mathcal{B}}, v_k^{\mathcal{B}} \mid \boldsymbol{z}_k) = b(u_k^{\mathcal{B}}; z_{1,k} + z_{3,k}, 1 - \phi) b(v_k^{\mathcal{B}}; z_{2,k} + z_{4,k}, 1 - \phi), \tag{10}$$

where

$$b(x; n, p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \tag{11}$$

is the formula for the binomial distribution. The probability that the sampled chromosome at a specific locus is of unknown type is usually unavailable, but we can estimate it with

$$\hat{\phi} = 1 - \frac{\sum_{k=1}^{K} (u_k^{\mathcal{A}} + v_k^{\mathcal{A}}) + \sum_{k=1}^{K} (u_k^{\mathcal{B}} + v_k^{\mathcal{B}})}{2 \sum_{k=1}^{K} n_k}. \tag{12}$$

2.2.2. Particle marginal Metropolis-Hastings

To obtain the marginal posterior $p(\boldsymbol{\vartheta} \mid \boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K})$, we resort to MCMC techniques since the posterior probability distribution in Eq. (4) is unavailable in a closed form. A Metropolis-Hastings (MH) scheme can be devised to explore the population genetic quantities of interest with a fairly arbitrary proposal probability distribution, *e.g.*, a random walk proposal, where a sample of new candidates of the parameters $\boldsymbol{\vartheta}^\star$ is drawn from the proposal $q(\boldsymbol{\vartheta}^\star \mid \boldsymbol{\vartheta})$ and is accepted with the Metropolis-Hastings ratio

$$A = \frac{p(\boldsymbol{\vartheta}^\star)}{p(\boldsymbol{\vartheta})} \frac{p(\boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K} \mid \boldsymbol{\vartheta}^\star)}{p(\boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K} \mid \boldsymbol{\vartheta})} \frac{q(\boldsymbol{\vartheta} \mid \boldsymbol{\vartheta}^\star)}{q(\boldsymbol{\vartheta}^\star \mid \boldsymbol{\vartheta})}. \tag{13}$$

Our problem now reduces to calculating the intractable marginal likelihood $p(\boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K} \mid \boldsymbol{\vartheta})$ in Eq. (13), which can be formulated as

$$p(\boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K} \mid \boldsymbol{\vartheta}) = \int_{\Omega_{\boldsymbol{X}}} \cdots \int_{\Omega_{\boldsymbol{X}}} p(\boldsymbol{x}_{1:K} \mid \boldsymbol{\vartheta}) p(\boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K} \mid \boldsymbol{x}_{1:K})\, d\boldsymbol{x}_{1:K}$$

and achieved with a Monte Carlo (MC) estimate (Beaumont, 2003; Andrieu & Roberts, 2009). This pseudo-marginal MCMC algorithm exploits the fact that the MC estimate of the marginal likelihood $p(\boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K} \mid \boldsymbol{\vartheta})$ is unbiased (or has a constant bias independent of the parameters $\boldsymbol{\vartheta}$) and targets the marginal posterior $p(\boldsymbol{\vartheta} \mid \boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K})$.

We adopt a closely related approach developed by Andrieu et al. (2010), which obtains an unbiased sequential Monte Carlo (SMC) estimate of the marginal likelihood $p(\boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K} \mid \boldsymbol{\vartheta})$ and targets the joint posterior $p(\boldsymbol{\vartheta}, \boldsymbol{x}_{1:K} \mid \boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K})$. This method is called particle marginal Metropolis-Hastings (PMMH) and permits a joint update of the population genetic quantities of interest and the latent population haplotype frequency trajectories. The co-estimation of the haplotype frequency trajectories of the underlying population is interesting in its own right, but our interest here lies only in the population genetic parameters. We therefore employ a special case of the PMMH algorithm in this work, where we do not generate and store the haplotype frequency trajectories of the underlying population in the state of the Markov chain. Full details about the PMMH algorithm can be found in Andrieu et al. (2010). Fearnhead & Künsch (2018) provided a detailed review of MC methods for estimating parameters in the HMM based on the particle filter.

In our Bayesian inference procedure, the implementation of the PMMH algorithm requires the SMC estimate of the marginal likelihood in Eq. (13). This can be achieved with the bootstrap particle filter introduced by Gordon et al. (1993). More specifically, we first draw a sample of initial candidates of the parameters $\boldsymbol{\vartheta}$ from the prior $p(\boldsymbol{\vartheta})$, then we run a bootstrap particle filter with the proposed parameters $\boldsymbol{\vartheta}$ to obtain the SMC estimate of the marginal likelihood, denoted by $\hat{p}(\boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K} \mid \boldsymbol{\vartheta})$, which is equal to the product of average weights of all particles at the sampling time points $\boldsymbol{t}_{1:K}$. In the bootstrap particle filter, we generate particles from the Wright-Fisher diffusion by using the Euler-Maruyama method (see Appendix A). We repeat the following steps until a sufficient number of samples of the parameters $\boldsymbol{\vartheta}$ have been obtained:

Step 1: Draw a sample of new candidates of the parameters $\boldsymbol{\vartheta}^\star$ from the proposal $q(\boldsymbol{\vartheta}^\star \mid \boldsymbol{\vartheta})$.

Step 2: Run a bootstrap particle filter with the proposed parameters $\boldsymbol{\vartheta}^\star$ to obtain the SMC estimate of the marginal likelihood $\hat{p}(\boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K} \mid \boldsymbol{\vartheta}^\star)$.

Step 3: Accept the proposed parameters $\boldsymbol{\vartheta}^\star$ with the Metropolis-Hastings ratio

$$A = \frac{p(\boldsymbol{\vartheta}^\star)}{p(\boldsymbol{\vartheta})} \frac{\hat{p}(\boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K} \mid \boldsymbol{\vartheta}^\star)}{\hat{p}(\boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K} \mid \boldsymbol{\vartheta})} \frac{q(\boldsymbol{\vartheta} \mid \boldsymbol{\vartheta}^\star)}{q(\boldsymbol{\vartheta}^\star \mid \boldsymbol{\vartheta})}.$$

Once enough samples of the parameters $\boldsymbol{\vartheta}$ have been obtained, we can get the minimum mean square error (MMSE) estimates for the population genetic quantities of interest, defined by

$$\hat{\boldsymbol{\vartheta}}_{MMSE} = \mathbb{E}(\boldsymbol{\vartheta} \mid \boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K}) = \int \boldsymbol{\vartheta} p(\boldsymbol{\vartheta} \mid \boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K})\, d\boldsymbol{\vartheta}.$$

Alternatively, we can compute the posterior $p(\boldsymbol{\vartheta} \mid \boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K})$ from the samples of the parameters $\boldsymbol{\vartheta}$ using nonparametric density estimation techniques (see Izenman, 1991, for a detailed review) and achieve the maximum a posteriori probability (MAP) estimates for the population genetic quantities of interest, defined by

$$\hat{\boldsymbol{\vartheta}}_{MAP} = \arg\max_{\boldsymbol{\vartheta}} p(\boldsymbol{\vartheta} \mid \boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K}).$$

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-2L-MMSE-PMMH`.

3. Results

In this section, we show how our Bayesian inference method performs on simulated datasets with given population genetic parameter values, including a scenario where sampled chromosomes contain unknown alleles. We also give two examples to demonstrate the improvement in the accuracy of estimates of the selection coefficients achieved through explicitly modelling genetic recombination and local linkage. Finally, we apply our approach to the time series aDNA data associated with horse white spotting patterns from previous studies of Ludwig et al. (2009), Pruvost et al. (2011) and Wutke et al. (2016).

3.1. Analysis of simulated data

We run forward-in-time simulations of the two-locus Wright-Fisher model with selection and evaluate the performance of our approach on these replicate simulations by examining the bias and the root mean square error (RMSE) of our Bayesian estimates. In what follows, we take the dominance parameters to be $h_{\mathcal{A}} = 0.5$ and $h_{\mathcal{B}} = 0.5$ (*i.e.*, the heterozygous fitness is the arithmetic average of the homozygous fitness, called genic selection) and choose a population size of $N = 5000$ unless otherwise noted. In principle, however, the conclusions hold for any other values of the dominance parameters $h_{\mathcal{A}}, h_{\mathcal{B}} \in [0, 1]$ and the population size $N \in \mathbb{N}$.

For each simulated dataset, given the values of the population genetic parameters $\boldsymbol{\vartheta}$ and the initial population haplotype frequencies $\boldsymbol{x}_0$, we simulate the haplotype frequency trajectories of the underlying population according to the two-locus Wright-Fisher model with selection. After obtaining the simulated population haplotype frequency trajectories, we draw the unobserved sample haplotype counts independently at each sampling time point according to the multinomial distribution in Eq. (7) first and then we generate the observed sample mutant allele counts and ancestral allele counts with Eqs. (8)-(11).

3.1.1. Power for inferring natural selection

We vary selection coefficients with $s_{\mathcal{A}} \in \{0.003, 0.01\}$ and $s_{\mathcal{B}} \in \{0, 0.002, 0.008\}$, and recombination rate with $r \in \{0.00001, 0.01\}$. We run 100 replicates for each of the 12 possible combinations of selection coefficients and recombination rate. For each replicate, we take the initial population haplotype frequencies to be $\boldsymbol{x}_0 = (0.04, 0.08, 0.08, 0.8)$ and simulate the haplotype frequency trajectories of the underlying population according to the two-locus Wright-Fisher

model with selection (see Supplemental Material, Figures S1 and S2). We sample 50 chromosomes from the underlying population at every 50 generations throughout 500 generations.

The resulting boxplots of the empirical studies are shown in Figure 1 for the allele frequency datasets generated without missing values ($\phi = 0$ in Eqs. (9) and (10)) and Figure 2 for the allele frequency datasets generated with missing values ($\phi = 0.02$ in Eqs. (9) and (10)), respectively. In these two figures, 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. As can be seen from the boxplot results, the MMSE estimates for the selection coefficients at both loci show little bias across different parameter ranges, no matter whether sampled chromosomes contain unknown alleles or not, although one can discern a slight bias for small selection coefficients. With the increase of the selection coefficients, the MMSE estimates for the selection coefficients become more accurate. The bias and the RMSE of the resulting MMSE estimates can be found in Supplemental Material, Tables S1 and S2.

For each combination of the selection coefficients and the recombination rate, we calculate the proportion of the 95% highest posterior density (HPD) intervals that include the true values, shown in the bottom left corner of each boxplot in Figures 1 and 2. On average, 92.00% of runs result in the true values of the selection coefficients being within their 95% HPD intervals for the simulated datasets without missing values, *i.e.*, 93.33% for tightly linked loci and 90.67% for loosely linked loci, and 92.08% of runs result in the true values of the selection coefficients being within their 95% HPD intervals for the simulated datasets with missing values, *i.e.*, 93.33% for tightly linked loci and 90.83% for loosely linked loci. We can see that small recombination rates can lead to better results for both loci.

In Figure 3, we illustrate results where the simulated data is given as haplotype frequencies (instead of allele frequencies in Figures 1 and 2). Compared to the estimates from allele frequency data, the estimates from haplotype frequency data are closer to their true values with smaller variances, especially for tightly linked loci. On average, 92.50% of runs result in the true values of the selection coefficients being within their 95% HPD intervals on average, with 93.67% for tightly linked loci and 91.33% for loosely linked loci. The bias and the RMSE of the resulting MMSE estimates are summarised in Supplemental Material, Table S3. This improvement in the performance of the estimates is to be expected as all else being equal since

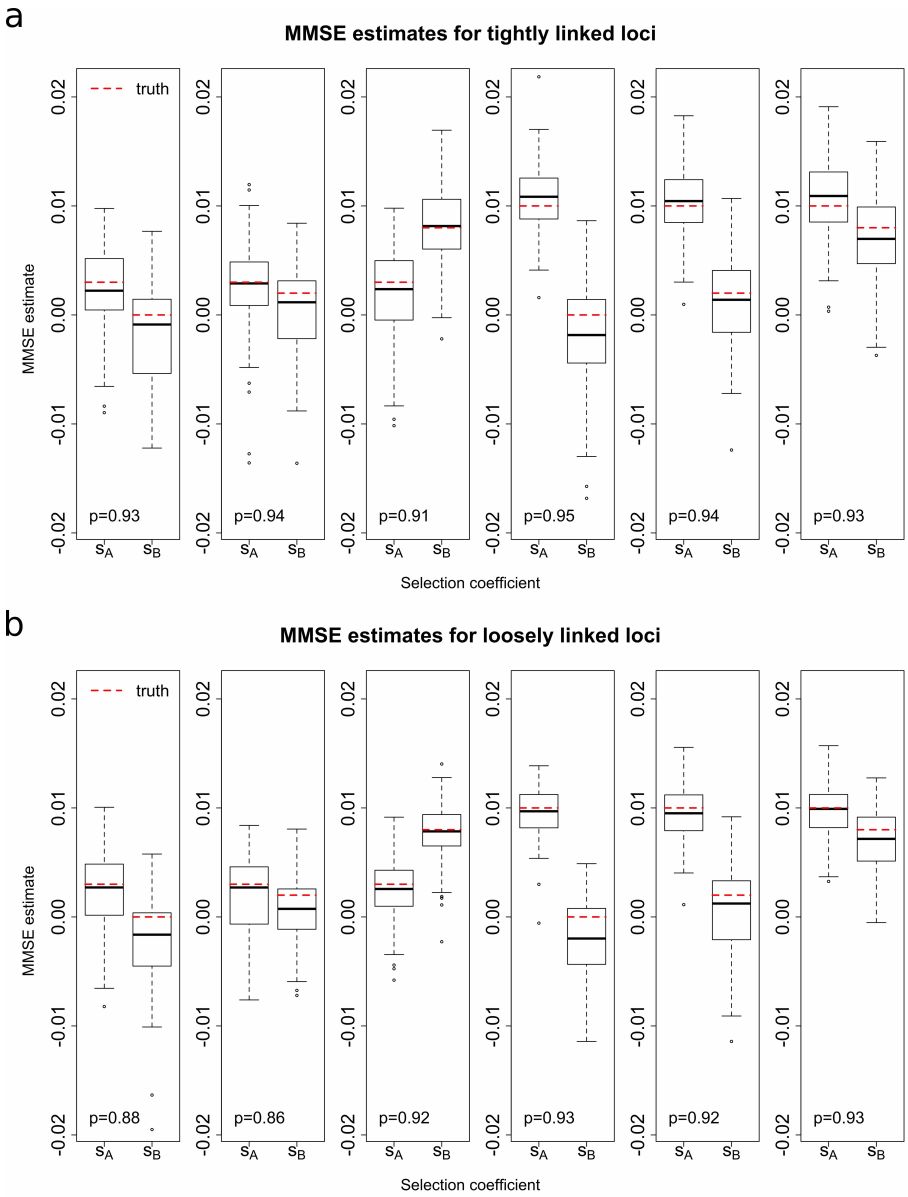

Figure 1: Empirical distributions of the MMSE estimates for 100 *allele frequency* datasets simulated *without* missing values. (a) Tightly linked loci with recombination rate $r = 0.00001$. (b) Loosely linked loci with the recombination rate $r = 0.01$. The p value in the bottom left corner indicates the proportion of runs where the true value of the selection coefficients falls within the 95% HPD.

haplotype frequency data contain more information than allele frequency data since the complex interplay between the four haplotypes in the sample can be directly observed in haplotype frequency data but only partially observed in allele frequency data.

In summary, our method can deliver accurate estimates of the selection coefficients using time series data of allele frequencies across different parameter ranges, regardless of whether sampled chromosomes contain unknown alleles or not. We have also investigated the bias and the RMSE of the resulting MAP estimates (see Supplemental Material, Tables S4-S6), which

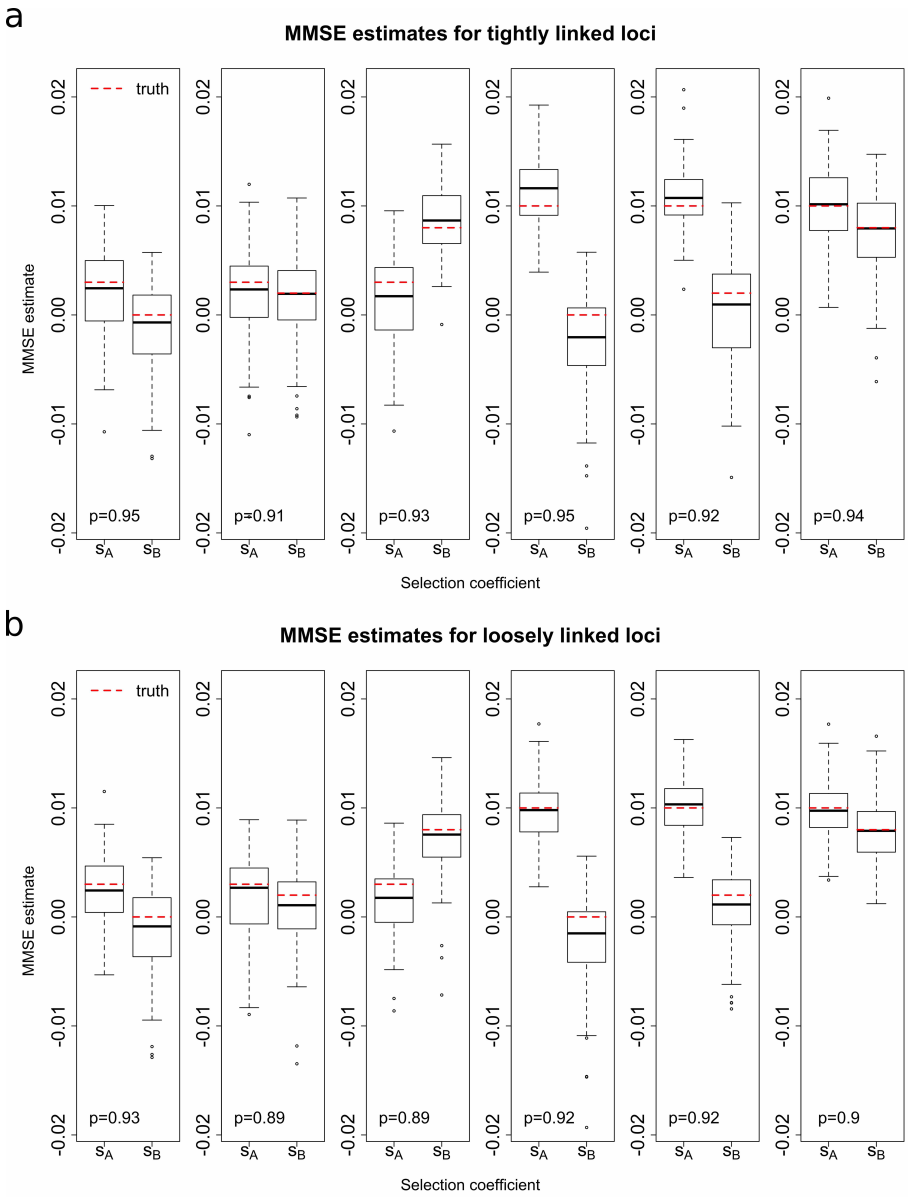

Figure 2: Empirical distributions of the MMSE estimates for 100 *allele frequency* datasets simulated *with* missing values. (a) Tightly linked loci with recombination rate $r = 0.00001$. (b) Loosely linked loci with the recombination rate $r = 0.01$. The p value in the bottom left corner indicates the proportion of runs where the true value of the selection coefficients falls within the 95% HPD.

display very similar characteristics to the MMSE estimates. The boxplots for the MAP estimates show little bias, with the upper and lower quartiles of the MAP estimates being similar to those of MMSE estimates (see Supplemental Material, Figures S3-S5).

3.1.2. Improvement from modelling genetic recombination and local linkage

In the case where a pair of loci are both suspected to be favoured by natural selection, one can still apply a single-locus approach to each locus to estimate the selection coefficient. There

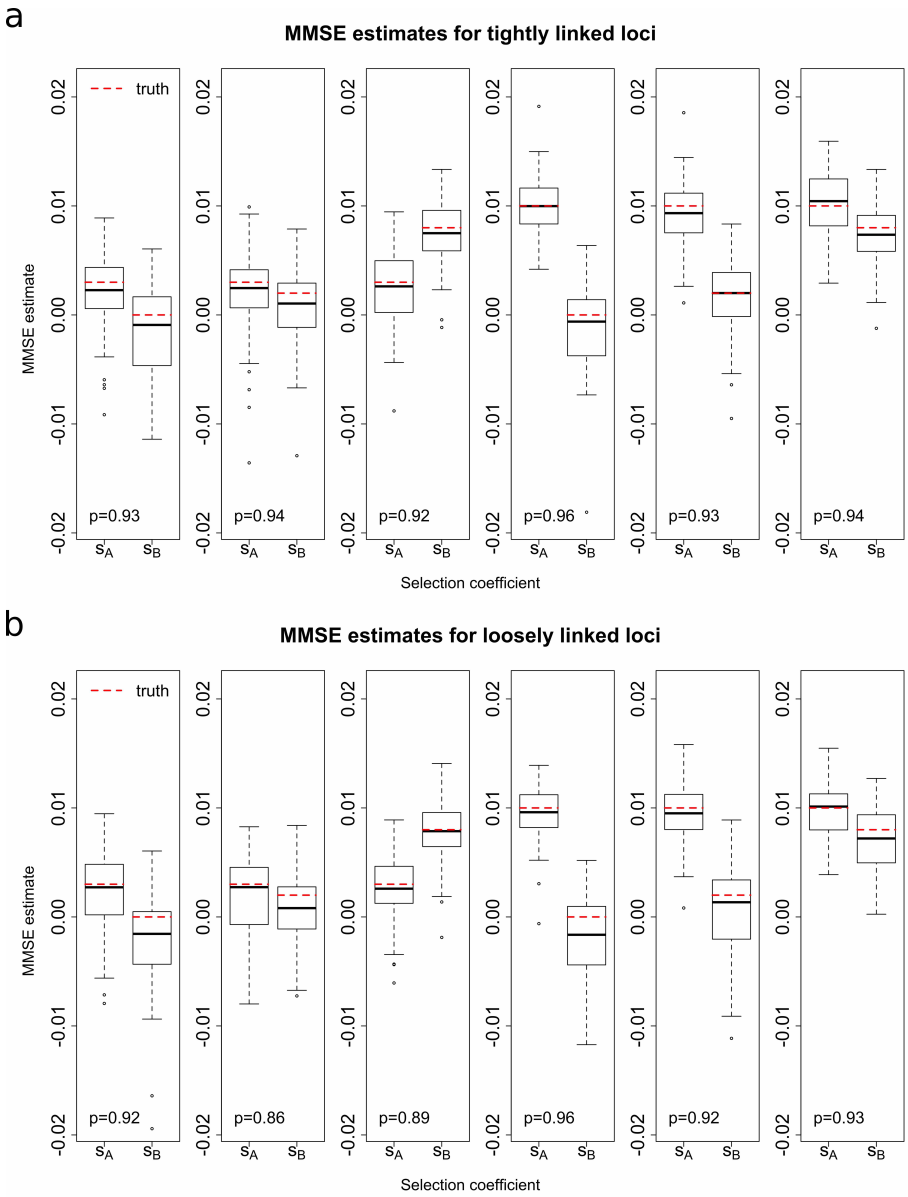

Figure 3: Empirical distributions of the MMSE estimates for 100 *haplotype frequency* datasets. (a) Tightly linked loci with recombination rate $r = 0.00001$. (b) Loosely linked loci with the recombination rate $r = 0.01$. The p value in the bottom left corner indicates the proportion of runs where the true value of the selection coefficients falls within the 95% HPD.

has been a considerable amount of work on the statistical inference of natural selection at a single locus from time series data of allele frequencies (*e.g.*, Bollback et al., 2008; Malaspinas et al., 2012; Steinrücken et al., 2014; Schraiber et al., 2016; He et al., 2019b). However, applying a single-locus approach may lead to inaccurate estimates of selection coefficients when the loci are linked (He et al., 2019a). In the case of tightly linked loci, modelling genetic recombination and local linkage becomes necessary, hence our two-locus method is far more desirable. Below we show with two examples with the recombination rate $r = 0.00001$. We simulate the haplotype

frequency trajectories of the underlying population through the two-locus Wright-Fisher model with selection and draw 200 chromosomes from the underlying population at generations 0, 100, 200, 300, 400 and 500.

In the first example, we consider a positively selected locus $\mathcal{A}$ ($s_{\mathcal{A}} = 0.01$) linked with a neutral locus $\mathcal{B}$ ($s_{\mathcal{B}} = 0$). We set the initial haplotype frequencies of the underlying population to be $\boldsymbol{x}_0 = (0.2, 0.1, 0.3, 0.4)$. The mutant allele frequency trajectories of the sample are shown in Figure 4a. The posterior probability distributions obtained with our single-locus approach, described in Supplemental Material, File S3, are shown in Figure 4b, and the posterior probability distributions estimated with our two-locus method, described in Section 2.2, are shown in Figure 4c. Bayesian estimates of the selection coefficients $s_{\mathcal{A}}$ and $s_{\mathcal{B}}$ are summarised in Table 1.

		single-locus method	two-locus method
selection coefficient $s_{\mathcal{A}}$	MAP ($\times 10^{-2}$)	1.153	1.099
	MMSE ($\times 10^{-2}$)	1.125	1.216
	95% HPD ($\times 10^{-2}$)	$[0.678, 1.506]$	$[0.684, 1.837]$
selection coefficient $s_{\mathcal{B}}$	MAP ($\times 10^{-2}$)	0.501	0.192
	MMSE ($\times 10^{-2}$)	0.507	0.125
	95% HPD ($\times 10^{-2}$)	$[0.090, 0.905]$	$[-0.905, 1.003]$

Table 1: A comparison of the Bayesian estimates obtained by using the single-locus method and the two-locus method from the simulated dataset of a positively selected locus tightly linked with a neutral locus.

One can observe that with a single-locus method, the estimate for the selection coefficient $s_{\mathcal{A}}$ is reasonably accurate, but the estimate for the selection coefficient $s_{\mathcal{B}}$ is off by a large amount. The true value for the selection coefficient $s_{\mathcal{B}}$ is 0, but the single-locus approach produces an estimate of roughly 0.005 and a 95% HPD interval that only encompasses positive values, which is strong evidence for the presence of natural selection. In comparison, the estimates for both of the selection coefficients $s_{\mathcal{A}}$ and $s_{\mathcal{B}}$ are fairly accurate with the two-locus method.

To understand the poor performance of the single-locus method in this example, we plot the mutant allele frequency trajectories of the underlying population in Figure 4d and the haplotype frequency trajectories of the underlying population in Figure 4e. The increase in the frequency of the $\mathcal{B}_1$ allele with time, despite it having a selection coefficient of 0, is caused by the $\mathcal{A}_1\mathcal{B}_1$ haplotype, which has a selection coefficient of 0.01. This compensates the decrease in the frequency of the $\mathcal{A}_2\mathcal{B}_1$ haplotype, resulting in an increasing trajectory for the $\mathcal{B}_1$ allele, albeit with a slower rate of increase than the $\mathcal{A}_1$ allele. With the two-locus approach, however, the interplay between all four haplotypes are taken into account and it produces accurate estimates

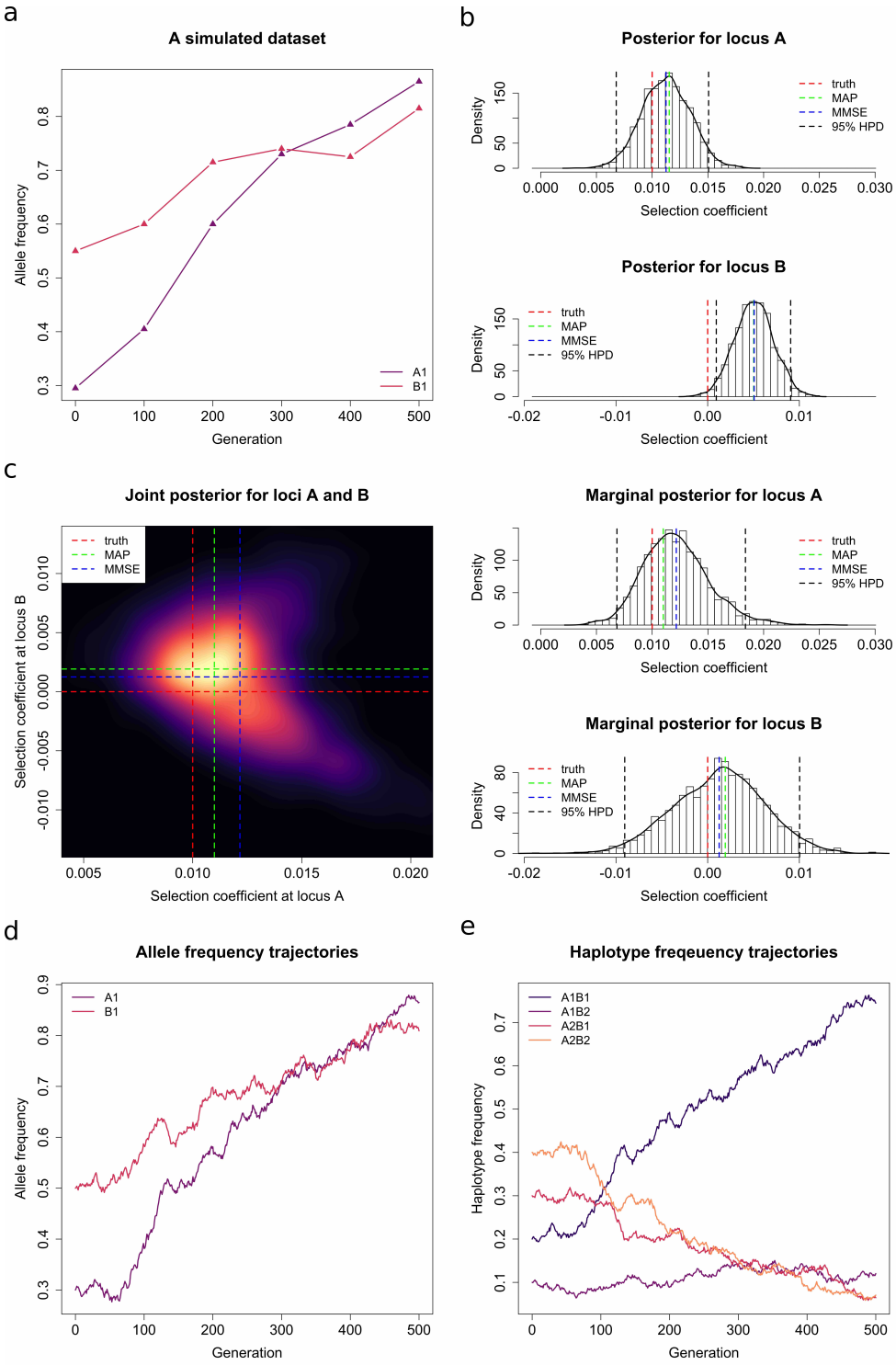

Figure 4: A comparison of the performance differences of the single-locus method and the two-locus method on the simulated dataset of a positively selected locus tightly linked with a neutral locus. (a) The simulated dataset. (b) Posteriors obtained with a single-locus method. (c) Posteriors obtained with a two-locus method. (d) Population mutant allele frequency trajectories. (e) Population haplotype frequency trajectories.

for both of the selection coefficients $s_{\mathcal{A}}$ and $s_{\mathcal{B}}$.

In the second example, we consider two positively selected and tightly linked loci $\mathcal{A}$ and $\mathcal{B}$,

where we take the selection coefficients to be $s_{\mathcal{A}} = 0.01$ and $s_{\mathcal{B}} = 0.005$, respectively, and set the initial haplotype frequencies of the underlying population to be $\boldsymbol{x}_0 = (0.05, 0.05, 0.7, 0.2)$. The results are illustrated in Figure 5 and summarised in Table 2. In this example, with the single-locus method, the estimate for the selection coefficient $s_{\mathcal{A}}$ is reasonably accurate, but the estimate for the selection coefficient $s_{\mathcal{B}}$ is off by a large amount, *i.e.*, its true value lies outside the 95% HPD interval. In fact, although the $\mathcal{B}_1$ allele is favoured by natural selection with a selection coefficient of 0.005, the resulting estimate for the selection coefficient $s_{\mathcal{B}}$ is roughly -0.0007 with a 95% HPD interval that includes the value 0, which implies no strong evidence for natural selection. In comparison, the two-locus method again produces fairly accurate estimates for both of the selection coefficients $s_{\mathcal{A}}$ and $s_{\mathcal{B}}$.

		single-locus method	two-locus method
selection coefficient $s_{\mathcal{A}}$	MAP ($\times 10^{-2}$)	0.732	0.775
	MMSE ($\times 10^{-2}$)	0.739	1.013
	95% HPD ($\times 10^{-2}$)	$[0.299, 1.154]$	$[0.340, 1.912]$
selection coefficient $s_{\mathcal{B}}$	MAP ($\times 10^{-2}$)	-0.117	0.194
	MMSE ($\times 10^{-2}$)	-0.071	0.423
	95% HPD ($\times 10^{-2}$)	$[-0.510, 0.387]$	$[-0.462, 1.531]$

Table 2: A comparison of the Bayesian estimates obtained by using the single-locus method and the two-locus method from the simulated dataset of a pair of positively selected and tightly linked loci.

As shown in Figure 5, the frequency of the $\mathcal{A}_1$ allele increases with time due to the increase in the frequencies of the $\mathcal{A}_1\mathcal{B}_1$ and $\mathcal{A}_1\mathcal{B}_2$ haplotypes, which are the two most selected haplotypes, with the selection coefficients of 0.015 and 0.01, respectively. The $\mathcal{B}_1$ allele is made up of the $\mathcal{A}_1\mathcal{B}_1$ and $\mathcal{A}_2\mathcal{B}_1$ haplotypes, with the selection coefficients of 0.01 and 0.005, respectively, which are the second and third most selected haplotypes. As a result of their initial conditions and selection coefficients, the frequency of the $\mathcal{B}_1$ allele roughly holds constant in time, since it is somewhat out-competed by the $\mathcal{A}_1$ allele. Viewing the trajectory of the $\mathcal{B}_1$ allele in isolation does not give strong evidence that it is selectively advantageous, which results in an estimate of roughly 0 in its selection coefficient through the single-locus approach. Moreover, even the 95% HPD interval for the single-locus method does not include the true selection coefficient of 0.005 for the $\mathcal{B}_1$ allele. Using the two-locus approach, we are again able to obtain accurate estimates for both of the selection coefficients $s_{\mathcal{A}}$ and $s_{\mathcal{B}}$.

Compared to existing single-locus approaches, our two-locus method accounts for the effect of genetic recombination and the information of local linkage through the two-locus Wright-

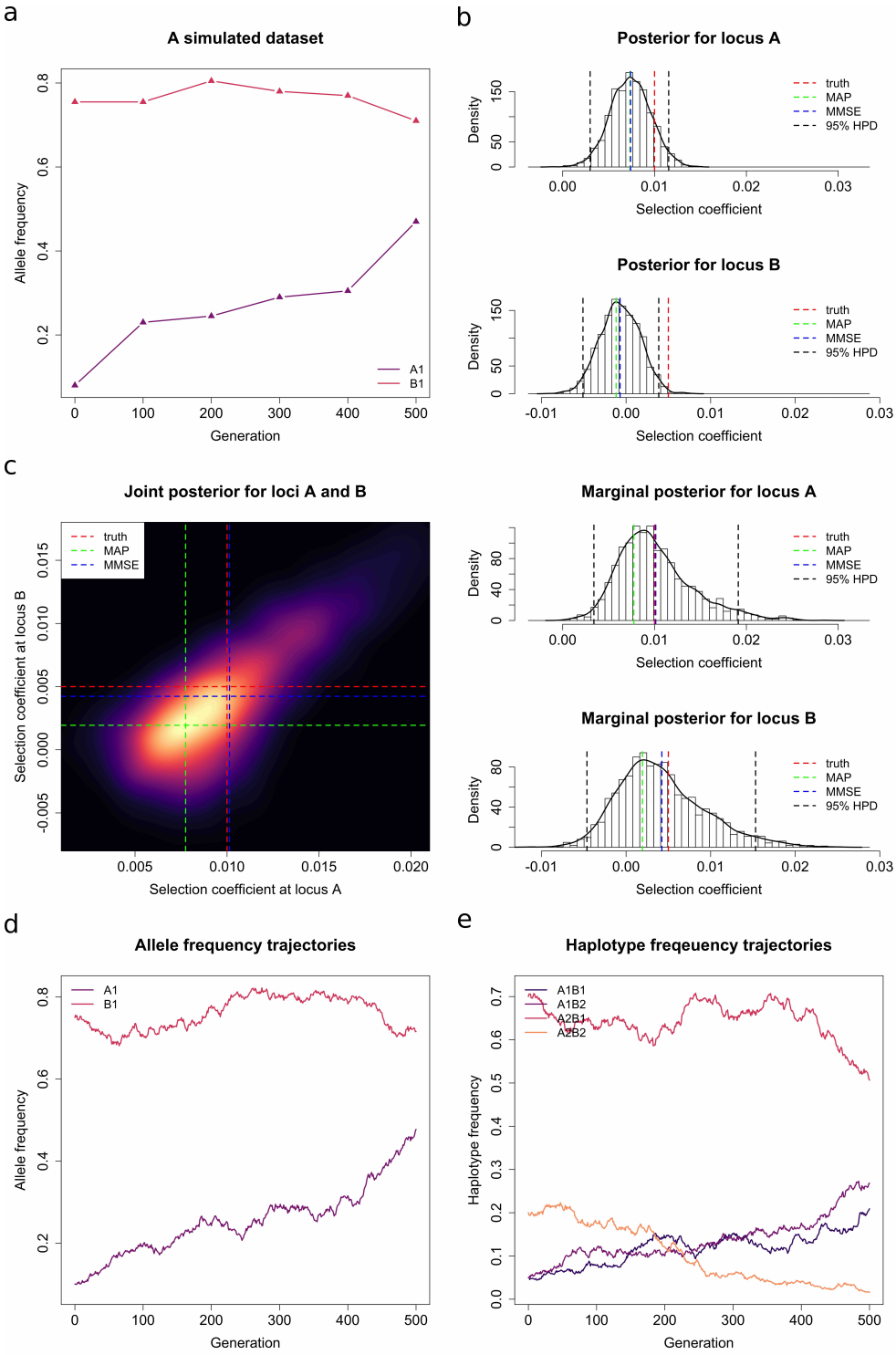

Figure 5: A comparison of the performance differences of the single-locus method and the two-locus method on the simulated dataset of a pair of positively selected and tightly linked loci. (a) The simulated dataset. (b) Posteriors obtained with a single-locus method. (c) Posteriors obtained with a two-locus method. (d) Population mutant allele frequency trajectories. (e) Population haplotype frequency trajectories.

Fisher diffusion with selection. Indeed, the dynamics of the two-locus Wright-Fisher diffusion with selection can demonstrate complex and unpredictable behaviour (see, *e.g.*, Yu & Etheridge,

2010; Cuthbertson et al., 2012), which can lead to inaccurate estimates of selection coefficients if one simply applies a single-locus approach. In contrast, applying our two-locus method can yield accurate estimates of both selection coefficients.

3.2. Analysis of real data

We apply our Bayesian inference method to real data by re-analysing the time serial samples of segregating alleles of the equine homologue of proto-oncogene c-kit (*KIT*). This data comes from previous studies of Ludwig et al. (2009), Pruvost et al. (2011) and Wutke et al. (2016), and can be found in Table 3. The *KIT* gene in horses resides on the long arm of chromosome 3 and presents two intervals associated to white spotting patterns, one in the intron 13 which codes for tobiano (*KIT13*), with the other in intron 16 which codes for sabino (*KIT16*). At the *KIT13* locus, the ancestral allele is designated *KM0*, while the mutant allele, associated with the tobiano pattern and acting as dominant (Brooks et al., 2007), is designated *KM1*. The tobiano pattern is characterised by depigmented patches of skin and associated hair that often cross the dorsal midline and cover the legs. At the *KIT16* locus, the ancestral allele is designated *sb1*, while the mutant allele associated with the sabino pattern and acting as semi-dominant (Brooks & Bailey, 2005), is designated *SB1*. The sabino pattern is characterised by irregularly bordered white patches of skin and associated hair that begin at the extremities and face and may extend up to the belly and midsection.

		KIT13	*KIT16*
sample time	sample size	*KM0*/*KM1*	*sb1*/*SB1*
17146	22	22/0	22/0
7029	14	14/0	14/0
5472	48	45/3	44/2
4442	24	24/0	24/0
3916	28	28/0	28/0
3352	56	53/3	52/4
2624	30	26/4	24/0
2330	14	11/3	12/0
1134	100	77/3	86/0

Table 3: Time serial samples of segregating alleles at the *KIT13* and *KIT16* loci. The unit of the sampling time is the year before present (BP).

We set the dominance parameters $h = 0$ for *KIT13* as the *KM1* allele is dominant, and $h = 0.5$ for *KIT16* as the *SB1* allele is semi-dominant. Following Der Sarkissian et al. (2015), we take the population size to be $N = 16000$ and the average length of a generation of horse to

be 8 years, the same as in Schraiber et al. (2016). We show all possible mutant allele frequency trajectories of the sample at the *KIT13* and *KIT16* loci in Figure 6, starting from 5472 years BP (*i.e.*, the third sampling time point). As can be seen in Table 3, there are various sampling time points when the sequencing of the aDNA material yielded a number of unknown alleles at loci *KIT13* and/or *KIT16*. Neither mutant allele was found in the first two samples dated 17146 and 7029 years BP. Indeed, both sabino and tobiano patterns are only present in domestic horses. As in Wutke et al. (2016), we assume that both mutant alleles, *KM1* and *SB1*, arose after the domestication of the horse, which is thought to have started in the Eurasian Steppes around 5500 years BP (Outram et al., 2009). We therefore discard the first two samples from our analysis. For completeness, in Supplemental Material, Figures S6 and S7, we present the results of natural selection inference when these two samples are also taken into account.

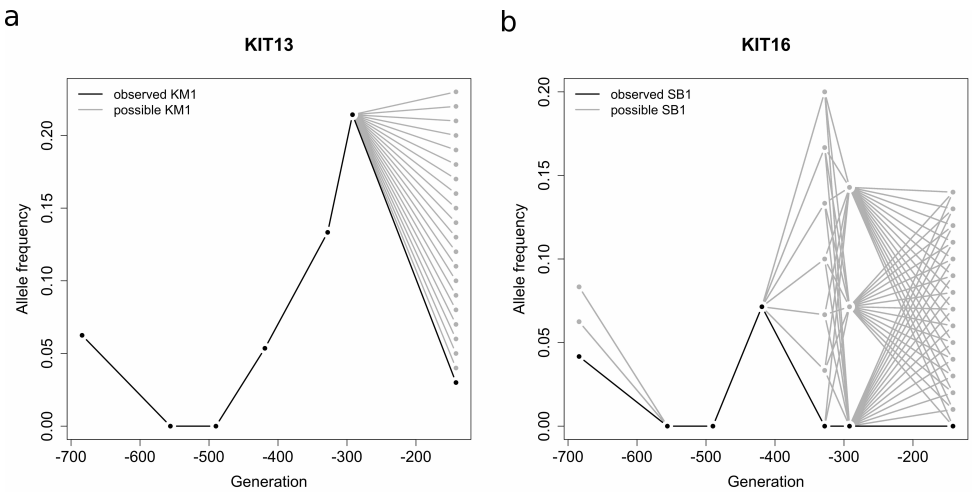

Figure 6: Potential changes in the mutant allele frequencies of the sample over time at the *KIT13* and *KIT16* loci. Ancient horse samples were taken at generations -684, -556, -490, -419, -328, -292 and -142. (a) Sample mutant allele frequency trajectories for *KIT13*. (b) Sample mutant allele frequency trajectories for *KIT16*.

As a result of the low quality of the *KIT* dataset, it becomes more difficult to intuit whether either or both mutant alleles at the *KIT13* and *KIT16* loci are selectively advantageous by simply eyeballing the mutant allele frequency trajectories of the sample. Using our two-locus Bayesian inference procedure, described in Section 2.2, we co-estimate the selection coefficients for the mutant alleles at the *KIT13* and *KIT16* loci under the case that sampled chromosomes contain variants with unknown alleles. For the recombination rate, we choose three average rates of recombination, 5×10^{-9}, 1×10^{-8} and 5×10^{-8} crossovers/bp, as suggested in Dumont & Payseur (2008), and multiply them by the genetic distance 4688 bp to obtain the recombination

rates between the *KIT13* and *KIT16* loci. The resulting posterior probability distributions are illustrated in Figure 7, and the MAP and MMSE estimates, as well as the 95% HPD intervals, are summarised in Table 4.

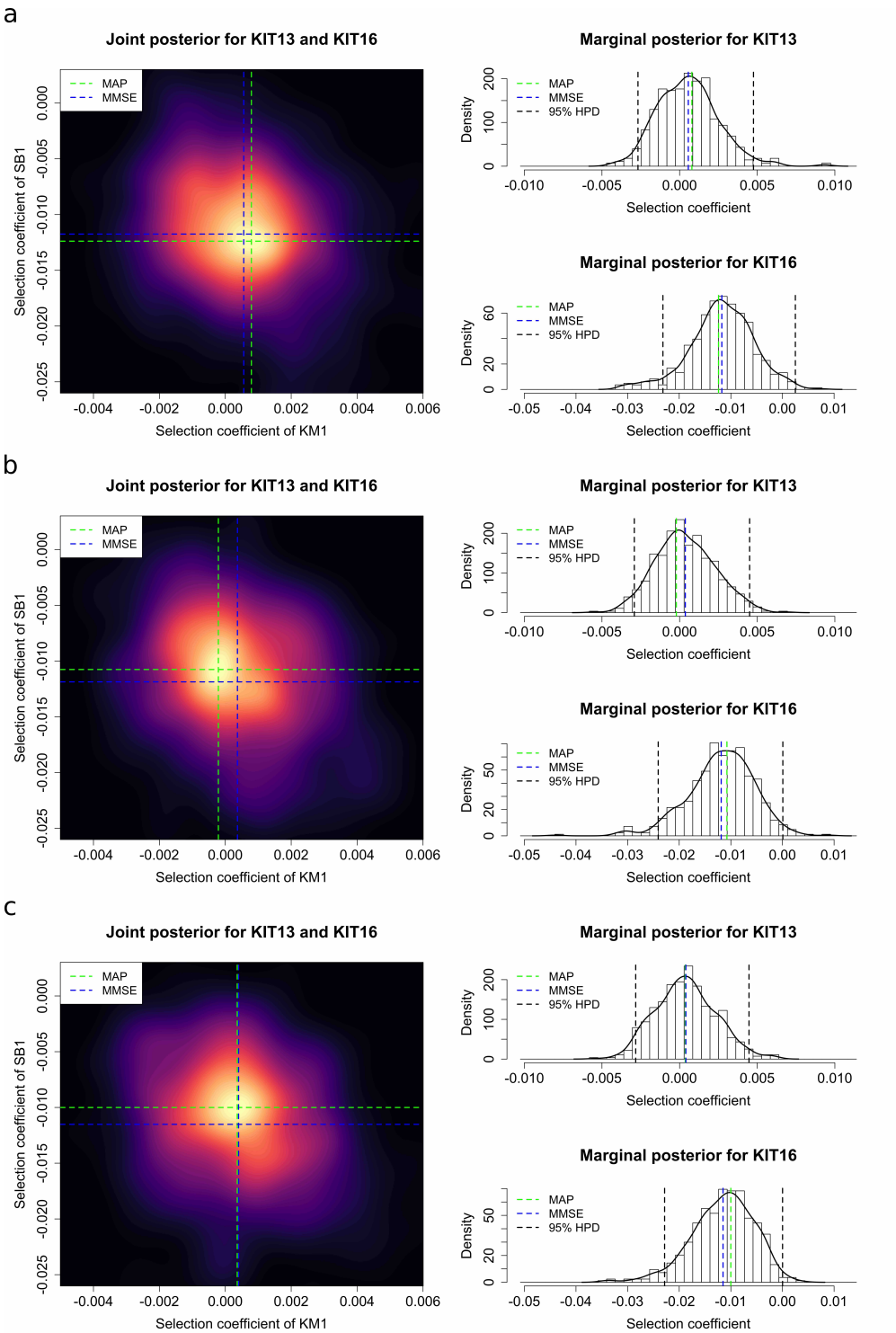

Figure 7: Posterior probability distributions for *KIT13* and *KIT16* obtained using the two-locus method with the population size of 16000 from the samples dated from 5472 years BP, with average rate of recombination (a) 5×10^{-9} crossovers/bp. (b) 1×10^{-8} crossovers/bp. (c) 5×10^{-8} crossovers/bp.

	recombination rate	MAP ($\times 10^{-2}$)	MMSE ($\times 10^{-2}$)	95% HPD ($\times 10^{-2}$)
KIT13	0.234×10^{-4}	0.079	0.056	$[-0.268, 0.476]$
	0.469×10^{-4}	-0.021	0.037	$[-0.292, 0.451]$
	2.340×10^{-4}	0.036	0.040	$[-0.283, 0.447]$
KIT16	0.234×10^{-4}	-1.238	-1.175	$[-2.316, 0.025]$
	0.469×10^{-4}	-1.076	-1.187	$[-2.407, 0.007]$
	2.340×10^{-4}	-1.001	-1.152	$[-2.283, 0.002]$

Table 4: MAP and MMSE estimates, as well as the 95% HPD intervals, for *KIT13* and *KIT16* obtained by using the two-locus method with the population size of 16000 from the samples dated from 5472 years BP.

As can be seen in Table 4, the MMSE estimates with different selected values of the recombination rate are essentially unchanged, while the MAP estimates vary a bit more than the MMSE estimates. This can be caused by the way we obtain our MAP estimates, where the posterior probability distribution is approximated using nonparametric density estimation techniques. Hence, the MAP estimates may depend on the number of the iterations of the PMMH. The resulting Bayesian estimates of the selection coefficients suggest that the *KM1* allele at the *KIT13* locus is weakly positively selected whereas the *SB1* allele at the *KIT16* locus is strongly negatively selected. The 95% HPD intervals for both selection coefficients include the value 0. For the *KIT13* locus, the posterior probability for positive selection is 0.564, not strong evidence for it being positively selected. However, for the *KIT16* locus, the posterior probability for negative selection is 0.982, strong evidence to support the *SB1* allele at the *KIT16* locus being negatively selected. This conclusion is 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 S8 and S9. We have also run an empirical study of *KIT13* and *KIT16* with the estimated selection coefficients (see Supplemental Material, Figure S10).

Our finding can be explained by that the decreasing attractiveness of spotted horses since the Middle Ages due to the religious and cultural ideas (Wutke et al., 2016). According to ancient Roman records, solid horses were preferred to spotted horses as the latter were considered to be of inferior quality. During medieval times, spotted horses had a negative connotation after several epidemics, resulting in a lower religious prestige for these patterns. Additionally, people might no longer see the need to distinguish wild (solid) horses from domesticated (spotted) horses as wild populations gradually became scarcer and approached extinction. Further reasons for the spotted horses being negatively selected might have been novel developments in weaponry such as the longbow, with these horses being an easier target than solid ones.

We have also used our single-locus Bayesian inference procedure, described in Supplemental Material, File S3, to estimate the selection coefficients for the mutant alleles at the *KIT13* and *KIT16* loci under the case that sampled chromosomes contain unknown alleles. The resulting posterior probability distributions are shown in Figure 8, and the MAP and MMSE estimates, as well as the 95% HPD intervals, are summarised in Table 5. The resulting Bayesian estimates of the selection coefficients suggest that the *KM1* allele at the *KIT13* locus is weakly selectively advantageous whereas the *SB1* allele at the *KIT16* locus is weakly selectively deleterious. However, as can be observed in Figure 5, the posterior probability distributions for the *KIT13* and *KIT16* loci are both approximately symmetric about 0. This suggests that there is no evidence to support the *KM1* allele at the *KIT13* locus or the *SB1* allele at the *KIT16* locus being selected. Compared to the results shown in Figure 7 and Table 4, we fail to tease apart negative selection at the *KIT16* locus without modelling genetic recombination and local linkage.

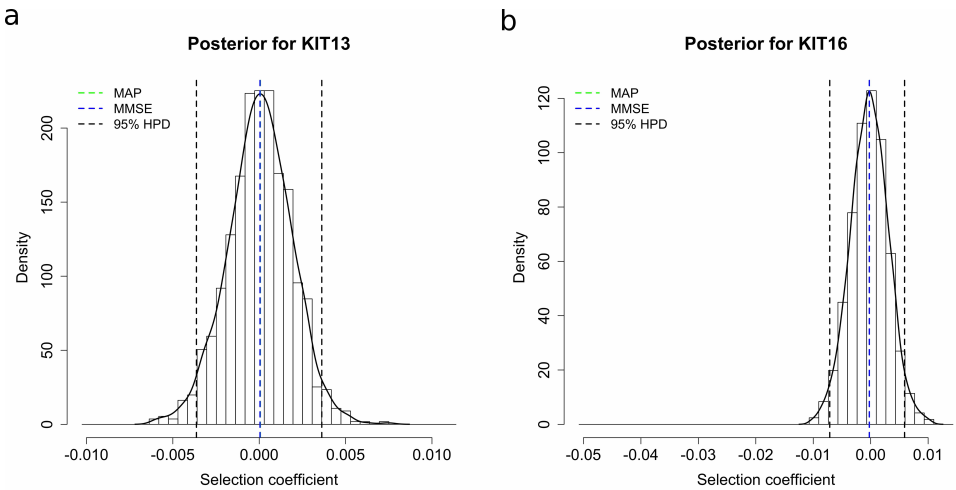

Figure 8: Posterior probability distributions for *KIT13* and *KIT16* obtained by using the single-locus method with the population size of 16000 from the samples dated from 5472 years BP. (a) *KIT13*. (b) *KIT16*.

	MAP ($\times 10^{-2}$)	MMSE ($\times 10^{-2}$)	95% HPD ($\times 10^{-2}$)
KIT13	0.006	0.005	$[-0.363, 0.362]$
KIT16	-0.023	-0.024	$[-0.713, 0.590]$

Table 5: MAP and MMSE estimates, as well as the 95% HPD intervals, for *KIT13* and *KIT16* obtained by using the single-locus method with the population size of 16000 from the samples dated from 5472 years BP.

3.3. Computational issues

In the PMMH algorithm, it is desirable to generate a large number of particles in the bootstrap particle filter to yield an accurate estimate of the marginal likelihood $p(\boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K} \mid \boldsymbol{\vartheta})$.

However, this can be computational burdensome since each iteration of the PMMH algorithm requires a run of the bootstrap particle filter even though fewer iterations are required. Balancing the particle number and the MCMC iteration number to obtain good performance at a reasonable computational cost was investigated by Pitt et al. (2012) and Doucet et al. (2015). In pseudo-marginal algorithms, if the estimates of the marginal likelihood are too noisy the chain tends to be 'sticky' with excessive autocorrelation (Beaumont, 2003). A simple rule-of-thumb is to select a particle number such that the standard deviation of the log-likelihood estimates is in the range from 1.0 to 1.7. Nevertheless, the PMMH algorithm exactly targets the marginal posterior $p(\boldsymbol{\vartheta} \mid \boldsymbol{u}_{1:K}, \boldsymbol{v}_{1:K})$ for any number of particles.

In each run of the bootstrap particle filter, we simulate the particles according to the Wright-Fisher diffusion with the Euler-Maruyama scheme. It is desirable to take a large L in the Euler-Maruyama scheme to get an accurate approximation of the Wright-Fisher diffusion, but large L increases the computational load. Stramer & Bognar (2011) suggested choosing L to be $L^\star$ such that the estimates of the marginal likelihood are approximately the same for any $L > L^\star$, where $L^\star$ can be obtained through extensive simulations.

In practice, we divide each generation into 5 subintervals in the Euler-Maruyama scheme, *i.e.*, $L = 5$. Our running time for a single iteration of the PMMH algorithm with 1500 particles (see Figure 9), achieving the standard deviation of the log-likelihood at approximately 1.504, on a single core of an Intel Core i7 processor at 4.2 GHz, is around 12.360 seconds for the *KIT* dataset. In principle, every particle can be simulated in parallel on a different core. Running 10000 iterations of the PMMH is sufficient for a relatively smooth resulting posterior surface, as shown in Figure 4. We discard the initial 2000 iterations as the burn-in period and then thin the remaining PMMH output, taking every fourth observation and regarding these as independent. Dahlin & Schön (2015) outlined a selected number of possible improvements and best practices for implementation. All of our code in this work is written in R with C++ by using Rcpp and RcppArmadillo.

Exact-approximate particle filtering approaches such as the PMMH algorithm we apply in this work seem to be useful for the inference of population genetic parameters from time series data of allele frequencies. This methodology can be generalised to a range of complex evolutionary scenarios, *e.g.*, non-constant population size. Although computationally demanding,

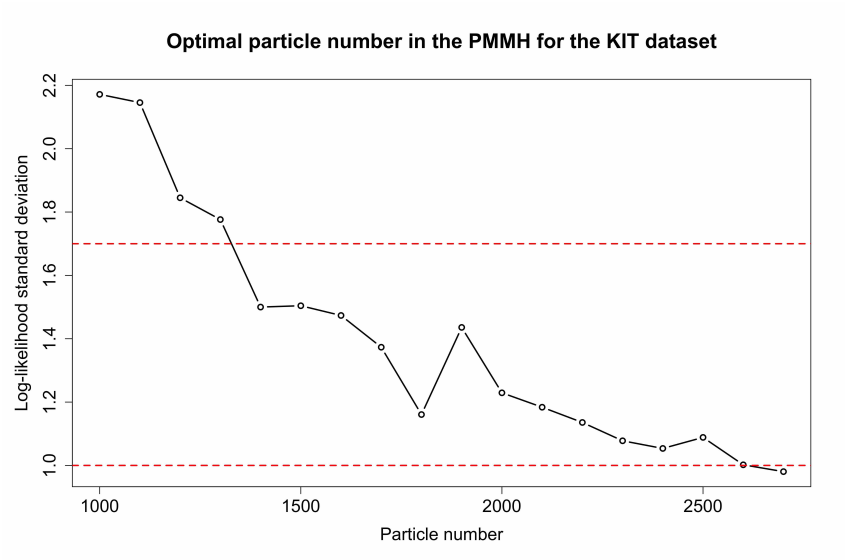

Figure 9: Changes in the standard deviation of the log-likelihood with the number of particles adopted in the PMMH for the *KIT* dataset.

improvements to the PMMH algorithm continue to be developed (*e.g.*, Yıldırım et al., 2018).

4. Discussion

In this work, we developed a novel MCMC-based method to infer natural selection at a pair of linked loci from allele frequency time series data while accounting for genetic recombination effect and local linkage information. Our Bayesian inference procedure was built on an HMM framework incorporating the two-locus Wright-Fisher diffusion with selection, and our Bayesian estimates of selection coefficients were achieved with the PMMH algorithm. We demonstrated that our approach could accurately and efficiently estimate selection coefficients from simulated data, regardless of whether sampled chromosomes contain unknown alleles or not. We found that under certain circumstances, existing single-locus methods failed to deliver precise estimates for selection coefficients, but our two-locus approach still worked well, especially for tightly linked loci. We applied our Bayesian inference procedure to the *KIT* gene in horses involved white spotting patterns and found strong evidence showing that the sabino pattern caused by the *SB1* allele at locus *KIT16* was selectively deleterious but no evidence showing that the tobiano pattern caused by the *KM1* allele at locus *KIT13* was selectively advantageous.

In addition to our method, Terhorst et al. (2015) is the only existing approach that can model linked loci and genetic drift for the inference of natural selection from temporal changes in allele frequencies. In Terhorst et al. (2015), the underlying population dynamics at multiple

linked loci was modelled using the Wright-Fisher model in their HMM framework, and the likelihood computation was carried out by approximating the Wright-Fisher model through a deterministic path with added Gaussian noise, which aims to fit a mathematically convenient transition probability density function by equating the first two moments of the Wright-Fisher model. Such a moment-based approximation works well for many applications when modelling the allele frequencies with intermediate values, but as soon as the allele frequencies get close to their boundaries (*i.e.*, allele fixation or loss), the Wright-Fisher model is poorly approximated by a Gaussian distribution that has infinite support hence will leak probability mass into the frequency values that are larger than 1 and smaller than 0, which is not mathematically possible. This can be problematic in the inference of natural selection because natural selection is expected to rapidly drive allele frequencies towards the boundaries 0 or 1. In practice, the approach of Terhorst et al. (2015) is tailored toward Pool-Seq time series data from E&R experiments and is restricted to analysing only a small number of linked loci simultaneously, *i.e.*, typically three linked loci.

The MCMC-based method we have developed in this work is built on the standard diffusion limit of the Wright-Fisher model of the stochastic evolutionary dynamics under natural selection at a pair of linked loci, which is shown to be a good approximation even if the allele frequencies get close to their boundaries (He et al., 2019a). The diffusion approximation allows our method to work well for the allele frequencies with all possible values. As our method relies on simulating the Wright-Fisher diffusion, it can be readily extended to model a range of complex evolutionary scenarios, *e.g.*, time-varying population size and selection coefficients. Our approach can handle sampled chromosomes that contain unknown alleles, which one might expect to encounter in real data. Even though we only illustrated the utility of our method on aDNA data in this work, our approach can also be used to analyse Pool-Seq time series data from E&R experiments, as in Terhorst et al. (2015).

One limitation of our approach is that it assumes that mutant alleles were created before the initial sampling time point. Once a sample contains at least one copy of the mutant allele, we can reasonably assume that the mutant allele arose before the time of that sample. However in the case of earlier samples without the mutant allele, there is uncertainty in pinpointing when the mutant allele arose. This problem can be remedied by co-estimating the allele age as in

e.g., Malaspinas et al. (2012); Schraiber et al. (2016); He et al. (2019b). These works all deal with natural selection at a single locus. Co-estimating selection coefficients at linked loci along with their allele ages can be expected to be cumbersome as there are many cases to consider. In the case of the ancient horse data, we did not wish to make the assumption that the mutant alleles, *KM1* and *SB1*, arose earlier than the time that horses were domesticated. However, we can compare the inference results obtained with different choices of the initial sampling time point and reach the same conclusion that there is no strong evidence for the *KM1* allele at locus *KIT13* to be positively selected, but there is strong evidence for the *SB1* allele at locus *KIT16* to be negatively selected.

Importantly, our Bayesian statistical framework lends itself to being extended to infer natural selection at multiple linked loci from time series data of allele frequencies, which might further improve the results of natural selection inference. The challenge is that with the increase in the number of linked loci, modelling the underlying population dynamics under natural selection becomes increasingly difficult. For example, there are eight haplotypes to consider in the case of three loci each with two alleles. In practice, it will be necessary to find a good approximation for the method to be computationally feasible, which will be the topic of future investigation. An important consideration is to what degree inferences are affected by the choice of stochastic or deterministic dynamics for the allele frequency trajectories (Jewett et al., 2016), and whether in many scenarios approximation with a deterministic model is satisfactory.

Acknowledgements

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

References

Andrieu, C., Doucet, A., & Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*, 269–342.

Andrieu, C., & Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, *37*, 697–725.

Andrieu, C., & Vihola, M. (2016). Establishing some order amongst exact approximations of MCMCs. *The Annals of Applied Probability*, *26*, 2661–2696.

Bank, C., Ewing, G. B., Ferrer-Admettla, A., Foll, M., & Jensen, J. D. (2014). Thinking too positive? Revisiting current methods of population genetic selection inference. *Trends in Genetics*, *30*, 540–546.

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, *164*, 1139–1160.

Bollback, J. P., & Huelsenbeck, J. P. (2007). Clonal interference is alleviated by high mutation rates in large populations. *Molecular Biology and Evolution*, *24*, 1397–1406.

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

Brooks, S. A., & Bailey, E. (2005). Exon skipping in the KIT gene causes a Sabino spotting pattern in horses. *Mammalian Genome*, *16*, 893–902.

Brooks, S. A., Lear, T. L., Adelson, D. L., & Bailey, E. (2007). A chromosome inversion near the KIT gene and the Tobiano spotting pattern in horses. *Cytogenetic and Genome Research*, *119*, 225–230.

Burke, M. K., Dunham, J. P., Shahrestani, P., Thornton, K. R., Rose, M. R., & Long, A. D. (2010). Genome-wide analysis of a long-term evolution experiment with *Drosophila*. *Nature*, *467*, 587–590.

Cuthbertson, C., Etheridge, A., & Yu, F. (2012). Fixation probability for competing selective sweeps. *Electronic Journal of Probability*, *17*, 1–36.

Dahlin, J., & Schön, T. B. (2015). Getting started with particle Metropolis-Hastings for inference in nonlinear dynamical models. *arXiv preprint arXiv:1511.01707*, .

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.

Doucet, A., Pitt, M. K., Deligiannidis, G., & Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, *102*, 295–313.

Dumont, B. L., & Payseur, B. A. (2008). Evolution of the genomic rate of recombination in mammals. *Evolution*, *62*, 276–294.

Fearnhead, P., & Künsch, H. R. (2018). Particle filters and data assimilation. *Annual Review of Statistics and Its Application*, *5*, 421–449.

Feder, A. F., Kryazhimskiy, S., & Plotkin, J. B. (2014). Identifying signatures of selection in genetic time series. *Genetics*, *196*, 509–522.

Ferrer-Admetlla, A., Leuenberger, C., Jensen, J. D., & Wegmann, D. (2016). An approximate Markov model for the Wright-Fisher diffusion and its application to time series data. *Genetics*, *203*, 831–846.

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

Foll, M., Poh, Y.-P., Renzette, N., Ferrer-Admetlla, A., Bank, C., Shim, H., Malaspinas, A.-S., Ewing, G., Liu, P., Wegmann, D., Caffrey, D. R., Zeldovich, K. B., Bolon, D. N., Wang, J. P., Kowalik, T. F., Schiffer, C. A., Finberg, R. W., & Jensen, J. D. (2014). Influenza virus drug resistance: a time-sampled population genetics perspective. *PLoS Genetics*, *10*, e1004185.

Foll, M., Shim, H., & Jensen, J. D. (2015). WFABC: a Wright-Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. *Molecular Ecology Resources*, *15*, 87–98.

Gordon, N. J., Salmond, D. J., & Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, *140*, 107–113.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, *5*, e1000695.

He, Z., Beaumont, M. A., & Yu, F. (2017). Effects of the ordering of natural selection and population regulation mechanisms on Wright-Fisher models. *G3: Genes, Genomes, Genetics*, *7*, 2095–2106.

He, Z., Beaumont, M. A., & Yu, F. (2019a). A numerical solution of the Wright-Fisher stochastic differential equation with application to transition probability density approximation. Manuscript submitted for publication.

He, Z., Dai, X., Beaumont, M. A., & Yu, F. (2019b). Maximum likelihood estimation of natural selection and allele age from time series data of allele frequencies. *bioRxiv*, (p. 837310).

Holder, K. K., & Bull, J. J. (2001). Profiles of adaptation in two similar viruses. *Genetics*, *159*, 1393–1404.

Hummel, S., Schmidt, D., Kremeyer, B., Herrmann, B., & Oppermann, M. (2005). Detection of the CCR5-Δ32 HIV resistance gene in Bronze Age skeletons. *Genes and Immunity*, *6*, 371–374.

Izenman, A. J. (1991). Review papers: Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, *86*, 205–224.

Jewett, E. M., Steinrücken, M., & Song, Y. S. (2016). The effects of population size histories on estimates of selection coefficients from time-series genetic data. *Molecular Biology and Evolution*, *33*, 3002–3027.

Kloeden, P. E., & Platen, E. (1992). *Numerical Solution of Stochastic Differential Equations*. Berlin: Springer-Verlag.

Lacerda, M., & Seoighe, C. (2014). Population genetics inference for longitudinally-sampled mutants under strong selection. *Genetics*, *198*, 1237–1250.

Lang, G. I., Rice, D. P., Hickman, M. J., Sodergren, E., Weinstock, G. M., Botstein, D., &

Desai, M. M. (2013). Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature*, *500*, 571–574.

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.

Mathieson, I., & McVean, G. (2013). Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics*, *193*, 973–984.

Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert, M., Cappellini, E., Petersen, B., Moltke, I., Johnson, P. L. F., Fumagalli, M., Vilstrup, J. T., Raghavan, M., Korneliussen, T., Malaspinas, A.-S., Vogt, J., Szklarczyk, D., Kelstrup, C. D., Vinther, J., Dolocan, A., Stenderup, J., Velazquez, A. V. V., Cahill, J., Rasmussen, M., Wang, X., Min, J., Zazula, G. D., Seguin-Orlando, A., Mortensen, C., Magnussen, K., Thompson, J. F., Weinstock, J., Gregersen, K., Røed, K. H., Eisenmann, V., Rubin, C. J., Miller, D. C., Antczak, D. F., Bertelsen, M. F., Brunak, S., Al-Rasheid, K. A. S., Ryder, O., Andersson, L., Mundy, J., Krogh, A., Gilbert, M. T. P., Kjær, K., Sicheritz-Ponten, T., Jensen, L. J., Olsen, J. V., Hofreiter, M., Nielsen, R., Shapiro, B., Wang, J., & Willerslev, E. (2013). Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*, *499*, 74–78.

Orozco-terWengel, P., Kapun, M., Nolte, V., Kofler, R., Flatt, T., & Schlötterer, C. (2012). Adaptation of *Drosophila* to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. *Molecular Ecology*, *21*, 4931–4941.

Outram, A. K., Stear, N. A., Bendrey, R., Olsen, S., Kasparov, A., Zaibert, V., Thorpe, N., & Evershed, R. P. (2009). The earliest horse harnessing and milking. *Science*, *323*, 1332–1335.

Pitt, M. K., dos Santos Silva, R., Giordani, P., & Kohn, R. (2012). On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, *171*, 134–151.

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.

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

Shim, H., Laurent, S., Matuszewski, S., Foll, M., & Jensen, J. D. (2016). Detecting and quantifying changing selection intensities from time-sampled polymorphism data. *G3: Genes, Genomes, Genetics*, *6*, 893–904.

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.

Stramer, O., & Bognar, M. (2011). Bayesian inference for irreducible diffusion processes using the pseudo-marginal approach. *Bayesian Analysis*, *6*, 231–258.

Tataru, P., Simonsen, M., Bataillon, T., & Hobolth, A. (2017). Statistical inference in the Wright-Fisher model using allele frequency data. *Systematic Biology*, *66*, e30–e46.

Terhorst, J., Schlötterer, C., & Song, Y. S. (2015). Multi-locus analysis of genomic time series data from experimental evolution. *PLoS Genetics*, *11*, e1005069–e1005069.

Wichman, H. A., Badgett, M. R., Scott, L. A., Boulianne, C. M., & Bull, J. J. (1999). Different trajectories of parallel evolution during viral adaptation. *Science*, *285*, 422–424.

Wichman, H. A., Millstein, J., & Bull, J. J. (2005). Adaptive molecular evolution for 13,000 phage generations. *Genetics*, *170*, 19–31.

Williamson, E. G., & Slatkin, M. (1999). Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics*, *152*, 755–761.

Wiser, M. J., Ribeck, N., & Lenski, R. E. (2013). Long-term dynamics of adaptation in asexual populations. *Science*, *342*, 1364–1367.

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.

Yıldırım, S., Andrieu, C., & Doucet, A. (2018). Scalable Monte Carlo inference for state-space models. *arXiv preprint arXiv:1809.02527*, .

Yu, F., & Etheridge, A. (2010). The fixation probability of two competing beneficial mutations. *Theoretical Population Biology*, *78*, 36–45.

Zhang, X. (2006). Euler-Maruyama approximations for SDEs with non-Lipschitz coefficients and applications. *Journal of Mathematical Analysis and Applications*, *316*, 447–458.

Appendix A. Euler-Maruyama method for the solution of the Wright-Fisher SDE

A number of numerical methods for SDEs have been developed (see Kloeden & Platen, 1992, for an excellent introduction). The numerical approach we adopt in this work is the commonly used Euler-Maruyama scheme, one of the most popular numerical methods for SDEs in practice due to its high efficiency and low complexity. Moreover, the Wright-Fisher SDE formulated in the form of He et al. (2019a) enables us to avoid boundary issues and reduce computational costs in the Euler-Maruyama scheme.

We divide each generation into L subintervals by setting $\Delta t = \frac{1}{2NL}$, and then the Euler-Maruyama approximation of the Wright-Fisher diffusion can be written as

$$\hat{X}_i(t + \Delta t) = \hat{X}_i(t) + \mu_i(\hat{\boldsymbol{X}}(t))\Delta t + \sum_{j=1}^{6} \sigma_{ij}(\hat{\boldsymbol{X}}(t))\Delta W_j(t),$$

for $i = 1, 2, 3, 4$, where $\Delta W_j(t) = W_j(t + \Delta t) - W_j(t)$ are independent and normally distributed with mean 0 and variance Δt for $j = 1, 2, \ldots, 6$. The Euler-Maruyama scheme is numerically stable and strongly consistent (see, *e.g.*, Kloeden & Platen, 1992), and the convergence of the Euler-Maruyama approximation of the Wright-Fisher diffusion is guaranteed by Zhang (2006).