An MCMC-based method for Bayesian inference of natural selection from time series DNA data across linked loci

Zhangyi He[a,*], 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 rapid improvement of DNA sequencing technology has made it possible to monitor genomes in great detail over time, which presents an opportunity for investigating natural selection based on time serial samples of genomes while accounting for genetic recombination. Such time series genetic data allow for more precise estimates for population genetic quantities and hypothesis testing on the recent action of natural selection. In the present work, we develop a novel Bayesian statistical framework for the inference of natural selection by capitalising on the temporal aspect of DNA data across linked loci, regardless of whether sampled chromosomes have been phased or not. Our approach relies on a hidden Markov model incorporating the two-locus Wright-Fisher diffusion with selection, which enables us to explicitly model genetic recombination. The posterior probability distribution for selection coefficients is obtained by using the particle marginal Metropolis-Hastings algorithm, which also allows for co-estimation of the population haplotype frequency trajectories. The performance of our method is evaluated through extensive simulations, showing that our estimates for selection coefficients are unbiased. Moreover, under certain circumstances, we find that our method can deliver precise estimates for selection coefficients whereas existing single-locus approaches fail, especially for tightly linked loci. The utility of our method is illustrated with an application to the loci encoding white spotting patterns in horses, where we also show that our method can handle missing values.

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

Author summary

We develop a novel Bayesian approach to estimate the selection coefficients of DNA data sampled at different times, known as time series genetic data, which can originate from ancient DNA (aDNA) sampled from archaeological sites or evolve-and-resequence (E&R) experiments. Our method improves upon previous work on this topic by considering the possibility of two selected loci along the same chromosomes, which will be competing against each other as well as subject to genetic recombination. We demonstrate through extensive simulations that our method produces essentially unbiased estimates for the selection coefficients, whereas considering the two loci independently, *i.e.*, using a single-locus approach, can lead to incorrect conclusions. We apply our method to ancient horse DNA data, where the *KIT13* and *KIT16* loci are both found on chromosome 13 and are associated with white spotting patterns in horses. We arrive at the conclusion that the mutant allele at the *KIT13* locus is weakly advantageous, whereas the mutant allele at the *KIT16* locus is weakly deleterious.

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 estimating 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; Allentoft 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 time series genetic 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 (see Bank et al., 2014; Malaspinas, 2016, for detailed reviews). A common method for analysing time series genetic data is based on the hidden Markov model (HMM) framework of Bollback et al. (2008), 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 approaches can become computationally infeasible for large populations 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 the 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) or the 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 based on the observed samples 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 investigating natural selection from time series genetic data across linked loci while accounting for the process of genetic recombination. Properly taking genetic recombination into account can be expected to provide a more precise estimate for the selection coefficient and more accurate hypothesis testing on the recent action of natural selection since genetic recombination may either reinforce or oppose the changes in allele frequencies caused by natural selection (He et al., 2017a). However, with the exception of Terhorst et al. (2015), all existing methods based on the Wright-Fisher model for inferring natural selection from time series data of allele frequencies are limited to a single locus, *i.e.*, a selected locus where allele frequency data are available is assumed to be independent of all other loci. Terhorst et al. (2015) extends a moment-based approximation

of the Wright-Fisher model introduced by Feder et al. (2014) to multiple loci, where the gene frequency transition between two given time points is modelled deterministically, with added Gaussian noise. 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 develop a novel Bayesian method for the inference of natural selection from time series genetic data across two linked loci while accounting for genetic recombination. The key innovation of our Bayesian inference procedure is that we use the Wright-Fisher diffusion of the stochastic evolutionary dynamics under natural selection at two linked loci as the hidden Markov process to characterise the temporal changes in the haplotype frequencies of the underlying population, which allows for explicitly modelling the process of genetic recombination. In addition, the diffusion approximation of the Wright-Fisher model enables us to overcome the restriction imposed by the Gaussian approximation utilised by Terhorst et al. (2015) that the allele frequency trajectory of the underlying population remains far away from allele fixation or loss. Our posterior computation is carried out with the particle marginal Metropolis-Hastings (PMMH) algorithm developed by Andrieu et al. (2010), which also allows for co-estimation of the haplotype frequency trajectories of the underlying population. Compared to Terhorst et al. (2015), our method is limited to two linked loci but can handle different types of DNA data, *e.g.*, aDNA data. Moreover, we model both phased and unphased chromosomes in our approach with the additional flexibility of modelling potential missing values, which is common in aDNA data due to postmortem DNA damage.

We evaluate the performance of our Bayesian inference procedure with extensive simulations, showing that our approach allows for efficient and accurate estimation of selection coefficients from time series DNA data across two linked loci, regardless of whether sampled chromosomes are phased or not. We also employ it to analyse the aDNA data associated with white spotting patterns in horses from Wutke et al. (2016), where unknown alleles in the sample are allowed. Moreover, we give two examples to show that our approach can deliver precise estimates for selection coefficients whereas existing single-locus methods fail under certain circumstances, especially for tightly linked loci. This suggests that the improvement in natural selection inference accuracy can be attained by modelling the process of genetic recombination. Without accounting for genetic recombination, we can always get the estimates of selection coefficients with a

single-locus method, which are often unfalsifiable and reasonable but sometimes not correct.

2. Materials and methods

In this section, we begin with a short review of the Wright-Fisher diffusion for two linked loci evolving under natural selection presented in He et al. (2017a), and then we describe our Bayesian method for the inference of natural selection from the time series genetic data sampled from two linked loci, *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 MCMC techniques.

2.1. Wright-Fisher diffusion

Consider a monoecious population of randomly mating diploid individuals at two linked autosomal 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., 2017b, for more details), 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 on both loci $\mathcal{A}_1\mathcal{B}_1$, $\mathcal{A}_1\mathcal{B}_2$, $\mathcal{A}_2\mathcal{B}_1$ and $\mathcal{A}_2\mathcal{B}_2$, labelled haplotypes 1, 2, 3 and 4, respectively. We attach the symbols $\mathcal{A}_1$ and $\mathcal{B}_1$ to the mutant alleles, which are assumed to arise only once in the population (*i.e.*, there is no recurrent mutation) and be selectively advantageous, and we attach the 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 constant 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 (*i.e.*, directional selection) 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 designate the recombination rate of two loci on the same chromosome by $r \in [0, 0.5]$ (*i.e.*, the fraction of recombinant offspring showing a crossover between two loci), and we assume that the population size is fixed to be N individuals over time (corresponding to $2N$ chromosomes).

2.1.1. Stochastic differential equation formulation

In this work, 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 $N \to \infty$. As the population size goes to infinity, the haplotype frequency trajectories follow a standard diffusion limit of the two-locus Wright-Fisher model with selection (He et al., 2017a). The Wright-Fisher diffusion is typically formulated in terms of the partial differential equation (PDE) satisfied by its transition probability density function (*e.g.*, Bollback et al., 2008; Gutenkunst et al., 2009; Steinrücken et al., 2014), but in this work we characterise it as the solution of the stochastic differential equation (SDE) instead (*e.g.*, Schraiber et al., 2016).

More specifically, we let $X_i(t)$ be the frequency of haplotype i in the population at time t for $i = 1, 2, 3, 4$ and designate the haplotype frequencies of the four possible types in the population by $\boldsymbol{X}(t)$, which evolves in the state space

$$\Omega_{\boldsymbol{X}} = \left\{ \boldsymbol{x} \in [0,1]^4 : x_1 + x_2 + x_3 + x_4 = 1 \right\},$$

and satisfies the SDE in the form of He et al. (2017a)

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

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 drift term $\boldsymbol{\mu}(\boldsymbol{x})$ and the diffusion term $\boldsymbol{\sigma}(\boldsymbol{x})$ satisfy the following non-Lipschitz conditions:

$$|\boldsymbol{\sigma}(\boldsymbol{x}) - \boldsymbol{\sigma}(\boldsymbol{x}')|^2 \leq |\boldsymbol{x} - \boldsymbol{x}'|^2 \cdot g(|\boldsymbol{x} - \boldsymbol{x}'|) \tag{4}$$

$$|\boldsymbol{\mu}(\boldsymbol{x}) - \boldsymbol{\mu}(\boldsymbol{x}')| \leq |\boldsymbol{x} - \boldsymbol{x}'| \cdot g(|\boldsymbol{x} - \boldsymbol{x}'|) \tag{5}$$

for all $\boldsymbol{x}, \boldsymbol{x}' \in \Omega_{\boldsymbol{X}}$, where $|\cdot|$ denotes the Euclidean vector norm or the Frobenius matrix norm as appropriate and

$$g(x) = \left(\log\left(\frac{1}{x}\right) \vee 1\right)^{\frac{1}{2}}.$$

The non-Lipschitz conditions (4) and (5) are sufficient to ensure the existence and uniqueness of the solution for the Wright-Fisher SDE in Eqs. (1)-(3) with Yamada & Watanabe (1971) and Taniguchi (1992).

2.1.2. Euler-Maruyama scheme

A number of numerical methods for SDEs have been developed (see Kloeden & Platen, 1992, for an excellent introduction). The numerical approach we adopt here 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. In addition, the Wright-Fisher SDE formulated in the form of He et al. (2017a) 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

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

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 $j = 1, 2, \ldots, 6$. According to Kloeden & Platen (1992), the Euler-Maruyama scheme is numerically stable and strongly consistent, and the convergence of the Euler-Maruyama approximation of the Wright-Fisher diffusion in Eq. (6) is guaranteed by the non-Lipschitz conditions in Eq. (5) (Zhang, 2006).

2.2. Bayesian inference of natural selection

We consider any two loci along the chromosome, labelled $\mathcal{A}$ and $\mathcal{B}$, and recall that the mutant alleles at the $\mathcal{A}$ and $\mathcal{B}$ loci are both favoured by natural selection. Let $\boldsymbol{\vartheta} = (\alpha_{\mathcal{A}}, h_{\mathcal{A}}, \alpha_{\mathcal{B}}, h_{\mathcal{B}}, \rho)$ denote the population genetic quantities of interest, where $\alpha_{\mathcal{A}}$ and $\alpha_{\mathcal{B}}$ are the scaled selection coefficients, $h_{\mathcal{A}}$ and $h_{\mathcal{B}}$ are the dominance parameters, and ρ is the scaled recombination rate. In this section, we introduce the HMM framework incorporating the two-locus Wright-Fisher diffusion with selection and discuss the Bayesian procedure for the inference of natural selection from time serial samples of phased or unphased chromosomes in detail, including how to handle unknown alleles in the sample. It should be noted that in this work we assume that the mutant alleles have already existed in the population before the sampling period.

2.2.1. Hidden Markov model

In our HMM framework, we assume that the underlying population evolves according to the Wright-Fisher diffusion in Eqs. (1)-(3), and the observations are modelled as independent samples drawn from the underlying population at time points $t_1 < t_2 < \ldots < t_K$. The sampling time points are measured in units of $2N$ generations to be consistent with the time scale in the Wright-Fisher diffusion. At the k-th sampling time point, there are n_k chromosomes drawn from the underlying population. If sampled chromosomes are unphased, we let $\boldsymbol{c}_k = (c_k^{\mathcal{A}}, c_k^{\mathcal{B}})$ denote the observed counts of mutant alleles at loci $\mathcal{A}$ and $\mathcal{B}$, and $\boldsymbol{m}_k = (m_k^{\mathcal{A}}, m_k^{\mathcal{B}})$ denote the observed counts of missing values at loci $\mathcal{A}$ and $\mathcal{B}$. On the other hand, if sampled chromosomes are phased, we let the symbol $*$ indicate the unknown allele at the corresponding locus and designate $\boldsymbol{c}_k = (c_{1,k}, c_{2,k}, c_{3,k}, c_{4,k})$ as the observed 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 and $\boldsymbol{m}_k = (m_{1,k}^{\mathcal{A}}, m_{2,k}^{\mathcal{A}}, m_{1,k}^{\mathcal{B}}, m_{2,k}^{\mathcal{B}}, m_k^{\mathcal{AB}})$ as the observed counts of missing values in the $*\mathcal{B}_1$, $*\mathcal{B}_2$, $\mathcal{A}_1*$, $\mathcal{A}_2*$ and $**$ haplotypes.

To find the Bayesian estimates of the population genetic parameters $\boldsymbol{\vartheta}$, we need to compute the posterior probability distribution $p(\boldsymbol{\vartheta} \mid \boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K})$, which can be obtained by conditioning

and integrating over all possible haplotype frequency trajectories of the underlying population at each sampling time point. Let $\boldsymbol{x}_{1:K} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K)$ be the haplotype frequency trajectories of the underlying population at the sampling time points $\boldsymbol{t}_{1:K}$, and then the posterior probability distribution for the population genetic quantities of interest can be formulated as

$$p\left(\boldsymbol{\vartheta} \mid \boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K}\right) = \int p\left(\boldsymbol{\vartheta}, \boldsymbol{x}_{1:K} \mid \boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K}\right) d\boldsymbol{x}_{1:K}. \tag{7}$$

Under our HMM framework, the joint posterior probability distribution for the population genetic quantities of interest and the haplotype frequency trajectories of the underlying population can be represented as

$$p\left(\boldsymbol{\vartheta}, \boldsymbol{x}_{1:K} \mid \boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K}\right) \propto p\left(\boldsymbol{\vartheta}\right) p\left(\boldsymbol{x}_{1:K} \mid \boldsymbol{\vartheta}\right) p\left(\boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K} \mid \boldsymbol{x}_{1:K}\right),$$

where $p(\boldsymbol{\vartheta})$ is the prior probability distribution for the population genetic quantities of interest and can be taken to be an uniform prior if their 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 $p(\boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K} \mid \boldsymbol{x}_{1:K})$ 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 formulated as

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

where $p(\boldsymbol{x}_1 \mid \boldsymbol{\vartheta})$ is the prior probability distribution for the haplotype frequencies of the underlying population at the first sampling time point and can be taken to be an uniform prior if the prior knowledge of the initial haplotype frequencies of the underlying population is poor, and $p(\boldsymbol{x}_{k+1} \mid \boldsymbol{x}_k; \boldsymbol{\vartheta})$ is the transition probability density function of the Wright-Fisher diffusion between two consecutive sampling time points for $k = 1, 2, \ldots, K-1$. The transition probability density function $p(\boldsymbol{x}_{k+1} \mid \boldsymbol{x}_k; \boldsymbol{\vartheta})$ can be obtained by numerically solving the Kolmogorov backward equation (or its adjoint) associated with the Wright-Fisher diffusion, but this requires a fine

enough discretisation of the state space and is strongly dependent on the underlying population genetic quantities of interest. Furthermore, it is computationally challenging and prohibitively expensive to numerically solve such a PDE in three dimensions for our posterior computation. Our 'exact-approximate' Monte Carlo procedure (Andrieu & Vihola, 2016) 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

$$p\left(\boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K} \mid \boldsymbol{x}_{1:K}\right) = \prod_{k=1}^{K} p\left(\boldsymbol{c}_k, \boldsymbol{m}_k \mid \boldsymbol{x}_k\right),$$

where $p(\boldsymbol{c}_k, \boldsymbol{m}_k \mid \boldsymbol{x}_k)$ 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 $p(\boldsymbol{c}_k, \boldsymbol{m}_k \mid \boldsymbol{x}_k)$, we need to introduce auxiliary variables $\boldsymbol{z}_k = (z_{1,k}, z_{2,k}, z_{3,k}, z_{4,k})$ to represent the true 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 may not be fully observed. We then have

$$p\left(\boldsymbol{c}_k, \boldsymbol{m}_k \mid \boldsymbol{x}_k\right) = \sum_{\boldsymbol{z}_k \in \Omega_{\boldsymbol{Z}_k}} p\left(\boldsymbol{z}_k \mid \boldsymbol{x}_k\right) p\left(\boldsymbol{c}_k, \boldsymbol{m}_k \mid \boldsymbol{z}_k\right), \tag{8}$$

where

$$\Omega_{\boldsymbol{Z}_k} = \left\{\boldsymbol{z}_k \in \mathbb{N}^4 : z_{1,k} + z_{2,k} + z_{3,k} + z_{4,k} = n_k\right\}.$$

Conditional on the haplotype frequency trajectories of the underlying population at the k-th sampling time point, the true haplotype counts of the sample can then be modelled through multinomial sampling from the underlying population at the $\mathcal{A}$ and $\mathcal{B}$ loci with sample size n_k, which means that we can formulate the first term in the summation of Eq. (8) as

$$p\left(\boldsymbol{z}_k \mid \boldsymbol{x}_k\right) = \frac{n_k!}{\prod_{i=1}^{4} z_{i,k}!} \prod_{i=1}^{4} x_{i,k}^{z_{i,k}}. \tag{9}$$

For the second term in the summation of Eq. (8), we need to consider the following two cases: the available data are drawn with (1) unphased chromosomes or (2) phased chromosomes.

If the available data are drawn with unphased chromosomes, we introduce auxiliary variables $\boldsymbol{u}_k = (u_k^{\mathcal{A}}, u_k^{\mathcal{B}})$, where $u_k^{\mathcal{A}}$ denotes the number of the chromosomes in the sample at the k-th

sampling time point that have unknown alleles (*i.e.*, missing values) at locus $\mathcal{A}$ but actually have the mutant allele. We can then write the second term in the summation of Eq. (8) as

$$p\left(\boldsymbol{c}_k, \boldsymbol{m}_k \mid \boldsymbol{z}_k\right) = \sum_{\boldsymbol{u}_k \in \Omega_{\boldsymbol{U}_k}} \mathbb{1}_{\{z_{1,k}+z_{2,k}=c_k^{\mathcal{A}}+u_k^{\mathcal{A}}\}}(\boldsymbol{u}_k) \cdot \mathbb{1}_{\{z_{1,k}+z_{3,k}=c_k^{\mathcal{B}}+u_k^{\mathcal{B}}\}}(\boldsymbol{u}_k), \tag{10}$$

where

$$\Omega_{\boldsymbol{U}_k} = \left\{\boldsymbol{u}_k \in \mathbb{N}^2 : u_k^{\mathcal{A}} \leq m_k^{\mathcal{A}}, u_k^{\mathcal{B}} \leq m_k^{\mathcal{B}}\right\}.$$

Note that $\mathbb{1}_A$ is the indicator function that equals to one if condition A holds and zero otherwise.

If the available data are drawn with phased chromosomes, we introduce auxiliary variables $\boldsymbol{u}_k = (u_{1,k}^{\mathcal{A}}, u_{2,k}^{\mathcal{A}}, u_{3,k}^{\mathcal{A}}, u_{4,k}^{\mathcal{A}}, u_{1,k}^{\mathcal{B}}, u_{2,k}^{\mathcal{B}}, u_{3,k}^{\mathcal{B}}, u_{4,k}^{\mathcal{B}}, u_{1,k}^{\mathcal{AB}}, u_{2,k}^{\mathcal{AB}}, u_{3,k}^{\mathcal{AB}}, u_{4,k}^{\mathcal{AB}})$, where $u_{1,k}^{\mathcal{A}}$ denotes the number of the chromosomes in the sample at the k-th sampling time point that have unknown alleles (*i.e.*, missing values) at locus $\mathcal{A}$ but actually have the $\mathcal{A}_1\mathcal{B}_1$ haplotype, and $u_{3,k}^{\mathcal{AB}}$ denotes the number of the chromosomes in the sample at the k-th sampling time point that have unknown alleles (*i.e.*, missing values) at loci $\mathcal{A}$ and $\mathcal{B}$ but actually have the $\mathcal{A}_2\mathcal{B}_1$ haplotype. We can then write the second term in the summation of Eq. (8) as

$$p\left(\boldsymbol{c}_k, \boldsymbol{m}_k \mid \boldsymbol{z}_k\right) = \sum_{\boldsymbol{u}_k \in \Omega_{\boldsymbol{U}_k}} \prod_{i=1}^{4} \mathbb{1}_{\{z_{i,k}=c_{i,k}+u_{i,k}^{\mathcal{A}}+u_{i,k}^{\mathcal{B}}+u_{i,k}^{\mathcal{AB}}\}}(\boldsymbol{u}_k), \tag{11}$$

where

$$\begin{aligned}\Omega_{\boldsymbol{U}_k} = \Big\{\boldsymbol{u}_k \in \mathbb{N}^{12} : & u_{1,k}^{\mathcal{AB}} + u_{2,k}^{\mathcal{AB}} + u_{3,k}^{\mathcal{AB}} + u_{4,k}^{\mathcal{AB}} = m_k^{\mathcal{AB}}, \\ & u_{1,k}^{\mathcal{A}} + u_{3,k}^{\mathcal{A}} = m_{1,k}^{\mathcal{A}}, u_{2,k}^{\mathcal{A}} + u_{4,k}^{\mathcal{A}} = m_{2,k}^{\mathcal{A}}, u_{1,k}^{\mathcal{B}} + u_{2,k}^{\mathcal{B}} = m_{1,k}^{\mathcal{B}}, u_{3,k}^{\mathcal{B}} + u_{4,k}^{\mathcal{B}} = m_{2,k}^{\mathcal{B}}\Big\}.\end{aligned}$$

2.2.2. Particle marginal Metropolis-Hastings

To obtain the marginal posterior probability distribution $p(\boldsymbol{\vartheta} \mid \boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K})$, we resort to MCMC techniques here as the posterior probability distribution in Eq. (7) is unavailable in a closed form. A Metropolis-Hastings (MH) scheme can be devised, with a fairly arbitrary proposal probability distribution, to explore the population genetic quantities of interest, where a sample of new candidates of the parameters $\boldsymbol{\vartheta}^\star$ is drawn from the proposal probability distri-

bution $q(\boldsymbol{\vartheta}^\star \mid \boldsymbol{\vartheta})$ and is accepted with the Metropolis-Hastings ratio

$$A = \frac{p\left(\boldsymbol{\vartheta}^\star\right)}{p\left(\boldsymbol{\vartheta}\right)} \frac{p\left(\boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K} \mid \boldsymbol{\vartheta}^\star\right)}{p\left(\boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K} \mid \boldsymbol{\vartheta}\right)} \frac{q\left(\boldsymbol{\vartheta} \mid \boldsymbol{\vartheta}^\star\right)}{q\left(\boldsymbol{\vartheta}^\star \mid \boldsymbol{\vartheta}\right)}. \tag{12}$$

Our problem reduces to calculating the intractable marginal likelihood $p(\boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K} \mid \boldsymbol{\vartheta})$, which can be 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 $p(\boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K} \mid \boldsymbol{\vartheta})$ is unbiased (or has a constant bias independent of the parameters $\boldsymbol{\vartheta}$) and targets the marginal posterior probability distribution $p(\boldsymbol{\vartheta} \mid \boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K})$.

We adopt a closely related method developed by Andrieu et al. (2010), which obtains an unbiased sequential Monte Carlo (SMC) estimate of the marginal likelihood $p(\boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K} \mid \boldsymbol{\vartheta})$ and targets the joint posterior probability distribution $p(\boldsymbol{\vartheta}, \boldsymbol{x}_{1:K} \mid \boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K})$. This approach is called particle marginal Metropolis-Hastings (PMMH), which permits a joint update of the population genetic quantities of interest and the haplotype frequency trajectories of the underlying population. 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 marginal posterior probability distribution $p(\boldsymbol{\vartheta} \mid \boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K})$. We therefore apply 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).

In our Bayesian inference procedure, the implementation of the PMMH algorithm requires the SMC estimate of the marginal likelihood in Eq. (12). 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 $p(\boldsymbol{\vartheta})$, and then we run a bootstrap particle filter with the proposed parameters $\boldsymbol{\vartheta}$ to obtain the SMC estimate of the marginal likelihood, denoted by $\hat{p}(\boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K} \mid \boldsymbol{\vartheta})$, which is equal to the product of average weights of all particles at the sampling time points $\boldsymbol{t}_{1:K}$. 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 $\hat{p}(\boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K} \mid \boldsymbol{\vartheta}^\star)$.

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

$$A = \frac{p\left(\boldsymbol{\vartheta}^\star\right)}{p\left(\boldsymbol{\vartheta}\right)} \frac{\hat{p}\left(\boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K} \mid \boldsymbol{\vartheta}^\star\right)}{\hat{p}\left(\boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K} \mid \boldsymbol{\vartheta}\right)} \frac{q\left(\boldsymbol{\vartheta} \mid \boldsymbol{\vartheta}^\star\right)}{q\left(\boldsymbol{\vartheta}^\star \mid \boldsymbol{\vartheta}\right)}.$$

Once enough samples of the parameters $\boldsymbol{\vartheta}$ have been obtained, we can compute the posterior probability distribution $p(\boldsymbol{\vartheta} \mid \boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K})$ and achieve the maximum a posteriori probability (MAP) estimates for the population genetic quantities of interest. Alternatively, we can obtain the minimum mean square error (MMSE) estimates for the population genetic quantities of interest directly from the samples of the parameters $\boldsymbol{\vartheta}$.

2.3. Data availability

The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article. Source code implementing the method described in this paper is available at `https://github.com/zhangyi-he/2L-WFD-PMMH`.

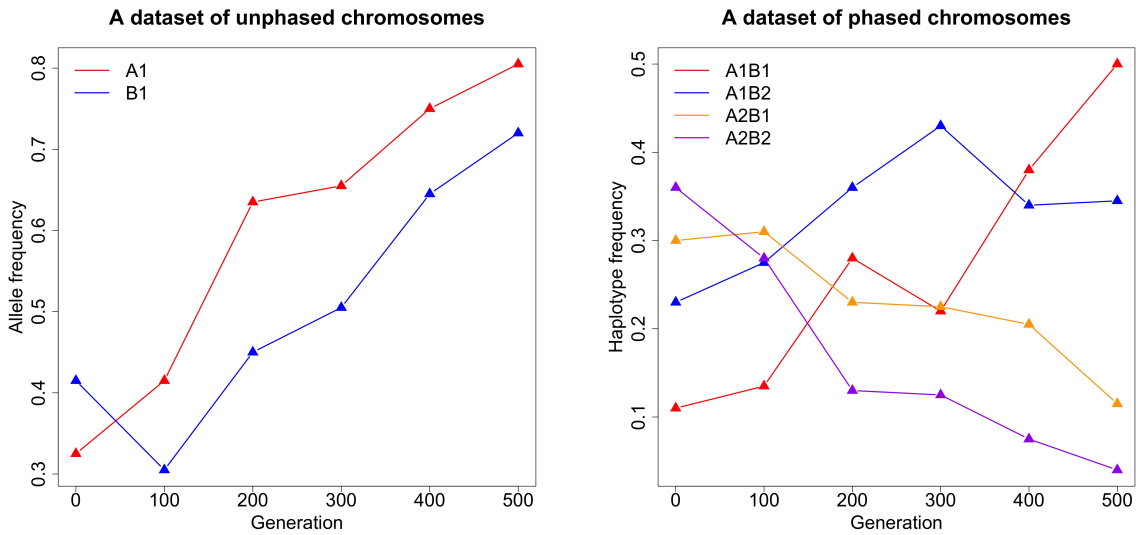
3. Results

In this section, we first demonstrate how our Bayesian inference procedure performs on several simulated datasets with given population genetic parameter values, including datasets simulated with phased and unphased chromosomes, respectively. We then apply our approach to time series aDNA data 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 ancient samples ranging from a pre- to a post-domestication period were sequenced.

3.1. Analysis of simulated data

To evaluate the performance of our Bayesian inference procedure, we run forward-in-time simulations of the two-locus Wright-Fisher model with selection (see, *e.g.*, He et al., 2017b) and assess the performance of our approach on these replicate simulations by examining the bias and root mean square error (RMSE) of our Bayesian estimates for selection coefficients. 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 dominance parameters $h_{\mathcal{A}}, h_{\mathcal{B}} \in [0, 1]$ and population size $N \in \mathbb{N}$.

For each simulated dataset, given values of population genetic parameters $\boldsymbol{\vartheta}$ and initial population haplotype frequencies $\boldsymbol{x}_0$, we simulate latent haplotype frequency trajectories of the underlying population according to the two-locus Wright-Fisher model with selection. After obtaining the simulated haplotype frequency trajectories of the underlying population, we draw the true sample haplotype counts independently at each sampling time point according to the multinomial distribution in Eq. (9). We generate the observations in the following two cases: (1) for unphased chromosomes we use Eq. (10) with $\boldsymbol{m}_{1:K} = \mathbf{0}$ to obtain sample allele counts $\boldsymbol{c}_{1:K}$, and (2) for phased chromosomes we use Eq. (11) with $\boldsymbol{m}_{1:K} = \mathbf{0}$ to obtain sample haplotype counts $\boldsymbol{c}_{1:K}$. We take observations every 100 generations from generation 0 to 500 each with a sample size of 200 chromosomes, and we set initial haplotype frequencies of the underlying population $\boldsymbol{x}_0 = (0.1, 0.2, 0.3, 0.4)$, selection coefficients $s_{\mathcal{A}} = 0.01$ and $s_{\mathcal{B}} = 0.005$, and recombination rate $r = 0.001$. See Figure 1a for the simulated dataset of unphased chromosomes and Figure 1b for the simulated dataset of phased chromosomes.



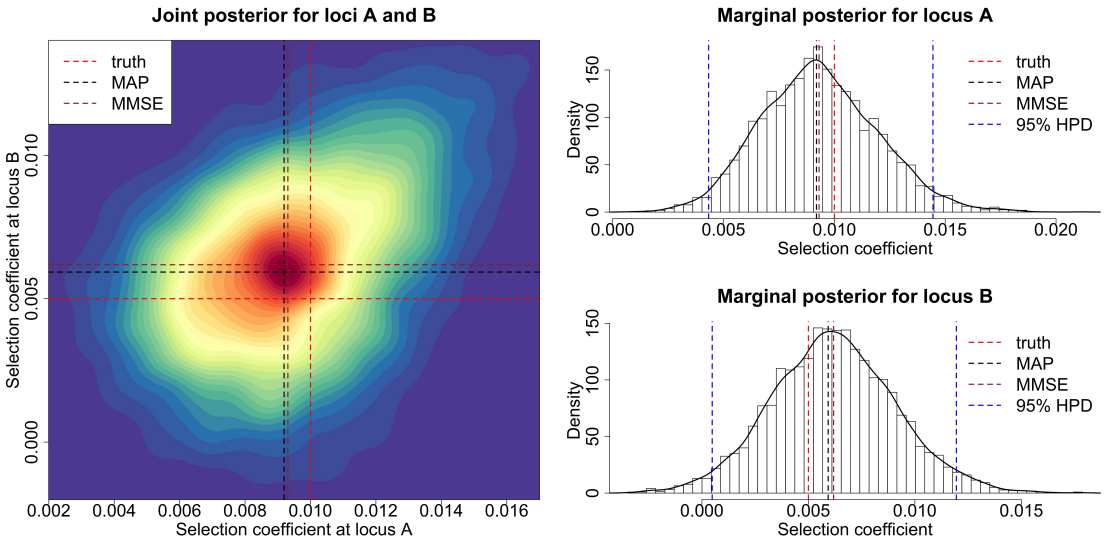
(a) A simulated dataset of unphased chromosomes.

(b) A simulated dataset of phased chromosomes.

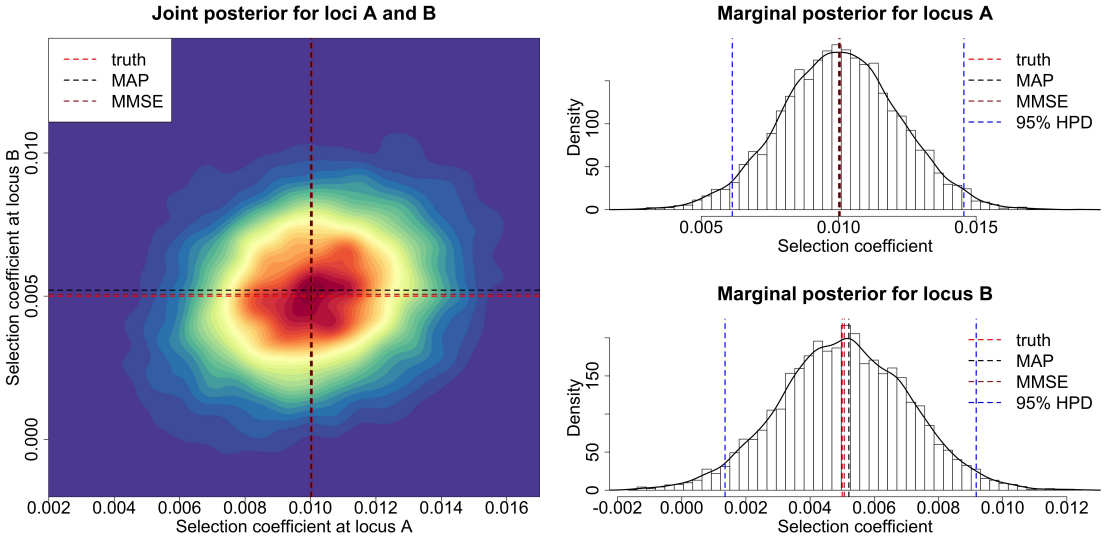
Figure 1: Simulated datasets of unphased and phased chromosomes. The filled triangles joined by lines denote the observed mutant allele frequency trajectories of the sample of unphased chromosomes or the observed haplotype frequency trajectories of the sample of phased chromosomes.

We now apply our Bayesian inference procedure, described in Section 2.2, to the simulated datasets shown in Figure 1 as an illustration. See Figure 2a for a plot of the resulting posterior surface for the selection coefficients at both loci generated from the time series data of unphased chromosomes illustrated in Figure 1a. In Figure 2b, we plot the resulting posterior surface for

selection coefficients at both loci generated from the time series data of phased chromosomes shown in Figure 1b. For each simulated dataset, we also plot the marginal posterior probability distributions for the selection coefficient at each locus with their corresponding 95% highest posterior density (HPD) intervals, respectively, in Figure 2.



(a) Posterior probability distributions for the simulated dataset of unphased chromosomes shown in Figure 1a.



(b) Posterior probability distributions for the simulated dataset of phased chromosomes shown in Figure 1b.

Figure 2: Posterior probability distributions for the simulated datasets of unphased and phased chromosomes shown in Figure 1. The red dashed lines denote the true values of the selection coefficients. The black dashed lines denote the MAP estimates of the selection coefficients, the brown dashed lines denote the MMSE estimates of the selection coefficients, and the blue dashed lines denote the 95% HPD intervals for the selection coefficients.

For these two examples, the resulting estimates of selection coefficients at both loci, using either MAP or MMSE criteria, are close to their true values. The marginal posterior probability

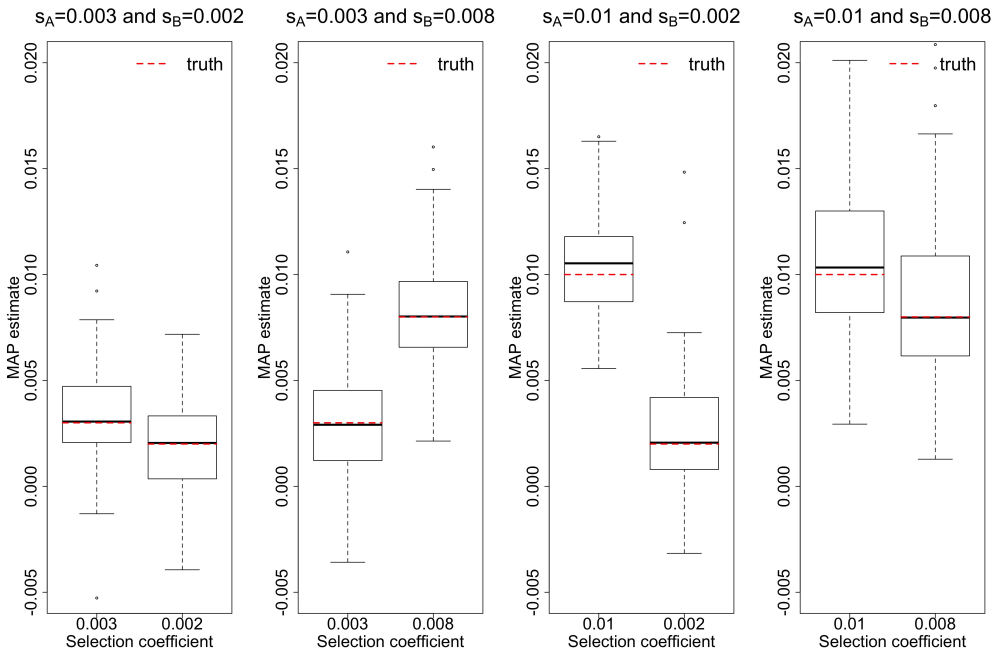
distribution for each selection coefficient shows a bell shape, and its true value is well within the corresponding 95% HPD interval. Compared to the estimates obtained from the dataset of unphased chromosomes, the estimates obtained from the dataset of phased chromosomes are close to their true values with narrower 95% HPD intervals. This is because phased chromosomes contain more information than unphased chromosomes. The complex interplay between the four haplotypes in the sample can be fully observed in the dataset of phased chromosomes but only partially observed in the dataset of unphased chromosomes.

To assess the accuracy of our Bayesian inference procedure more systematically, we run simulation studies with both phased and unphased chromosomes. We vary the selection coefficients with $s_{\mathcal{A}} \in \{0.003, 0.01\}$ and $s_{\mathcal{B}} \in \{0.002, 0.008\}$, and the recombination rate with $r \in \{0.00001, 0.01\}$, and run 100 replicates for each of the eight possible combinations of selection coefficients and recombination rate. For each replicate, we set the initial haplotype frequencies of the underlying population to $\boldsymbol{x}_0 = (0.1, 0.2, 0.3, 0.4)$ and simulate the haplotype frequency trajectories of the underlying population according to the two-locus Wright-Fisher model with selection. We draw 200 chromosomes from the underlying population at every 100 generations over the course of 500 generations. The resulting boxplots of the simulation studies with phased and unphased chromosomes are shown in Figures 3 and 4, respectively.

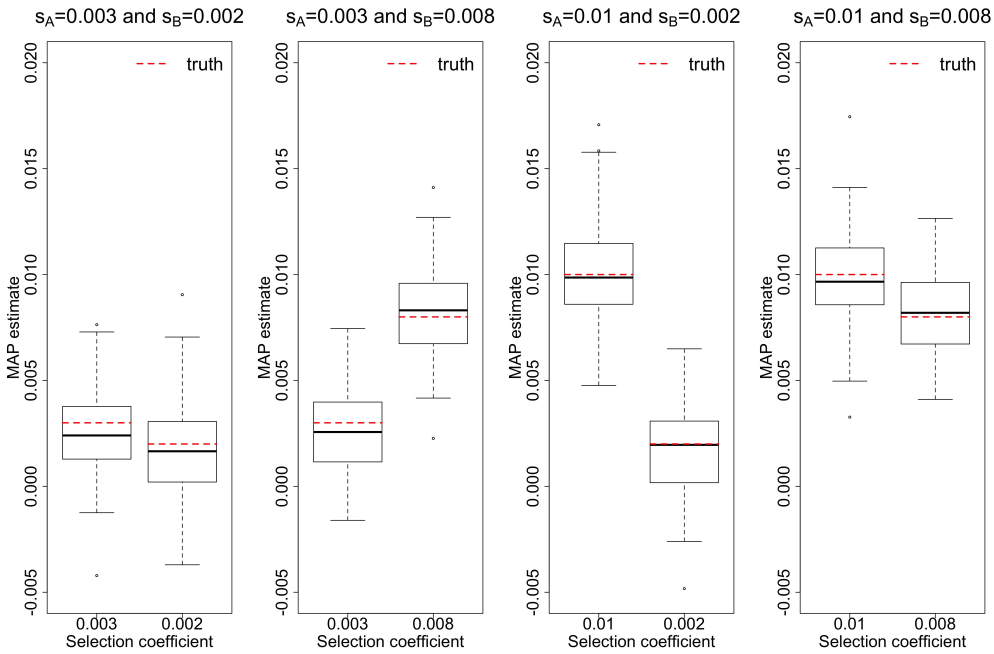
Figures 3 and 4 show that the MAP estimates for the selection coefficients at both loci have little bias across all eight combinations of selection coefficients and recombination rate we have chosen. Furthermore, as shown in S1 Table and S2 Table, MAP estimates obtained from phased data have lower variance than the same estimates obtained from unphased data. This is to be expected as all else being equal, phased chromosomes contain more information than unphased chromosomes. We have also investigated the bias of the MMSE estimates for the selection coefficients, which display very similar characteristics to the MAP estimates. The MMSE estimates show little bias, the upper and lower quantiles of the MMSE estimates are similar to those of MAP estimates. We therefore do not show them explicitly in the work.

3.2. Analysis of real data

We apply our Bayesian inference method to real data by reanalysing the time serial sample of segregating alleles of the equine homologue of proto-oncogene c-kit (KIT), which comes from a few previous studies, Ludwig et al. (2009), Pruvost et al. (2011) and Wutke et al. (2016),



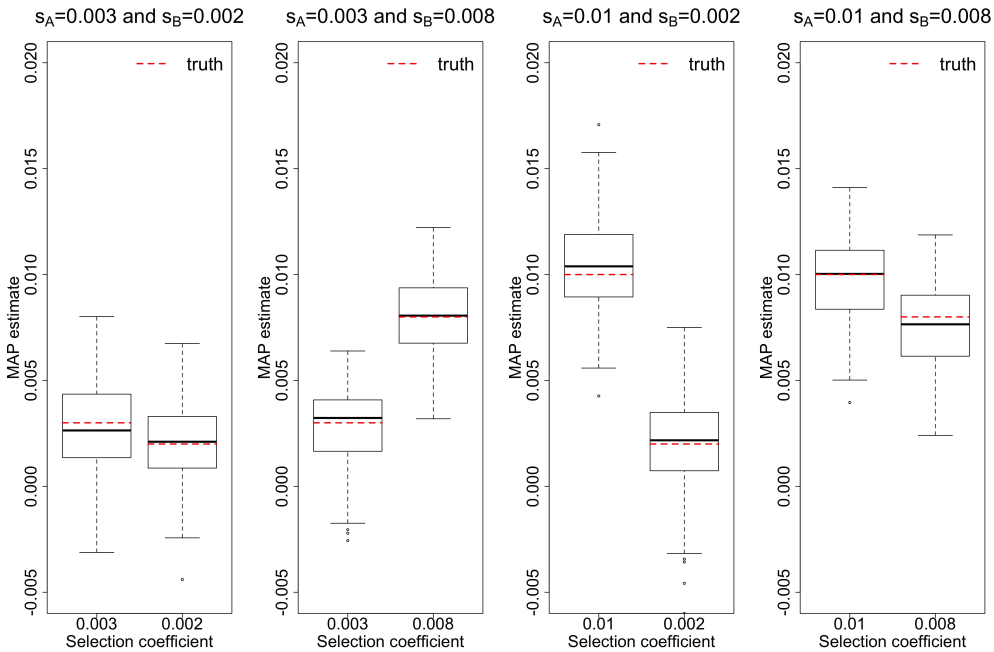
(a) MAP estimates for the selection coefficients at tightly linked loci ($r = 0.00001$).



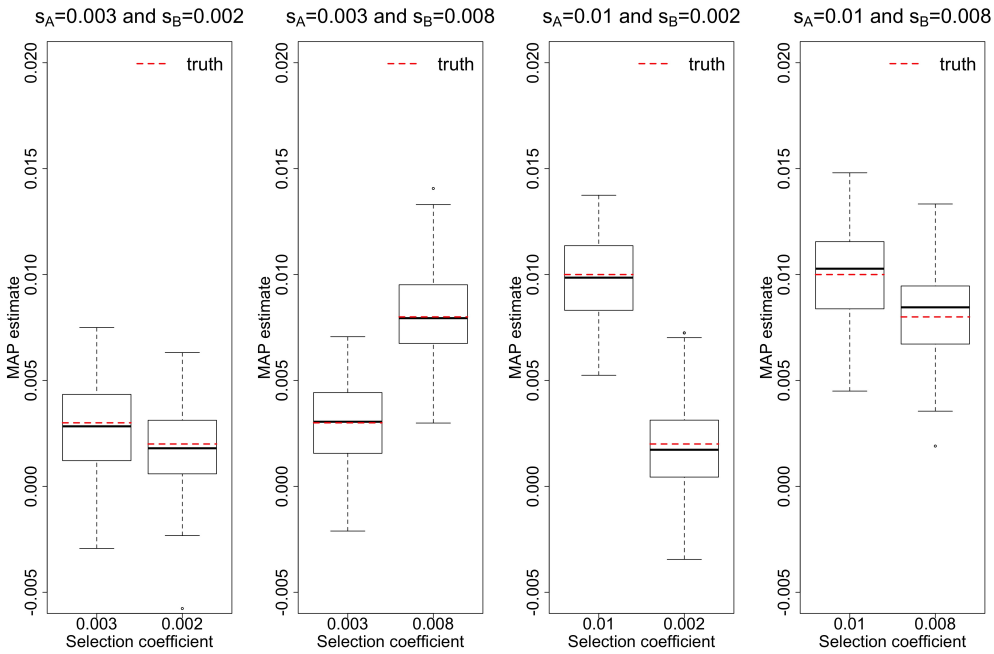
(b) MAP estimates for the selection coefficients at loosely linked loci ($r = 0.01$).

Figure 3: Empirical distributions of the MAP estimates of the selection coefficients at linked loci for 100 datasets of unphased chromosomes simulated for each combination of the selection coefficients and the recombination rate. The red dashed lines indicate the true values of selection coefficients.

and is shown in Table 1. The equine *KIT* gene 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*), and one in intron 16 which codes for sabino (*KIT16*). At the *KIT13* locus,



(a) MAP estimates for the selection coefficients at tightly linked loci ($r = 0.00001$).



(b) MAP estimates for the selection coefficients at loosely linked loci ($r = 0.01$).

Figure 4: Empirical distributions of the MAP estimates of the selection coefficients at linked loci for 100 datasets of phased chromosomes simulated for each combination of the selection coefficients and the recombination rate. The red dashed lines indicate the true values of selection coefficients.

the ancestral allele is designated *KM0*, while the mutant allele, associated with the tobiano pattern, which acts 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, which acts as semi-dominant (Brooks & Bailey, 2005), is designated *SB1*. The sabino pattern is characterised by irregularly bordered white patches 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*	unknown	*sb1*/*SB1*	unknown
-17146	22	22/0	0	22/0	0
-7029	14	14/0	0	14/0	0
-5472	48	45/3	0	46/2	2
-4442	24	24/0	0	24/0	0
-3916	28	28/0	0	28/0	0
-3352	56	53/3	0	52/4	0
-2624	30	26/4	0	24/0	6
-2330	14	11/3	0	12/0	2
-1134	100	77/3	20	86/0	14

Table 1: Time serial samples of segregating alleles at the *KIT13* and *KIT16* loci. Sampling time points are in the units of years before the present.

We set the dominance parameters to be $h = 0$ for *KIT13* since the *KM1* allele is dominant, and $h = 0.5$ for *KIT16* since the *SB1* allele is semi-dominant. Following Der Sarkissian et al. (2015), we set the population size to be $N = 16000$ and take the average length of a generation of horses to be eight years. As indicated in Table 1, there are various sampling time points when the sequencing of the aDNA material yielded a number of unknown alleles at the *KIT13* and/or *KIT16* loci. We discard the first two samples dated 17146 and 7029 years before present as neither mutant allele existed in the sample at that time. We plot the possible mutant allele frequency trajectories of the sample at loci *KIT13* and *KIT16* in Figure 5, starting from roughly 3500 BC. Indeed, according to Wutke et al. (2016), both sabino and tobiano patterns only occur in domestic horses. It can be reasonably assumed that both mutant alleles, *KM1* and *SB1*, arose after domestication of horse, which is thought to have started in the Eurasian Steppes around 3500 BC (Outram et al., 2009).

As a result of the low quality of the dataset, it is 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. Here we jointly estimate the selection coefficients for the mutant alleles at the *KIT13* and *KIT16* loci, using the method

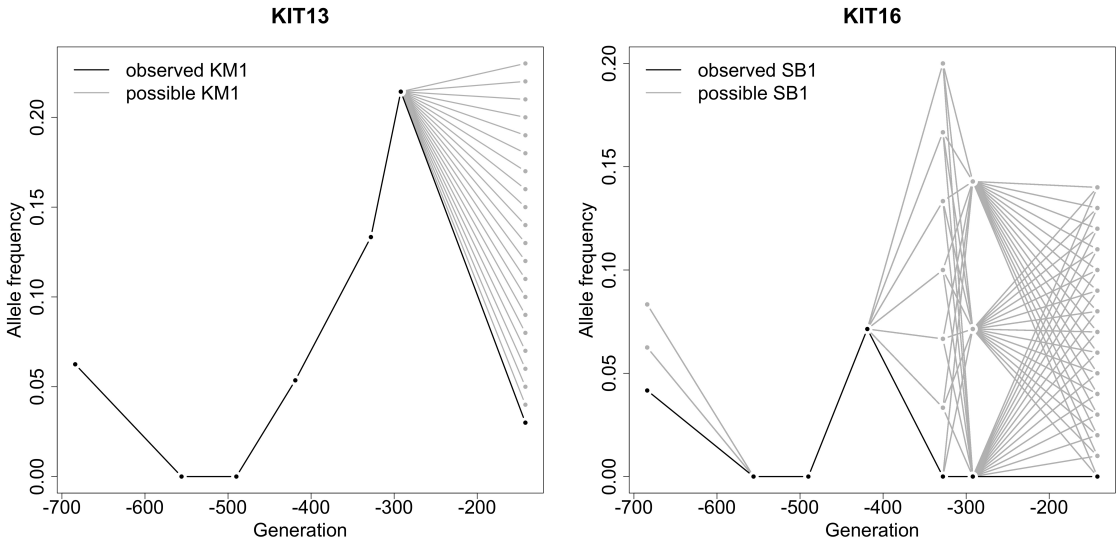


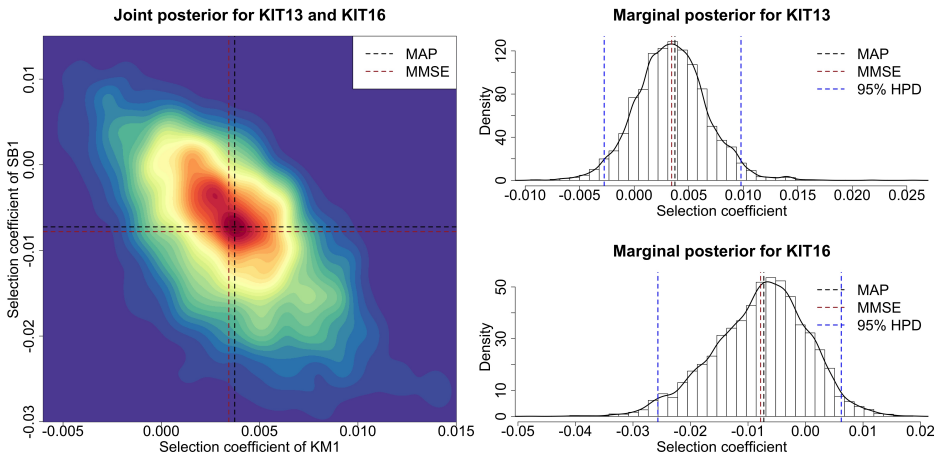
Figure 5: Changes in the mutant allele frequencies of the sample over time for the *KIT13* and *KIT16* loci. The sampling time points are at generations -684, -556, -490, -419, -328, -292 and -142. The black points joined by lines denote the observed mutant allele frequency trajectories of the sample, and the grey points joined by lines denote the possible mutant allele frequency trajectories of the sample while accounting unknown alleles in the sample.

we described in Section 2.2.2, under the case where sampled chromosomes are unphased and some of the chromosomes may have unknown alleles. Specifically, the emission probabilities are given by Eqs. (8) and (10). 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 loci *KIT13* and *KIT16*. The resulting posterior probability distributions are shown in Figure 6, and the MAP and MMSE estimates, as well as the 95% HPD intervals, are summarised in Table 2.

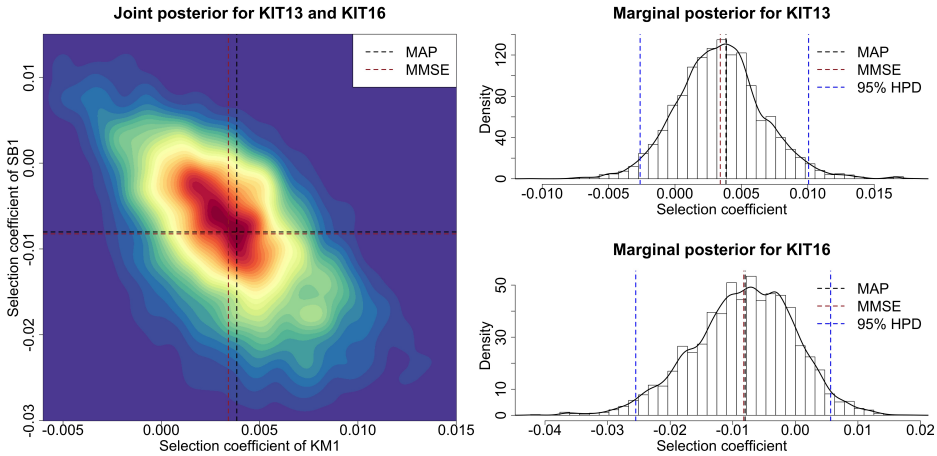
	recombination rate	MAP	MMSE	95% HPD
KIT13	2.344×10^{-5}	3.73×10^{-3}	3.43×10^{-3}	$[-2.76 \times 10^{-3}, 9.79 \times 10^{-3}]$
	4.688×10^{-5}	3.84×10^{-3}	3.41×10^{-3}	$[-2.63 \times 10^{-3}, 1.01 \times 10^{-2}]$
	2.344×10^{-4}	2.93×10^{-3}	3.63×10^{-3}	$[-2.82 \times 10^{-3}, 1.02 \times 10^{-2}]$
KIT16	2.344×10^{-5}	-7.25×10^{-3}	-7.81×10^{-3}	$[-2.57 \times 10^{-2}, 6.21 \times 10^{-3}]$
	4.688×10^{-5}	-8.02×10^{-3}	-8.24×10^{-3}	$[-2.55 \times 10^{-2}, 5.65 \times 10^{-3}]$
	2.344×10^{-4}	-4.82×10^{-3}	-8.46×10^{-3}	$[-2.39 \times 10^{-2}, 7.17 \times 10^{-3}]$

Table 2: The MAP and MMSE estimates, as well as the 95% HPD intervals, of the selection coefficients for the mutant alleles at the *KIT13* and *KIT16* loci.

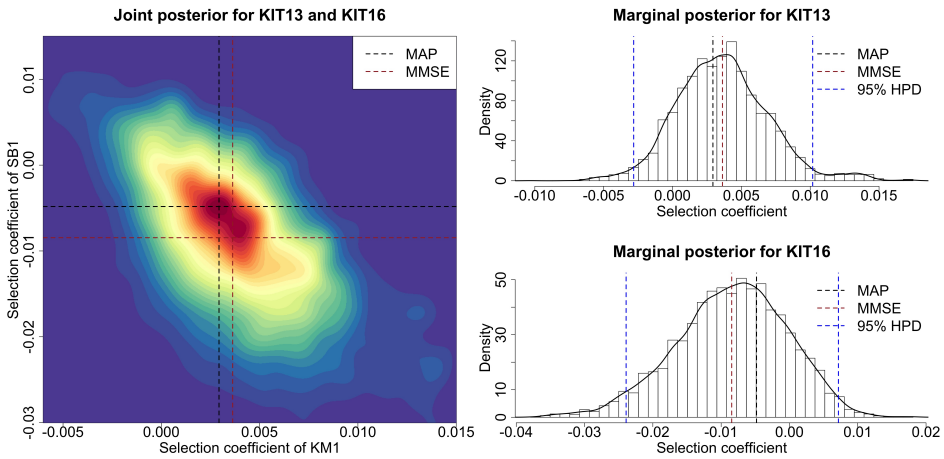
As can be seen in these results, the MMSE estimates with different choices of the recombination rate are essentially unchanged, while the MAP estimates vary a bit more than the MMSE



(a) Posterior probability distributions with the average rate of recombination 5×10^{-9} crossovers/bp.



(b) Posterior probability distributions with the average rate of recombination 1×10^{-8} crossovers/bp.



(c) Posterior probability distributions with the average rate of recombination 5×10^{-8} crossovers/bp.

Figure 6: Posterior probability distributions for *KIT13* and *KIT16*. The population size is set to be 16000, and the average rate of recombination is chosen to be 5×10^{-9}, 1×10^{-8} and 5×10^{-8} crossovers/bp, respectively. The black dashed lines denote the MAP estimates of the selection coefficients, the brown dashed lines denote the MMSE estimates of the selection coefficients, and the blue dashed lines denote the 95% HPD intervals for the selection coefficients.

estimates. The 95% HPD intervals for both selection coefficients include 0, thereby offering no strong evidence that either the *KM1* allele at the *KIT13* locus or the *SB1* allele at the *KIT16* locus is selected. However, as can be seen from the MAP and MMSE estimates of the selection coefficients, there is weak evidence to support the *KM1* allele at the *KIT13* locus being weakly positively selected and the *SB1* allele at the *KIT16* locus being weakly negatively selected. This conclusion is further confirmed with the estimates using different values of the population size ($N = 8000$ and $N = 32000$) as can be seen in S1 Figure and S2 Figure. Our findings may help to explain the phenomenon that the tobiano pattern is widespread in horses today, whereas the sabino pattern is found only in North American horses (Ransom & Kaczensky, 2016).

3.3. Computational issues

It is desirable to generate a large number of particles in the bootstrap particle filter to estimate the marginal likelihood $p(\boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K} \mid \boldsymbol{\vartheta})$ in the PMMH algorithm. However, this can become computational burdensome as each iteration of the PMMH algorithm requires a run of the bootstrap particle filter. We refer to Andrieu & Roberts (2009) for a discussion of the relationship between the computational accuracy of the estimator of the marginal likelihood and the computational efficiency of the pseudo-marginal approach. Nevertheless, the PMMH algorithm exactly targets the marginal posterior probability distribution $p(\boldsymbol{\vartheta} \mid \boldsymbol{c}_{1:K}, \boldsymbol{m}_{1:K})$ for any number of particles. Furthermore, 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 sufficiently large L in the Euler-Maruyama scheme to achieve an accurate approximation of the Wright-Fisher diffusion, but a 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 10 subintervals in the Euler-Maruyama scheme. The running time for a single iteration of the PMMH algorithm with 1000 particles, on a single core of an Intel Core i7 processor at 4.2GHz, is around 2.683 seconds for the unphased dataset shown in Figure 1a and is around 2.096 seconds for the phased dataset shown in Figure 1b. In principle, each particle can be simulated in parallel on a different core. Running 10000 iterations of the PMMH algorithm is sufficient for a relatively smooth resulting posterior surface plot. We

discard the first 2000 iterations of the PMMH output as burn-in and then thin the remaining PMMH output, taking every fifth observation and regarding these as independent.

Exact-approximate particle filtering approaches such as the PMMH algorithm we use in this work seem very useful for inference of population genetic quantities in time series genetic data since this methodology can be generalised to a range of scenarios. Although computationally demanding, improvements to the PMMH continue to be developed (*e.g.*, Yıldırım et al., 2018).

4. Discussion

In the present work, we have developed a novel MCMC-based method to analyse time series genetic data across linked loci under natural selection while accounting for genetic recombination. Our Bayesian inference procedure relies on an HMM framework incorporating the two-locus Wright-Fisher diffusion with selection, and the Bayesian estimates for selection coefficients are achieved through the PMMH algorithm. We have demonstrated that our approach can be used to accurately estimate selection coefficients on simulated data, regardless of whether sampled chromosomes are phased or not. We have applied our approach to two linked *KIT* loci involved in horse white spotting patterns and found some evidence that the tobiano pattern caused by the *KM1* allele at locus *KIT13* is weakly advantageous while the sabino pattern caused by the *SB1* allele at locus *KIT16* is weakly deleterious. Such an application also shows that our method is able to handle unknown alleles in the sample, which is common in aDNA data.

In the case where the two loci are both suspected to be favoured by natural selection, one could still apply a single-locus approach to each locus to estimate the selection coefficient, and there has been a considerable amount of work on the statistical inference of natural selection from time series data of allele frequencies at a single locus (*e.g.*, Bollback et al., 2008; Malaspinas et al., 2012; Steinrücken et al., 2014; Schraiber et al., 2016). However, applying a single-locus approach may lead to inaccurate estimates of selection coefficients when the two loci are close in terms of recombination rate (He et al., 2017a). In this case of tightly linked loci, accounting for genetic recombination becomes necessary, and our two-locus method is far more desirable. We illustrate this by the following two examples, where we set dominance parameters $h_{\mathcal{A}} = 0.5$ and $h_{\mathcal{B}} = 0.5$, recombination rate $r = 0.00001$ and population size $N = 5000$. We simulate the latent haplotype frequency trajectories of the underlying population through the two-locus

Wright-Fisher model with selection and draw 200 unphased chromosomes from the underlying population at generations 0, 100, 200, 300, 400 and 500.

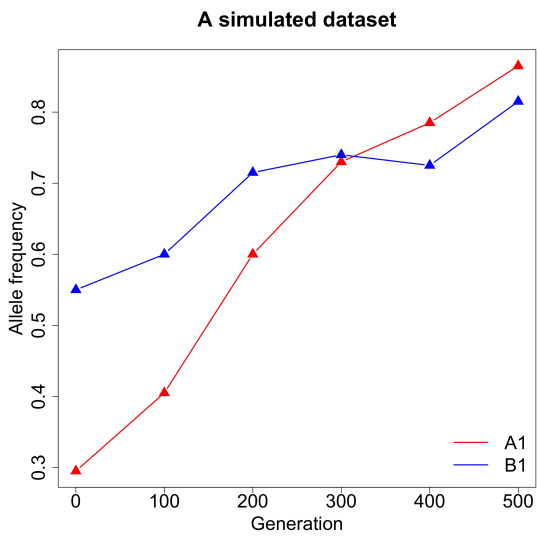
In the first example, we consider a positively selected locus $\mathcal{A}$ tightly linked with a neutral locus $\mathcal{B}$, where we take the selection coefficients to be $s_{\mathcal{A}} = 0.01$ and $s_{\mathcal{B}} = 0$, respectively, and set the initial haplotype frequencies of the underlying population to be $\boldsymbol{x}_0 = (0.2, 0.1, 0.3, 0.4)$. The mutant allele frequency trajectories of the sample are illustrated in Figure 7a. The posterior probability distributions obtained with our single-locus approach, described in S1 Appendix, are shown in Figure 7b, and the posterior probability distributions estimated with our two-locus method, described in Section 2.2, are shown in Figure 7c. These Bayesian estimates of the selection coefficients $s_{\mathcal{A}}$ and $s_{\mathcal{B}}$ are summarised in Table 3.

		single-locus method	two-locus method
selection coefficient $s_{\mathcal{A}}$	MAP	1.10×10^{-2}	8.18×10^{-3}
	MMSE	1.12×10^{-2}	9.71×10^{-3}
	95% HPD	$[6.99 \times 10^{-3}, 1.55 \times 10^{-2}]$	$[3.96 \times 10^{-3}, 1.56 \times 10^{-2}]$
selection coefficient $s_{\mathcal{B}}$	MAP	5.07×10^{-3}	1.94×10^{-3}
	MMSE	4.96×10^{-3}	1.53×10^{-3}
	95% HPD	$[6.49 \times 10^{-4}, 9.09 \times 10^{-3}]$	$[-8.09 \times 10^{-3}, 1.04 \times 10^{-2}]$

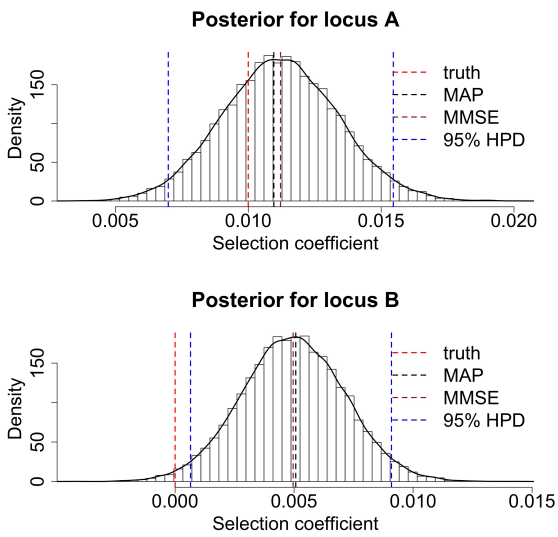
Table 3: Comparison of Bayesian estimates obtained from the simulated dataset in Figure 7a with a single-locus method vs. a two-locus method.

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 selection coefficients $s_{\mathcal{A}}$ and $s_{\mathcal{B}}$ are fairly accurate with the two-locus method.

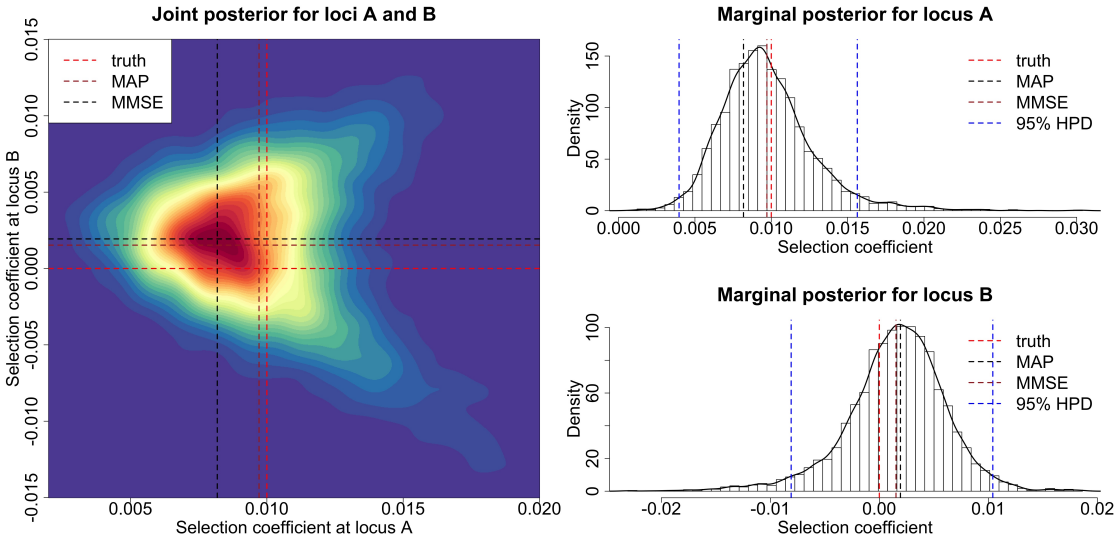
To understand the poor performance of the single-locus method in this example, we plotted the mutant allele frequency trajectories of the underlying population in Figure 8a and the haplotype frequency trajectories of the underlying population in Figure 8b. The increase in the frequency of the $\mathcal{B}_1$ 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 can compensate 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 on 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



(a) A simulated dataset.



(b) Posteriors obtained with a single-locus method.

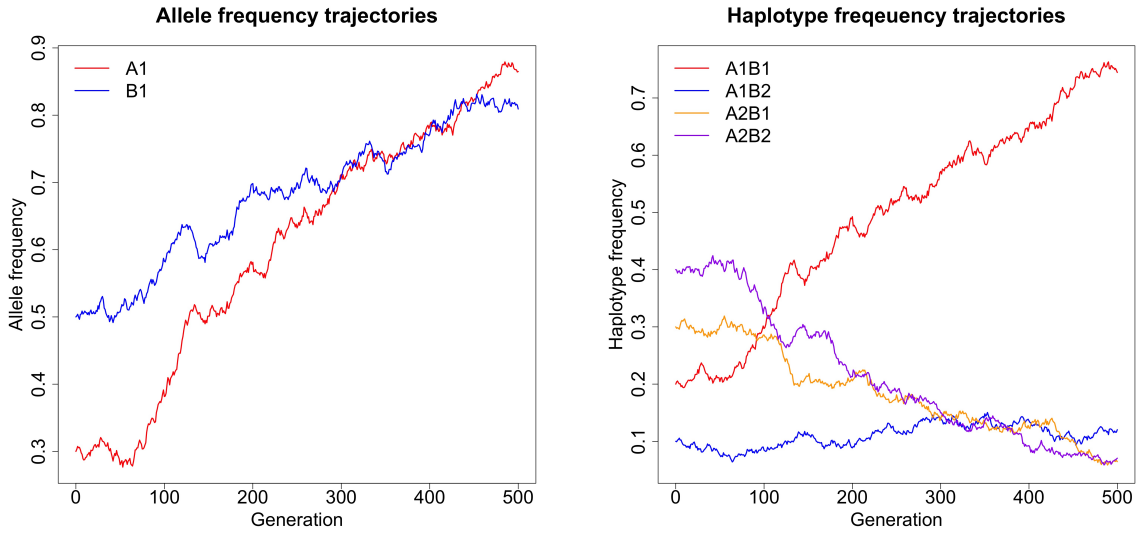


(c) Posteriors obtained with a two-locus method.

Figure 7: Posterior probability distributions for the simulated dataset of a selected locus tightly linked with a neutral locus. The red dashed lines denote the true values of the selection coefficients. The black dashed lines denote the MAP estimates of the selection coefficients, the brown dashed lines denote the MMSE estimates of the selection coefficients, and the blue dashed lines denote the 95% HPD intervals for the selection coefficients.

for both selection coefficients.

In the second example, we consider two tightly linked loci $\mathcal{A}$ and $\mathcal{B}$ under natural selection, 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 shown in Figure 9 and summarised in Table 4. In this example, with the single-locus method, the estimate for the selection coefficient $s_{\mathcal{A}}$ is reasonably accurate, but the estimate



(a) Mutant allele frequency trajectories

(b) Haplotype frequency trajectories

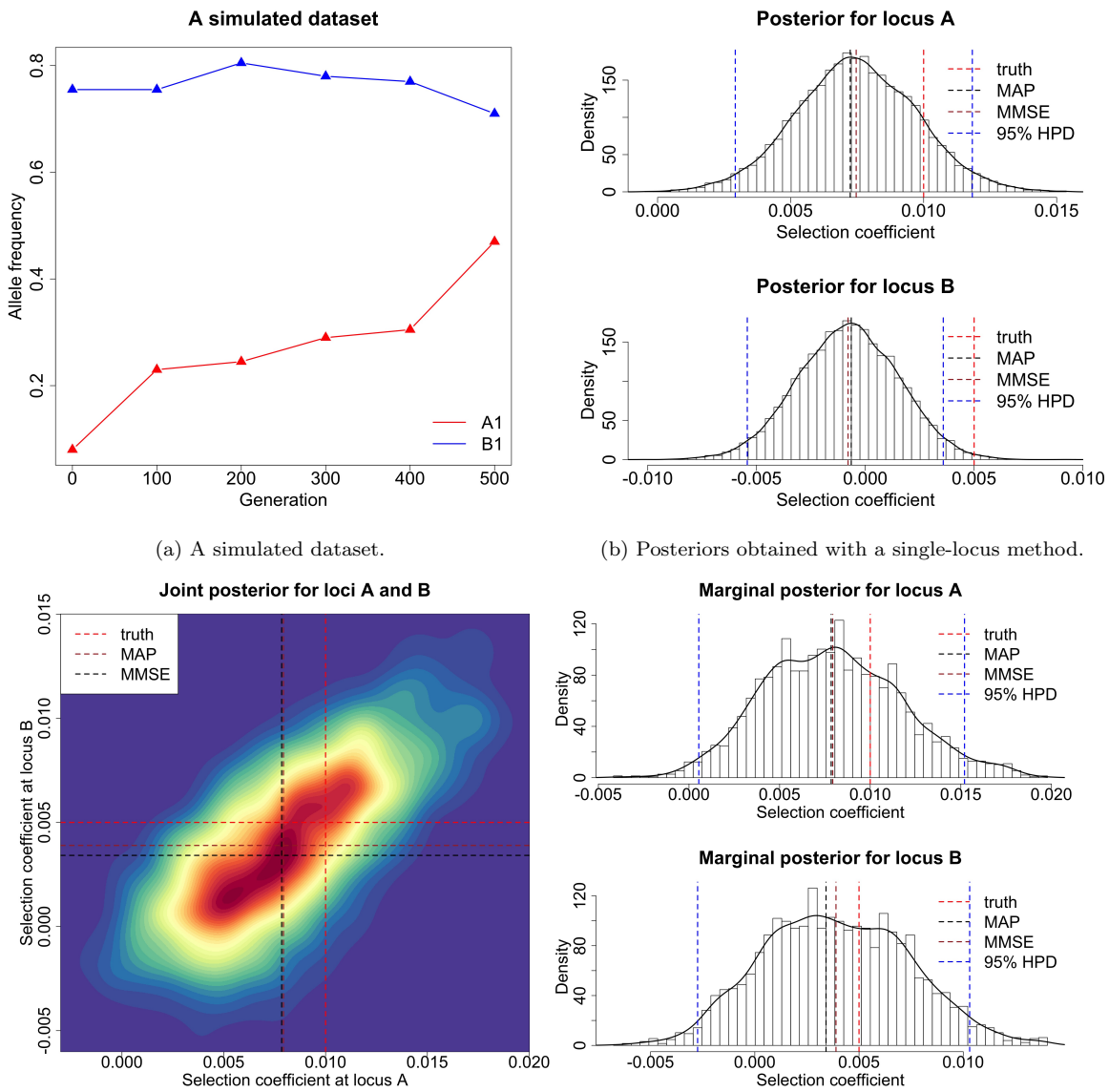
Figure 8: Mutant allele frequency trajectories and haplotype frequency trajectories of the underlying population for the simulated dataset in Figure 7a.

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, even though the $\mathcal{B}_1$ allele is positively selected 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 yields no strong evidence for natural selection. In comparison, the two-locus approach 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	7.24×10^{-3}	7.84×10^{-3}
	MMSE	7.46×10^{-3}	7.93×10^{-3}
	95% HPD	$[2.92 \times 10^{-3}, 1.18 \times 10^{-2}]$	$[5.41 \times 10^{-4}, 1.52 \times 10^{-2}]$
selection coefficient $s_{\mathcal{B}}$	MAP	-6.47×10^{-4}	3.42×10^{-3}
	MMSE	-7.94×10^{-4}	3.89×10^{-3}
	95% HPD	$[-5.43 \times 10^{-3}, 3.58 \times 10^{-3}]$	$[-2.74 \times 10^{-3}, 1.03 \times 10^{-2}]$

Table 4: Comparison of Bayesian estimates obtained from the simulated dataset in Figure 9a with a single-locus method vs. a two-locus method.

We plotted the mutant allele frequency trajectories of the underlying population in Figure 10a and the haplotype frequency trajectories of the underlying population in Figure 10b. The frequency of the $\mathcal{A}_1$ allele increases with time due to the increase in the frequencies of both $\mathcal{A}_1\mathcal{B}_1$ and $\mathcal{A}_1\mathcal{B}_2$ haplotypes, since they are the two most selected haplotypes, with selection coefficient 0.015 and 0.01, respectively. The $\mathcal{B}_1$ allele is made up of the $\mathcal{A}_1\mathcal{B}_1$ and $\mathcal{A}_2\mathcal{B}_1$ hap-



(a) A simulated dataset.

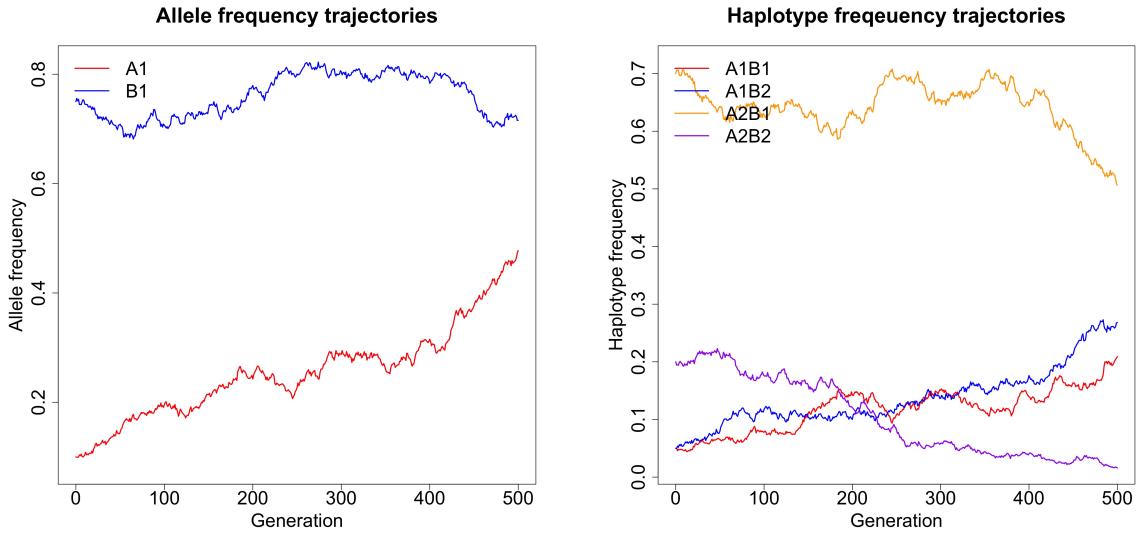
(b) Posteriors obtained with a single-locus method.

(c) Posteriors obtained with a two-locus method.

Figure 9: Posterior probability distributions for the simulated dataset of two tightly linked loci under natural selection. The red dashed lines denote the true values of the selection coefficients. The black dashed lines denote the MAP estimates of the selection coefficients, the brown dashed lines denote the MMSE estimates of the selection coefficients, and the blue dashed lines denote the 95% HPD intervals for the selection coefficients.

lotypes, with selection coefficients 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 a 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. With the two-locus approach, we are again able to obtain accurate estimates for both of the selection coefficients $s_{\mathcal{A}}$ and $s_{\mathcal{B}}$.



(a) Mutant allele frequency trajectories

(b) Haplotype frequency trajectories

Figure 10: Mutant allele frequency trajectories and haplotype frequency trajectories of the underlying population for the simulated dataset in Figure 9a.

Compared to existing single-locus approaches, our two-locus method accounts for the process of genetic recombination and the complex interplay between the four haplotypes. The dynamics of two-locus Wright-Fisher diffusion with selection can demonstrate complex and unpredictable behaviour (see, *e.g.*, Yu & Etheridge, 2010; Cuthbertson et al., 2012). We have demonstrated that this can lead to incorrect estimates of selection coefficients if one simply applies a single-locus approach. In contrast, applying our two-locus method yields accurate estimates of both selection coefficients.

In addition to our two-locus 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 characterised by 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. Such a moment-based approximation aims to fit a mathematically convenient transition probability density function by equating the first moment of the Wright-Fisher model, which can be problematic as natural selection is expected to

drive allele frequencies towards the boundaries faster than genetic drift. Moreover, in practice, the method 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 the present work makes few approximations. We use the diffusion limit to approximate the Wright-Fisher model of the stochastic evolutionary dynamics under natural selection at two linked loci. Since our procedure is based on simulating trajectories of the Wright-Fisher diffusion, we can handle natural selection of almost arbitrary types, whether dominant, semi-dominant or recessive, as well as possible loss and fixation of alleles. This also allows us to handle different types of DNA data, *e.g.*, aDNA data. In addition, we model both phased and unphased chromosomes in our approach with the additional flexibility of modelling potential missing values in data. Our procedure can also handle the Pool-Seq time series data from E&R experiments, as in Terhorst et al. (2015). In the future, we aim to find statistical approaches that can handle data from more than two linked loci. The challenge is that as the number of loci increases, modelling the underlying population dynamics 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 good approximations in order for the method to be computationally feasible.

Supporting information

S1 Table. **RMSE of the MAP estimator for the selection coefficients with unphased chromosomes.** RMSE of the MAP estimator for the selection coefficients at two linked loci for 120 datasets of unphased chromosomes simulated for each combination of selection coefficients and recombination rate.

S2 Table. **RMSE of the MAP estimator for the selection coefficients with phased chromosomes.** MRMSE of the MAP estimator for the selection coefficients at two linked loci for 120 datasets of phased chromosomes simulated for each combination of selection coefficients and recombination rate.

S1 Figure. **Posterior probability distributions for *KIT13* and *KIT16* with the population size $N = 8000$.** Posterior probability distributions for *KIT13* and *KIT16*. The

population size is set to be 8000, and the average rate of recombination is chosen to be 5×10^{-9}, 1×10^{-8} and 5×10^{-8} crossovers/bp, respectively. The black dashed lines denote the MAP estimates of the selection coefficients, the brown dashed lines denote the MMSE estimates of the selection coefficients, and the blue dashed lines denote the 95% HPD regions for the selection coefficients.

S2 Figure. **Posterior probability distributions for *KIT13* and *KIT16* with the population size $N = 32000$.** Posterior probability distributions for *KIT13* and *KIT16*. The population size is set to be 32000, and the average rate of recombination is chosen to be 5×10^{-9}, 1×10^{-8} and 5×10^{-8} crossovers/bp, respectively. The black dashed lines denote the MAP estimates of the selection coefficients, the brown dashed lines denote the MMSE estimates of the selection coefficients, and the blue dashed lines denote the 95% HPD regions for the selection coefficients.

S1 Appendix. **An MCMC-based method for Bayesian inference of natural selection from time series DNA data at a single locus.**

Acknowledgements

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

References

Allentoft, M. E., Sikora, M., Sjogren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P. B., Schroeder, H., Ahlstrom, T., Vinner, L., Malaspinas, A.-S., Margaryan, A., Higham, T., Chivall, D., Lynnerup, N., Harvig, L., Baron, J., Casa, P. D., Dabrowski, P., Duffy, P. R., Ebel, A. V., Epimakhov, A., Frei, K., Furmanek, M., Gralak, T., Gromov, A., Gronkiewicz, S., Grupe, G., Hajdu, T., Jarysz, R., Khartanovich, V., Khokhlov, A., Kiss, V., Kolar, J., Kriiska, A., Lasak, I., Longhi, C., McGlynn, G., Merkevicius, A., Merkyte, I., Metspalu, M., Mkrtchyan, R., Moiseyev, V., Paja, L., Palfi, G., Pokutta, D., Pospieszny, L., Price, T. D., Saag, L., Sablin, M., Shishlina, N., Smrcka, V., Soenov, V. I., Szeverenyi, V., Toth, G., Trifanova, S. V., Varul, L., Vicze, M., Yepiskoposyan, L., Zhitenev, V., Orlando,

L., Sicheritz-Ponten, T., Brunak, S., Nielsen, R., Kristiansen, K., & Willerslev, E. (2015). Population genomics of Bronze Age Eurasia. *Nature*, *522*, 167–172.

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.

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 Przewalskis horse. *Current Biology*, *25*, 2577–2583.

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

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

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

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. (2017a). A numerical solution of the Wright-Fisher stochastic differential equation with application to transition probability density approximation. Manuscript submitted for publication.

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

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.

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.

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.

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.

Ransom, J. I., & Kaczensky, P. (2016). *Wild Equids: Ecology, Management, and Conservation*. Baltimore: Johns Hopkins University Press.

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

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.

Taniguchi, T. (1992). Successive approximations to solutions of stochastic differential equations. *Journal of Differential Equations*, *96*, 152–169.

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.

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.

Yamada, T., & Watanabe, S. (1971). On the uniqueness of solutions of stochastic differential equations. *Journal of Mathematics of Kyoto University*, *11*, 155–167.

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.