



He, Z., Dai, X., Beaumont, M., & Yu, F. (2020). Detecting and quantifying natural selection at two linked loci from time series data of allele frequencies with forward-in-time simulations. *Genetics*, 216(2), 521-541. <https://doi.org/10.1534/genetics.120.303463>

Early version, also known as pre-print

Link to published version (if available):
[10.1534/genetics.120.303463](https://doi.org/10.1534/genetics.120.303463)

[Link to publication record in Explore Bristol Research](#)
PDF-document

This is the submitted manuscript (SM). The accepted author manuscript version (AAM) will be made available online via Genetics Society of America. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
<http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

Detecting and quantifying natural selection at two linked loci from time series data of allele frequencies with forward-in-time simulations

Zhangyi He^{a,1,*}, Xiaoyang Dai^b, Mark Beaumont^b, Feng Yu^{a,*}

^a*School of Mathematics, University of Bristol, Bristol BS8 1UG, United Kingdom*

^b*School of Biological Sciences, University of Bristol, Bristol BS8 1TQ, United Kingdom*

Abstract

Recent advances in DNA sequencing techniques have made it possible to monitor genomes in great detail over time. This improvement provides an opportunity for us to study natural selection based on time serial samples of genomes while accounting for genetic recombination effect and local linkage information. Such time series genomic data allow for more accurate estimation 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 where the underlying process is a two-locus Wright-Fisher diffusion with selection, which enables us to explicitly model genetic recombination and local linkage. The posterior probability distribution for selection coefficients is computed by applying the particle marginal Metropolis-Hastings algorithm, which allows us to efficiently calculate the likelihood. We evaluate the performance of our Bayesian inference procedure through extensive simulations, showing that our approach can deliver accurate estimates of selection coefficients, and the addition of genetic recombination and local linkage brings about significant improvement in the inference of natural selection. We also illustrate the utility of our method on real data with an application to ancient DNA data associated with white spotting patterns in horses.

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

*Corresponding author.

Email addresses: z.he@imperial.ac.uk (Zhangyi He), feng.yu@bristol.ac.uk (Feng Yu)

¹Present address: Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London W2 1PG, United Kingdom

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 that requires unrealistic assumptions of ancestral demography and selective regimes (see Bank et al., 2014, for a review). With the advances in DNA sequencing technologies, an ever-increasing amount of allele frequency data sampled at multiple time points are becoming available. Such time series genetic data can arise from experimental evolution (*e.g.*, Burke et al., 2010; Orozco-terWengel et al., 2012; Lang et al., 2013; Wisser 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). Temporally spaced samples 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 allele frequency 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; Paris et al., 2019; He et al., 2019), reviewed in Bank et al. (2014) and Malaspinas (2016). A common method to analysing time series allele frequency data is based on the hidden Markov model (HMM) framework of Williamson & Slatkin (1999), where the underlying population is assumed to evolve under 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 review of the statistical inference in the Wright-Fisher model using allele frequency

31 data). However, such approaches can become computationally infeasible for large populations
32 because they require a prohibitively large amount of computation and storage for the calculation
33 of the likelihood. Most existing HMM-based methods are therefore built on either the diffusion
34 approximation of the Wright-Fisher model (*e.g.*, Bollback et al., 2008; Malaspinas et al., 2012;
35 Steinrücken et al., 2014; Schraiber et al., 2016; Ferrer-Admetlla et al., 2016; He et al., 2019) or
36 the moment-based approximation of the Wright-Fisher model (*e.g.*, Lacerda & Seoighe, 2014;
37 Feder et al., 2014; Terhorst et al., 2015; Paris et al., 2019). Such approximations enable efficient
38 integration over all possible allele frequency trajectories of the underlying population, thereby
39 allowing the likelihood computation to be completed in a reasonable amount of time.

40 The recent advent of high-throughput sequencing technologies has made it possible to monitor
41 genomes in great detail over time. This provides an opportunity for detecting and estimating
42 natural selection at multiple linked loci from time series data of allele frequencies while taking
43 the process of genetic recombination and the information of local linkage into account. Properly
44 accounting for genetic recombination and local linkage can be expected to provide more precise
45 estimates for the selection coefficient and more accurate hypothesis testing on the recent action
46 of natural selection since genetic recombination may either reinforce or oppose changes in allele
47 frequencies caused by natural selection according to the levels of linkage disequilibrium (He
48 et al., 2020). However, with the exception of Terhorst et al. (2015), all existing methods built on
49 the Wright-Fisher model for inferring natural selection from time series allele frequency data are
50 limited to either a single locus (*e.g.*, Bollback et al., 2008; Malaspinas et al., 2012; Steinrücken
51 et al., 2014; Schraiber et al., 2016; He et al., 2019) or multiple independent loci (*e.g.*, Foll et al.,
52 2014, 2015; Shim et al., 2016; Ferrer-Admetlla et al., 2016; Paris et al., 2019), *i.e.*, genetic re-
53 combination effect and local linkage information are ignored in these approaches. The exception
54 amongst these methods, Terhorst et al. (2015), extended a moment-based approximation of the
55 Wright-Fisher model introduced by Feder et al. (2014) to the case of multiple linked loci with
56 an application to the pooled sequencing (Pool-Seq) data from evolve-and-resequence (E&R) ex-
57 periments, where the allele frequency transition between two consecutive sampling time points
58 is modelled deterministically, with added Gaussian noise.

59 In the present work, we propose a novel HMM-based method for Bayesian inference of natural
60 selection at two linked loci from time series data of allele frequencies while accounting for the

61 process of genetic recombination, thereby incorporating the information on local linkage. Our
62 key innovation is that a diffusion approximation to the Wright-Fisher model of the stochastic
63 evolutionary dynamics under natural selection at two linked loci is used as the hidden Markov
64 process to characterise the changes in the haplotype frequencies of the underlying population
65 over time, which enables us to explicitly model genetic recombination and local linkage. The
66 diffusion approximation we use in our approach allows us to avoid the restriction that the allele
67 frequency trajectory of the underlying population remains far away from allele fixation or loss,
68 which was imposed by the Gaussian approximation used in Terhorst et al. (2015). Our posterior
69 computation is carried out with the particle marginal Metropolis-Hastings (PMMH) algorithm
70 developed by Andrieu et al. (2010), which enables us to efficiently calculate the likelihood. Also,
71 our method can handle sampled chromosomes with unknown alleles, which is common in aDNA
72 data due to postmortem damage. In addition, our method can be readily extended to model a
73 range of complex evolutionary scenarios like non-constant demographic histories.

74 We evaluate the performance of our Bayesian inference procedure through extensive simula-
75 tions. We show that our method enables efficient and accurate estimation of selection coefficients
76 from time series genetic data, regardless of whether sampled chromosomes contain unknown al-
77 leles or not. We present two scenarios where existing single-locus methods fail to deliver precise
78 estimates for selection coefficients whereas our approach still works well, especially when the
79 loci are tightly linked. This shows the efficacy of our method in modelling genetic recombina-
80 tion and local linkage. Finally, we apply our Bayesian inference procedure to analyse the aDNA
81 data associated with white spotting patterns in horses from Wutke et al. (2016) and find that
82 in horses there is no evidence showing that the tobiano pattern is positively selected but strong
83 evidence of the sabino pattern being negatively selected.

84 **2. Materials and Methods**

85 In this section, we begin with a short review of the Wright-Fisher diffusion for two linked loci
86 evolving subject to natural selection over time, and then describe our Bayesian procedure for the
87 inference of natural selection at the two linked loci from temporally spaced samples, *e.g.*, how to
88 set up the HMM framework and how to compute the posterior probability distribution for the
89 population genetic quantities of interest with Markov chain Monte Carlo (MCMC) techniques.

90 *2.1. Wright-Fisher diffusion*

91 Consider a diploid population of randomly mating individuals at two linked loci \mathcal{A} and \mathcal{B}
92 evolving under natural selection according to the two-locus Wright-Fisher model with selection
93 (see, *e.g.*, He et al., 2017), for which we assume discrete time and nonoverlapping generations.
94 At each locus, there are two possible allele types, labelled \mathcal{A}_1 , \mathcal{A}_2 and \mathcal{B}_1 , \mathcal{B}_2 , respectively,
95 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
96 and 4, respectively. We attach symbols \mathcal{A}_1 and \mathcal{B}_1 to the mutant alleles, which are assumed to
97 arise only once in the population and be positively selected once it exists, and we attach symbols
98 \mathcal{A}_2 and \mathcal{B}_2 to the ancestral alleles, which are assumed to originally exist in the population.

99 We incorporate viability selection into the population dynamics and assume that the viability
100 is fixed from the time that the mutant allele arises in the population and is only determined by
101 the genotype at each single locus. More specifically, we assume that the relative viabilities of
102 the sixteen possible (ordered) genotypes at the two loci are determined multiplicatively from the
103 relative viabilities at individual loci, and the relative viabilities of the three possible genotypes
104 at each single 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
105 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 $h_{\mathcal{A}} \in [0, 1]$
106 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
107 $(1 - h_{\mathcal{A}}s_{\mathcal{A}})(1 - s_{\mathcal{B}})$. We let $r \in [0, 0.5]$ be the recombination rate of the two loci on the same
108 chromosome (*i.e.*, the fraction of recombinant offspring showing a crossover between the two
109 loci). We assume that the population size is fixed to be N diploid individuals for all time.

110 *2.1.1. Two-locus Wright-Fisher diffusion with selection*

111 We consider a scaling limit of the Wright-Fisher model, where the unit of time is rescaled by
112 $2N$. The scaled selection coefficients $\alpha_{\mathcal{A}} = 2Ns_{\mathcal{A}}$ and $\alpha_{\mathcal{B}} = 2Ns_{\mathcal{B}}$, and the scaled recombination
113 rate $\rho = 4Nr$ are kept constant while the population size N is taken to infinity. As the popu-
114 lation size approaches infinity, the haplotype frequency trajectories follow a standard diffusion
115 limit of the two-locus Wright-Fisher model with selection (see, *e.g.*, He et al., 2020), called the
116 two-locus Wright-Fisher diffusion with selection. The Wright-Fisher diffusion has already been
117 successfully applied in the inference of natural selection from allele frequency time series data.
118 The partial differential equation (PDE) satisfied by the transition probability density function
119 of the Wright-Fisher diffusion was used in *e.g.*, Bollback et al. (2008); Gutenkunst et al. (2009);

120 Steinrücken et al. (2014); He et al. (2019). In this work, as used in *e.g.*, Schraiber et al. (2016),
 121 we characterise the Wright-Fisher diffusion as the solution of the stochastic differential equation
 122 (SDE) instead.

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

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

126 and satisfies the SDE in the following form

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

127 with initial condition $\mathbf{X}(t_0) = \mathbf{x}_0$. In Eq. (1), the drift term $\boldsymbol{\mu}(\mathbf{x})$ is a four-dimensional vector
 128 being

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

129 the diffusion term $\boldsymbol{\nu}(\mathbf{x})$ is a four by three matrix satisfying

$$\boldsymbol{\nu}(\mathbf{x})\boldsymbol{\nu}(\mathbf{x})^\top = \boldsymbol{\Sigma}(\mathbf{x}) = \begin{pmatrix} x_1(1 - x_1) & -x_1x_2 & -x_1x_3 & -x_1x_4 \\ -x_2x_1 & x_2(1 - x_2) & -x_2x_3 & -x_2x_4 \\ -x_3x_1 & -x_3x_2 & x_3(1 - x_3) & -x_3x_4 \\ -x_4x_1 & -x_4x_2 & -x_4x_3 & x_4(1 - x_4) \end{pmatrix}, \quad (3)$$

130 and $\mathbf{W}(t)$ is a three-dimensional standard Brownian motion. The term $x_1x_4 - x_2x_3$ in Eq. (2)
 131 is a measure of the linkage disequilibrium between the \mathcal{A} and \mathcal{B} loci introduced by Lewontin &
 132 Kojima (1960), which quantifies the non-random association of the alleles at the two loci.

133 2.1.2. Forward-in-time simulation of the Wright-Fisher diffusion

134 To obtain a numerical solution of the Wright-Fisher SDE in Eq. (1), we need to compute
 135 the diffusion term $\boldsymbol{\nu}(\mathbf{x})$, which we have to perform at each time step in most existing numerical
 136 simulation schemes. The diffusion term $\boldsymbol{\nu}(\mathbf{x})$ can be analytically derived with the Cholesky de-
 137 composition (Sato, 1976), which however, explodes at the boundaries. There exist other matrix
 138 decompositions capable of computing the diffusion term $\boldsymbol{\nu}(\mathbf{x})$ such as spectral decomposition,
 139 which are valid for positive semi-definite matrices, typically at the expense of either additional
 140 numerical errors and computational costs, or limitations in applicability to the infinitesimal
 141 covariance matrix $\boldsymbol{\Sigma}(\mathbf{x})$ of the form in Eq. (3).

142 Following He et al. (2020), we reformulate the Wright-Fisher SDE in the following form

$$d\mathbf{X}(t) = \boldsymbol{\mu}(\mathbf{X}(t))dt + \boldsymbol{\sigma}(\mathbf{X}(t))d\mathbf{W}(t), \quad t \geq t_0 \quad (4)$$

143 with initial condition $\mathbf{X}(t_0) = \mathbf{x}_0$, where the diffusion term $\boldsymbol{\sigma}(\mathbf{x})$ can be explicitly written down
 144 as

$$\boldsymbol{\sigma}(\mathbf{x}) = \begin{pmatrix} \sqrt{x_1x_2} & \sqrt{x_1x_3} & \sqrt{x_1x_4} & 0 & 0 & 0 \\ -\sqrt{x_2x_1} & 0 & 0 & \sqrt{x_2x_3} & \sqrt{x_2x_4} & 0 \\ 0 & -\sqrt{x_3x_1} & 0 & -\sqrt{x_3x_2} & 0 & \sqrt{x_3x_4} \\ 0 & 0 & -\sqrt{x_4x_1} & 0 & -\sqrt{x_4x_2} & -\sqrt{x_4x_3} \end{pmatrix}, \quad (5)$$

145 and $\mathbf{W}(t)$ is a six-dimensional standard Brownian motion. Combining Eqs. (3) and (5), we have

$$\boldsymbol{\sigma}(\mathbf{x})\boldsymbol{\sigma}(\mathbf{x})^\top = \boldsymbol{\Sigma}(\mathbf{x}) = \boldsymbol{\nu}(\mathbf{x})\boldsymbol{\nu}(\mathbf{x})^\top,$$

146 which implies that the two Wright-Fisher SDE's have the same infinitesimal generator

$$\mathcal{L} = \sum_{i=1}^4 \mu_i(\mathbf{x}) \frac{\partial}{\partial x_i} + \frac{1}{2} \sum_{i=1}^4 \sum_{j=1}^4 \Sigma_{ij}(\mathbf{x}) \frac{\partial^2}{\partial x_i \partial x_j},$$

147 thus having the same weak solution. This guarantees that we can achieve the solution of the

148 Wright-Fisher SDE of the form in Eq. (1) by numerically solving the Wright-Fisher SDE of the
 149 form in Eq. (4), which enables us to avoid boundary issues and reduce computational costs.

150 There exist a number of numerical simulation schemes for SDE's (see Kloeden & Platen,
 151 1992, for an excellent introduction). The numerical approach we adopt in this work is the com-
 152 monly used Euler-Maruyama scheme, one of the most popular numerical methods for SDE's in
 153 practice due to its high efficiency and low complexity. More specifically, we divide each genera-
 154 tion into L subintervals by setting $\Delta t = 1/(2NL)$, and then the Euler-Maruyama approximation
 155 of the Wright-Fisher diffusion can be formulated as

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

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

160 2.2. Bayesian inference of natural selection

161 Suppose that the available data are always sampled from the underlying population at a
 162 finite number of distinct time points, say $t_1 < t_2 < \dots < t_K$, where the time is measured in units
 163 of $2N$ generations to be consistent with the timescale of the Wright-Fisher diffusion. At the
 164 k -th sampling time point, we let $\mathbf{u}_k = (u_k^A, u_k^B)$ and $\mathbf{v}_k = (v_k^A, v_k^B)$ denote the counts of mutant
 165 alleles and ancestral alleles observed at loci \mathcal{A} and \mathcal{B} in the sample of n_k chromosomes drawn
 166 from the underlying population, respectively. The population genetic quantities of interest in
 167 this work are the scaled selection coefficients $\alpha_{\mathcal{A}}$ and $\alpha_{\mathcal{B}}$, the dominance parameters $h_{\mathcal{A}}$ and
 168 $h_{\mathcal{B}}$, and the scaled recombination rate ρ , which are denoted by $\boldsymbol{\vartheta} = (\alpha_{\mathcal{A}}, h_{\mathcal{A}}, \alpha_{\mathcal{B}}, h_{\mathcal{B}}, \rho)$.

169 2.2.1. Hidden Markov model

170 Similar to Bollback et al. (2008), the underlying population is assumed to evolve according
 171 to the two-locus Wright-Fisher diffusion with selection in our HMM framework, and the ob-
 172 servations are modelled as independent samples drawn from the underlying population at each
 173 sampling time point (see Figure 1 for the graphical representation of our HMM framework).
 174 To compute the posterior probability distribution $p(\boldsymbol{\vartheta} \mid \mathbf{u}_{1:K}, \mathbf{v}_{1:K})$, we condition and integrate

175 over all possible haplotype frequency trajectories of the underlying population at each sampling
 176 time point. More specifically, we let $\mathbf{x}_{1:K} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_K\}$ denote the haplotype frequency
 177 trajectories of the underlying population at the sampling time points $\mathbf{t}_{1:K}$. The posterior prob-
 178 ability distribution for the population genetic quantities of interest can then be written down
 179 as

$$p(\boldsymbol{\vartheta} | \mathbf{u}_{1:K}, \mathbf{v}_{1:K}) = \int_{\Omega_{\mathbf{X}}} \cdots \int_{\Omega_{\mathbf{X}}} p(\boldsymbol{\vartheta}, \mathbf{x}_{1:K} | \mathbf{u}_{1:K}, \mathbf{v}_{1:K}) d\mathbf{x}_{1:K}, \quad (6)$$

180 where

$$p(\boldsymbol{\vartheta}, \mathbf{x}_{1:K} | \mathbf{u}_{1:K}, \mathbf{v}_{1:K}) \propto p(\boldsymbol{\vartheta})p(\mathbf{x}_{1:K} | \boldsymbol{\vartheta})p(\mathbf{u}_{1:K}, \mathbf{v}_{1:K} | \mathbf{x}_{1:K}). \quad (7)$$

181 In Eq. (7), $p(\boldsymbol{\vartheta})$ is the prior probability distribution for the population genetic quantities of
 182 interest and can be taken to be a uniform prior over the parameter space if prior knowledge
 183 is poor, $p(\mathbf{x}_{1:K} | \boldsymbol{\vartheta})$ is the probability distribution for the haplotype frequency trajectories
 184 of the underlying population at the sampling time points $\mathbf{t}_{1:K}$, and $p(\mathbf{u}_{1:K}, \mathbf{v}_{1:K} | \mathbf{x}_{1:K})$ is the
 185 conditional probability for the observations at the sampling time points $\mathbf{t}_{1:K}$ given the haplotype
 186 frequency trajectories of the underlying population.

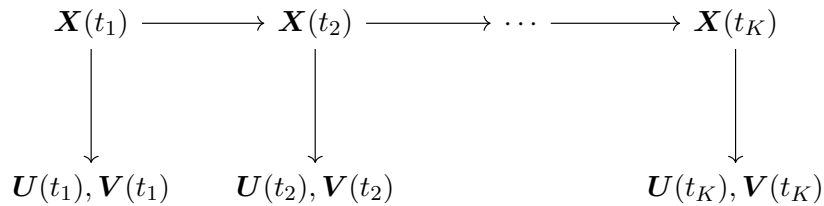


Figure 1: Graphical representation of the HMM framework for time series data of allele frequencies.

187 Since the Wright-Fisher diffusion is shown to be a Markov process, the probability distribu-
 188 tion for the haplotype frequency trajectories of the underlying population at the sampling time
 189 points $\mathbf{t}_{1:K}$ can be decomposed as

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

190 where $p(\mathbf{x}_1 | \boldsymbol{\vartheta})$ is the prior probability distribution for the haplotype frequencies of the underly-
 191 ing population at the initial sampling time point and can be taken to be a uniform prior over the
 192 state space $\Omega_{\mathbf{X}}$, known as the flat Dirichlet distribution, if prior knowledge is poor. The term
 193 in the product above, $p(\mathbf{x}_{k+1} | \mathbf{x}_k; \boldsymbol{\vartheta})$, is the transition probability density of the Wright-Fisher

194 diffusion between two consecutive sampling time points for $k = 1, 2, \dots, K - 1$, which can be
 195 obtained by numerically solving the Kolmogorov backward equation (or its adjoint) associated
 196 with the Wright-Fisher diffusion. However, this requires a fine enough discretisation of the state
 197 space $\Omega_{\mathcal{X}}$, if a finite difference method is used, and strongly depends on the underlying popula-
 198 tion genetic parameters (Ragsdale & Gutenkunst, 2017). In addition, numerically solving such
 199 a PDE in three dimensions for our posterior computation is computationally challenging and
 200 prohibitively expensive. We therefore resort to an ‘exact-approximate’ Monte Carlo procedure
 201 (Andrieu & Vihola, 2016) in this work that only involves simulating the Wright-Fisher SDE in
 202 the form of Eq. (4), as a tractable alternative.

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

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

205 where $p(\mathbf{u}_k, \mathbf{v}_k \mid \mathbf{x}_k)$ is the conditional probability for the observations at the k -th sampling time
 206 point given the haplotype frequency trajectories of the underlying population for $k = 1, 2, \dots, K$.
 207 To calculate the emission probability $p(\mathbf{u}_k, \mathbf{v}_k \mid \mathbf{x}_k)$, we let $\mathbf{z}_k = (z_{1,k}, z_{2,k}, z_{3,k}, z_{4,k})$ denote the
 208 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
 209 point, which are usually unobserved (see Figure 2 for the graphical representation of our HMM
 210 framework incorporating the additional level of sampling noise). We then have

$$p(\mathbf{u}_k, \mathbf{v}_k \mid \mathbf{x}_k) = \sum_{\mathbf{z}_k \in \Omega_{\mathbf{Z}_k}} p(\mathbf{z}_k \mid \mathbf{x}_k) p(\mathbf{u}_k, \mathbf{v}_k \mid \mathbf{z}_k), \quad (8)$$

211 where

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

212 Conditional on the haplotype frequency trajectories of the underlying population at the k -th
 213 sampling time point, the haplotype counts of the sample can be modelled through multinomial
 214 sampling from the underlying population with sample size n_k . We can then formulate the first

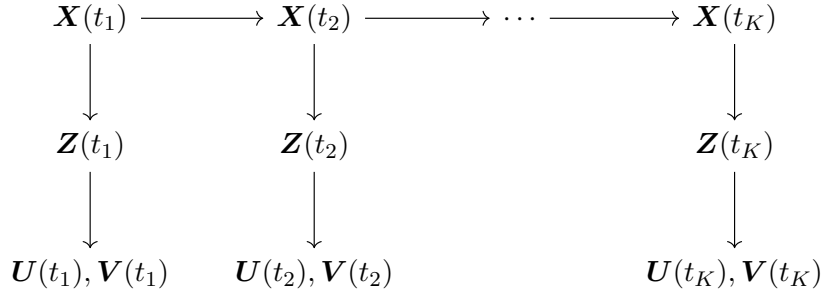


Figure 2: Graphical representation of the HMM framework for time series data of allele frequencies incorporating the additional level of sampling noise caused by the unobserved haplotype counts of the sample.

215 term in the summation of Eq. (8) as

$$p(\mathbf{z}_k | \mathbf{x}_k) = \frac{n_k!}{\prod_{i=1}^4 z_{i,k}!} \prod_{i=1}^4 x_{i,k}^{z_{i,k}}. \quad (9)$$

216 The second term in the summation of Eq. (8) can be decomposed as

$$p(\mathbf{u}_k, \mathbf{v}_k | \mathbf{z}_k) = p(u_k^A, v_k^A | \mathbf{z}_k) p(u_k^B, v_k^B | \mathbf{z}_k). \quad (10)$$

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

$$p(u_k^A, v_k^A | \mathbf{z}_k) = b(u_k^A; z_{1,k} + z_{2,k}, 1 - \phi) b(v_k^A; z_{3,k} + z_{4,k}, 1 - \phi) \quad (11)$$

$$p(u_k^B, v_k^B | \mathbf{z}_k) = b(u_k^B; z_{1,k} + z_{3,k}, 1 - \phi) b(v_k^B; z_{2,k} + z_{4,k}, 1 - \phi), \quad (12)$$

219 where

$$b(x; n, p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}. \quad (13)$$

220 The probability that the sampled chromosome at a single locus is of unknown type is usually
 221 unavailable, but we can estimate it with

$$\hat{\phi} = 1 - \frac{\sum_{k=1}^K (u_k^A + v_k^A) + \sum_{k=1}^K (u_k^B + v_k^B)}{2 \sum_{k=1}^K n_k}. \quad (14)$$

222 2.2.2. Particle marginal Metropolis-Hastings

223 To compute the marginal posterior $p(\boldsymbol{\vartheta} | \mathbf{u}_{1:K}, \mathbf{v}_{1:K})$, we resort to MCMC techniques since
 224 the posterior probability distribution in Eq. (6) is unavailable in a closed form. A Metropolis-

225 Hastings (MH) scheme can be devised to explore the population genetic quantities of interest
 226 with a fairly arbitrary proposal probability distribution, *e.g.*, a random walk proposal, where
 227 a sample of new candidates of the parameters $\boldsymbol{\vartheta}^*$ is drawn from the proposal $q(\boldsymbol{\vartheta}^* | \boldsymbol{\vartheta})$ and is
 228 accepted with the Metropolis-Hastings ratio

$$A = \frac{p(\boldsymbol{\vartheta}^*) p(\mathbf{u}_{1:K}, \mathbf{v}_{1:K} | \boldsymbol{\vartheta}^*) q(\boldsymbol{\vartheta} | \boldsymbol{\vartheta}^*)}{p(\boldsymbol{\vartheta}) p(\mathbf{u}_{1:K}, \mathbf{v}_{1:K} | \boldsymbol{\vartheta}) q(\boldsymbol{\vartheta}^* | \boldsymbol{\vartheta})}. \quad (15)$$

229 Our problem reduces to the calculation of the intractable marginal likelihood $p(\mathbf{u}_{1:K}, \mathbf{v}_{1:K} | \boldsymbol{\vartheta})$
 230 in Eq. (15), which can be formulated as

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

231 and achieved with a Monte Carlo (MC) estimate (Beaumont, 2003; Andrieu & Roberts, 2009).
 232 This pseudo-marginal MCMC algorithm exploits the fact that the MC estimate of the marginal
 233 likelihood $p(\mathbf{u}_{1:K}, \mathbf{v}_{1:K} | \boldsymbol{\vartheta})$ is unbiased (or has a constant bias independent of the parameters
 234 $\boldsymbol{\vartheta}$) and targets the marginal posterior $p(\boldsymbol{\vartheta} | \mathbf{u}_{1:K}, \mathbf{v}_{1:K})$.

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

247 In our Bayesian inference procedure, the implementation of the PMMH algorithm requires
 248 the SMC estimate of the marginal likelihood $p(\mathbf{u}_{1:K}, \mathbf{v}_{1:K} | \boldsymbol{\vartheta})$. This can be achieved by the boot-
 249 strap particle filter introduced by Gordon et al. (1993) in the following manner. For the sampling

250 time point t_1 , we first generate a sample of M particles, denoted by $\mathbf{x}_1^{1:M} = \{\mathbf{x}_1^1, \mathbf{x}_1^2, \dots, \mathbf{x}_1^M\}$,
 251 from the prior $p(\mathbf{x}_1 | \boldsymbol{\vartheta})$ and assign each particle \mathbf{x}_1^m a weight given by

$$w_1^m = p(\mathbf{u}_1, \mathbf{v}_1 | \mathbf{x}_1^m)$$

252 for $m = 1, 2, \dots, M$, where the superscript m denotes the particle label. We then calculate the
 253 SMC estimate of the marginal likelihood for the observations \mathbf{u}_1 and \mathbf{v}_1 by

$$\hat{p}(\mathbf{u}_1, \mathbf{v}_1 | \boldsymbol{\vartheta}) = \frac{1}{M} \sum_{m=1}^M w_1^m$$

254 and resample M times with replacement from the sample of particles $\mathbf{x}_1^{1:M}$ with the probabil-
 255 ities given by the normalised weights $\mathbf{w}_1^{1:M} / \sum_{m=1}^M w_1^m$. We repeat the following steps for the
 256 sampling time points $\mathbf{t}_{2:K}$:

257 Step 1: Generate each particle \mathbf{x}_k^m by simulating the Wright-Fisher diffusion $\mathbf{X}(t)$ on the time
 258 interval $[t_{k-1}, t_k]$ starting at the frequency $\mathbf{X}(t_{k-1}) = \mathbf{x}_{k-1}^m$ with the Euler-Maruyama
 259 scheme for $m = 1, 2, \dots, M$.

260 Step 2: Assign each particle \mathbf{x}_k^m a weight given by

$$w_k^m = p(\mathbf{u}_k, \mathbf{v}_k | \mathbf{x}_k^m)$$

261 for $m = 1, 2, \dots, M$.

262 Step 3: Calculate the SMC estimate of the marginal likelihood for the observations $\mathbf{u}_{1:k}$ and
 263 $\mathbf{v}_{1:k}$ by

$$\hat{p}(\mathbf{u}_{1:k}, \mathbf{v}_{1:k} | \boldsymbol{\vartheta}) = \hat{p}(\mathbf{u}_{1:k-1}, \mathbf{v}_{1:k-1} | \boldsymbol{\vartheta}) \frac{1}{M} \sum_{m=1}^M w_k^m.$$

264 Step 4: Resample M times with replacement from the sample of particles $\mathbf{x}_k^{1:M}$ with the prob-
 265 abilities given by the normalised weights $\mathbf{w}_k^{1:M} / \sum_{m=1}^M w_k^m$.

266 Our Bayesian inference procedure then consists in the followings. We first generate a sample
 267 of initial candidates of the parameters $\boldsymbol{\vartheta}$ from the prior $p(\boldsymbol{\vartheta})$ and then run a bootstrap particle
 268 filter with the proposed parameters $\boldsymbol{\vartheta}$ to obtain the SMC estimate of the marginal likelihood
 269 $\hat{p}(\mathbf{u}_{1:K}, \mathbf{v}_{1:K} | \boldsymbol{\vartheta})$. We repeat the following steps until a sufficient number of the samples of the
 270 parameters $\boldsymbol{\vartheta}$ have been obtained:

271 Step 1: Generate a sample of new candidates of the parameters $\boldsymbol{\vartheta}^*$ from the proposal $q(\boldsymbol{\vartheta}^* | \boldsymbol{\vartheta})$.

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

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

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

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

$$\hat{\boldsymbol{\vartheta}} = \mathbb{E}[\boldsymbol{\vartheta} | \mathbf{u}_{1:K}, \mathbf{v}_{1:K}] = \int \boldsymbol{\vartheta} p(\boldsymbol{\vartheta} | \mathbf{u}_{1:K}, \mathbf{v}_{1:K}) d\boldsymbol{\vartheta}.$$

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

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

281 2.3. Data availability

282 The authors state that all data necessary for confirming the conclusions of this work are
283 represented fully within the article. Source code implementing the approach described in this
284 work is available at <https://github.com/zhangyi-he/WFM-2L-DiffusApprox-FwdPMMH>. Sup-
285 plemental material available at FigShare: <https://doi.org/10.25386/genetics.12821585>.

286 3. Results

287 In this section, we show how our Bayesian inference method performs on simulated datasets
288 with known population genetic parameter values, including a scenario where sampled chromo-
289 somes contain unknown alleles. We also present two examples to show the improvement in the
290 inference of natural selection by explicitly modelling genetic recombination and local linkage.
291 Finally, we apply our approach to the aDNA data associated with horse white spotting patterns
292 from previous studies of Ludwig et al. (2009), Pruvost et al. (2011) and Wutke et al. (2016).

293 *3.1. Analysis of simulated data*

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

301 For each simulated dataset, given the values of the population genetic parameters $\boldsymbol{\vartheta}$ and the
302 initial population haplotype frequencies \boldsymbol{x}_0 , we simulate the haplotype frequency trajectories of
303 the underlying population according to the two-locus Wright-Fisher model with selection. After
304 obtaining the simulated population haplotype frequency trajectories, we draw the unobserved
305 sample haplotype counts independently at each sampling time point according to the multino-
306 mial distribution in Eq. (9) first and then we generate the observed sample mutant allele counts
307 and ancestral allele counts with Eqs. (10)-(13).

308 *3.1.1. Power to infer natural selection*

309 We vary the selection coefficients with $s_{\mathcal{A}} \in \{0.003, 0.01\}$ and $s_{\mathcal{B}} \in \{0, 0.002, 0.008\}$, and the
310 recombination rate with $r \in \{0.00001, 0.01\}$ in our simulation studies. We perform 100 replicates
311 for each of the 12 possible combinations of the selection coefficients and the recombination rate.
312 For each replicate, we set the initial population haplotype frequencies $\boldsymbol{x}_0 = (0.04, 0.08, 0.08, 0.8)$
313 and simulate the haplotype frequency trajectories of the underlying population according to the
314 two-locus Wright-Fisher model with selection. We sample 50 chromosomes from the underlying
315 population at every 50 generations throughout 500 generations.

316 We choose a uniform prior over the interval $[-1, 1]$ for the selection coefficients $s_{\mathcal{A}}$ and $s_{\mathcal{B}}$,
317 and a flat Dirichlet prior for the initial population haplotype frequencies \boldsymbol{x}_0 in our Bayesian
318 inference method. We divide each generation into 5 subintervals in the Euler-Maruyama scheme
319 and run the PMMH algorithm with 1500 particles and 10000 iterations. We discard the initial
320 2000 iterations as the burn-in period and then thin the remaining PMMH output by selecting
321 every fourth value.

322 The resulting boxplots of the empirical studies are shown in Figure 3 for the allele frequency

323 datasets generated without missing values ($\phi = 0$ in Eqs. (11) and (12)) and Figure 4 for
324 the allele frequency datasets generated with missing values ($\phi = 0.02$ in Eqs. (11) and (12)),
325 respectively. In the two figures, the tips of the whiskers denote the 2.5%-quantile and the 97.5%-
326 quantile, and the boxes represent the first and third quartile with the median in the middle. As
327 can be seen from the boxplot results, our estimates for the selection coefficients at both loci show
328 little bias across different parameter ranges, no matter whether sampled chromosomes contain
329 unknown alleles or not, although one can discern a slight bias for small selection coefficients.
330 With the increase of the selection coefficients, our estimates for the selection coefficients become
331 more accurate. See the bias and the RMSE of the resulting MMSE estimates in Supplemental
332 Material, Tables S1 and S2.

333 For each combination of the selection coefficients and the recombination rate, we calculate
334 the proportion of the 95% highest posterior density (HPD) intervals that include the true values,
335 shown in the bottom left corner of each boxplot in Figures 3 and 4. On average, 92.00% of the
336 runs result in the true values of the selection coefficients being within the 95% HPD intervals
337 for the simulated datasets without missing values, 93.33% for tightly linked loci and 90.67% for
338 loosely linked loci. For simulated datasets with 2% missing values, 92.08% of the runs result
339 in the true values of the selection coefficients being within the 95% HPD intervals, *i.e.*, 93.33%
340 for tightly linked loci and 90.83% for loosely linked loci. We can see that small recombination
341 rates can lead to better results for both loci.

342 In Figure 5, we illustrate the resulting boxplots of the empirical studies where the simulated
343 data is given as haplotype frequencies (instead of allele frequencies in Figures 3 and 4). Com-
344 pared to the estimates from allele frequency data, our estimates from haplotype frequency data
345 are closer to their true values with smaller variances, especially for tightly linked loci. On aver-
346 age, 92.50% of the runs result in the true values of the selection coefficients being within their
347 95% HPD intervals on average, with 93.67% for tightly linked loci and 91.33% for loosely linked
348 loci. The bias and the RMSE of the resulting MMSE estimates are summarised in Supplemental
349 Material, Table S3. This improvement in the performance of our estimates is to be expected
350 as all else being equal haplotype frequency data contain more information than allele frequency
351 data. The complex interplay between the four haplotypes in the sample can be directly observed
352 in haplotype frequency data but only partially observed in allele frequency data.

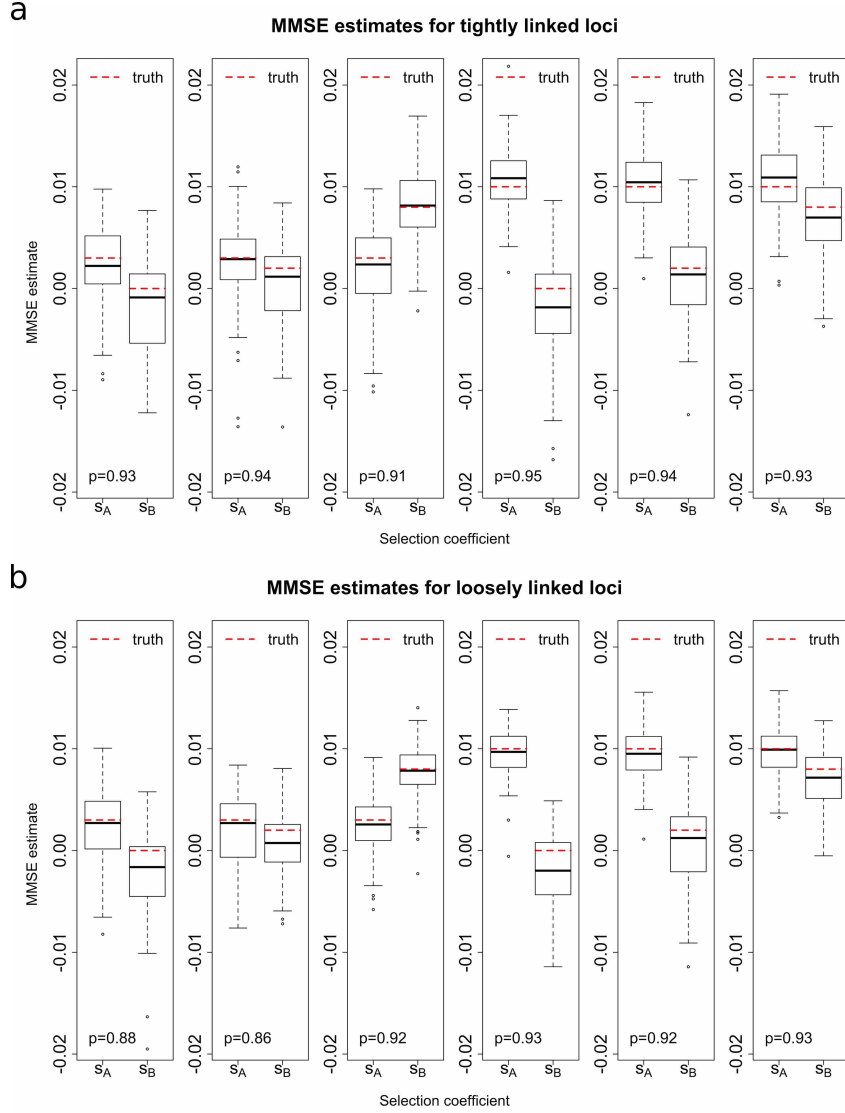


Figure 3: Empirical distributions of the MMSE estimates for 100 *allele frequency* datasets (*without* missing values) simulated with the initial population haplotype frequencies $\mathbf{x}_0 = (0.04, 0.08, 0.08, 0.8)$ and the dominance parameters $h_A = 0.5$ and $h_B = 0.5$ for the case of (a) tightly linked loci with the recombination rate $r = 0.00001$ and (b) loosely linked loci with the recombination rate $r = 0.01$. The p value in the bottom left corner indicates the proportion of the runs where the true values of the selection coefficients both fall within their 95% HPD intervals.

353 However, as illustrated in Figure 5, our estimates are still slightly biased for small selection
 354 coefficients. This may be caused by the initial population frequencies of the haplotypes that
 355 contain mutant alleles being close to 0 in our simulated data. In such a situation, the population
 356 frequency trajectories of these haplotypes will be, with high probability, near 0 during the
 357 sampling period for small selection coefficients (see Supplemental Material, Figures S1 and S2).
 358 This can cause a number of simulated datasets to have sample counts 0 for the haplotypes that
 359 contain mutant alleles, especially when the selection coefficients are small. Such datasets contain

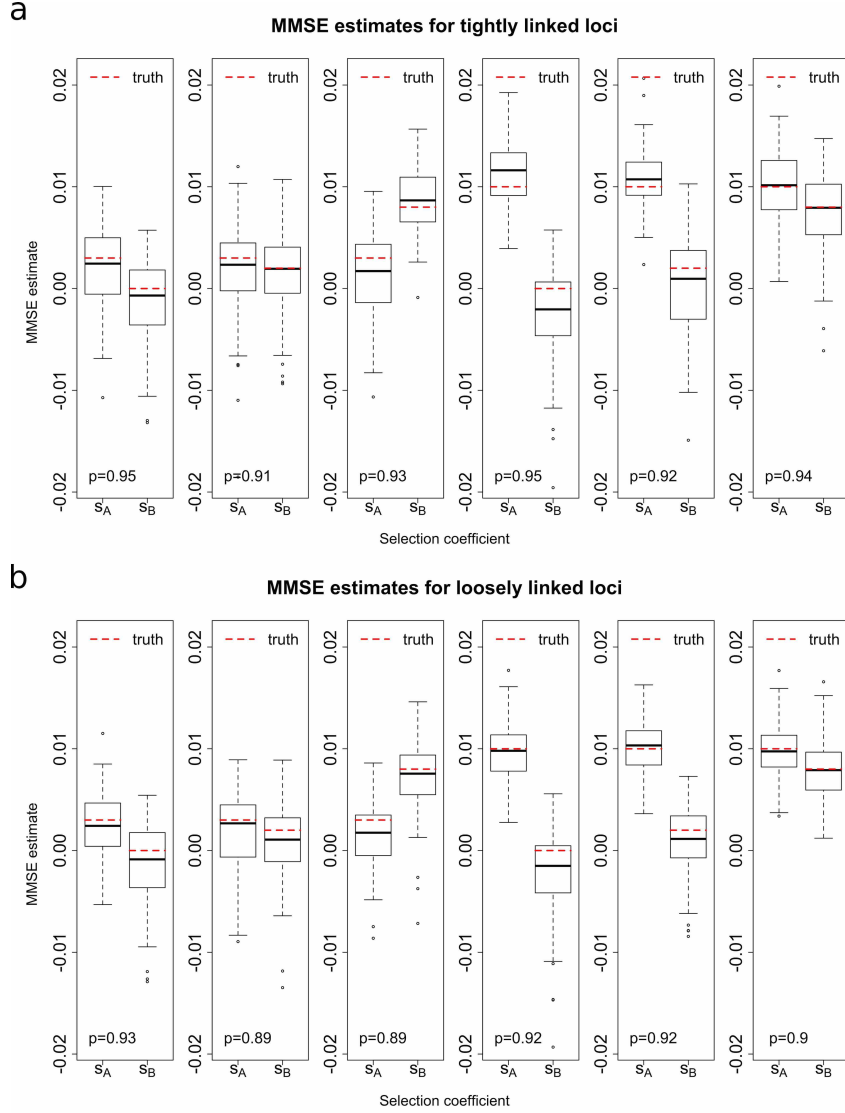


Figure 4: Empirical distributions of the MMSE estimates for 100 *allele frequency* datasets (*with 2% missing values*) simulated with the initial population haplotype frequencies $\mathbf{x}_0 = (0.04, 0.08, 0.08, 0.8)$ and the dominance parameters $h_A = 0.5$ and $h_B = 0.5$ for the case of (a) tightly linked loci with the recombination rate $r = 0.00001$ and (b) loosely linked loci with the recombination rate $r = 0.01$. The p value in the bottom left corner indicates the proportion of the runs where the true values of the selection coefficients both fall within their 95% HPD intervals.

360 little information on the underlying selection coefficients. As can be observed from Figure 6,
 361 the bias can be almost completely eliminated for all combinations of the selection coefficients
 362 and the recombination rate if the starting population frequencies of the haplotypes that contain
 363 mutant alleles are taken to be intermediate values like $\mathbf{x}_0 = (0.1, 0.2, 0.3, 0.4)$. The bias and the
 364 RMSE of the resulting MMSE estimates are summarised in Supplemental Material, Table S4.
 365 The haplotype frequency trajectories of the underlying population for the haplotype frequency
 366 datasets simulated with the initial population haplotype frequencies $\mathbf{x}_0 = (0.1, 0.2, 0.3, 0.4)$ can

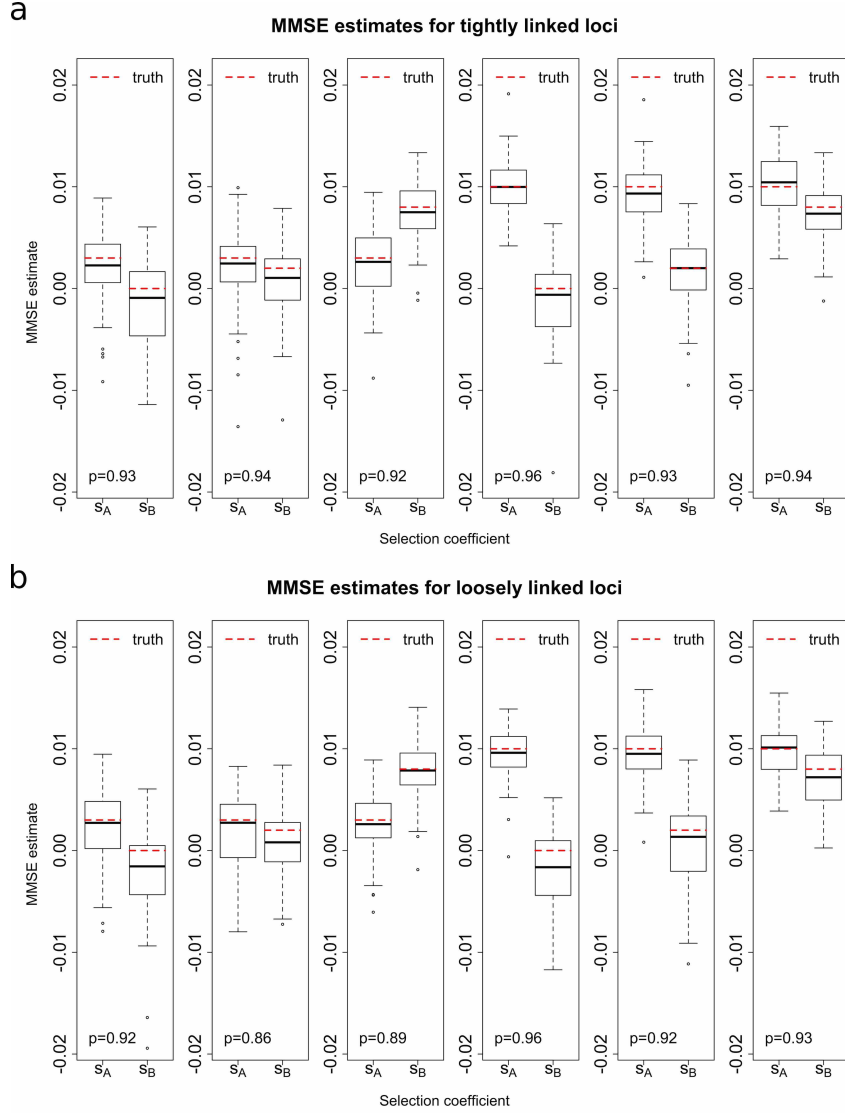


Figure 5: Empirical distributions of the MMSE estimates for 100 *haplotype frequency* datasets simulated with the initial population haplotype frequencies $\mathbf{x}_0 = (0.04, 0.08, 0.08, 0.8)$ and the dominance parameters $h_A = 0.5$ and $h_B = 0.5$ for the case of (a) tightly linked loci with the recombination rate $r = 0.00001$ and (b) loosely linked loci with the recombination rate $r = 0.01$. The p value in the bottom left corner indicates the proportion of the runs where the true values of the selection coefficients both fall within their 95% HPD intervals.

367 be found in Supplemental Material, Figures S3 and S4. We also assess the performance of our
 368 method for the case that a new mutation arose in the population (at frequency $1/(2N)$) at $t = 0$
 369 when the neighbouring mutation became established. See Supplemental Material, Figure S5 and
 370 Table S5 for boxplots of the resulting MMSE estimates with their bias and RMSE, which show
 371 that our approach can still produce precise estimates of the selection coefficients in this case.
 372 It should be noticed that in this case we condition the mutant alleles at both loci to survive
 373 until the most recent sampling time point and sample 50 chromosomes from the underlying

374 population at every 120 generations throughout 1200 generations so that a significant number
 375 of the realisations of the haplotype frequency trajectories of the underlying population can
 376 capture a significant proportion of the selective sweep.

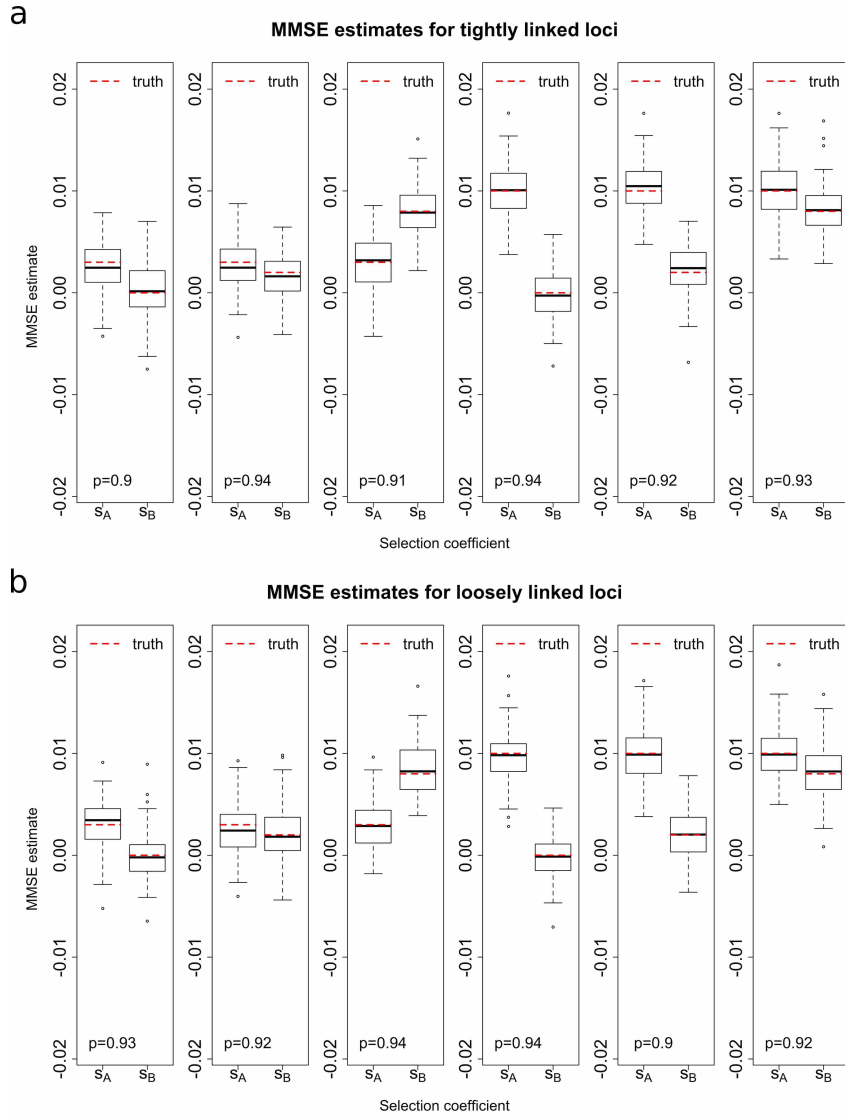


Figure 6: Empirical distributions of the MMSE estimates for 100 *haplotype frequency* datasets simulated with the initial population haplotype frequencies $\mathbf{x}_0 = (0.1, 0.2, 0.3, 0.4)$ and the dominance parameters $h_A = 0.5$ and $h_B = 0.5$ for the case of (a) tightly linked loci with the recombination rate $r = 0.00001$ and (b) loosely linked loci with the recombination rate $r = 0.01$. The p value in the bottom left corner indicates the proportion of the runs where the true values of the selection coefficients both fall within their 95% HPD intervals.

377 In summary, our Bayesian inference procedure can deliver accurate estimates of the selection
 378 coefficients based on time series data of allele frequencies across different parameter ranges,
 379 regardless of whether sampled chromosomes contain unknown alleles or not. We also generate
 380 datasets with other selection schemes, *e.g.*, the dominance parameters $h_A = 0$ and $h_B = 1$. The

381 resulting boxplots of the simulation studies are shown in Supplemental Material, Figure S6, with
382 the bias and the RMSE of the resulting MMSE estimates summarised in Supplemental Material,
383 Table S6. In addition to MMSE estimates, we present the bias and the RMSE of MAP estimates
384 (see Supplemental Material, Tables S7-S12), which display very similar characteristics to the
385 MMSE estimates. The boxplots for MAP estimates show little bias, with the upper and lower
386 quartiles of the MAP estimates being similar to those of the MMSE estimates (see Supplemental
387 Material, Figures S7-S12).

388 3.1.2. *Improvement from modelling genetic recombination and local linkage*

389 In the case where a pair of loci are both suspected to be subject to natural selection, one can
390 still use a single-locus method to each locus to estimate selection coefficient. To our knowledge,
391 there has been a considerable amount of work on the statistical inference of natural selection at
392 a single locus from time series data of allele frequencies (*e.g.*, Bollback et al., 2008; Malaspinas
393 et al., 2012; Steinrücken et al., 2014; Schraiber et al., 2016; Ferrer-Admetlla et al., 2016; He
394 et al., 2019). However, using a single-locus approach may lead to inaccurate estimates of the
395 selection coefficients when the two loci are linked (He et al., 2020). In the case of tightly linked
396 loci, modelling genetic recombination and local linkage becomes necessary, thus our two-locus
397 method is far more desirable. Below we illustrate with two examples of tightly linked loci with
398 the recombination rate $r = 0.00001$. We simulate the haplotype frequency trajectories of the
399 underlying population through the two-locus Wright-Fisher model with selection and draw 200
400 chromosomes from the underlying population at generations 0, 100, 200, 300, 400 and 500.

401 In the first example, we consider a positively selected locus \mathcal{A} tightly linked with a selectively
402 neutral locus \mathcal{B} , where we set the selection coefficients $s_{\mathcal{A}} = 0.01$ and $s_{\mathcal{B}} = 0$, respectively. We
403 take the initial haplotype frequencies of the underlying population to be $\mathbf{x}_0 = (0.2, 0.1, 0.3, 0.4)$.
404 The mutant allele frequency trajectories of the sample are shown in Figure 7a. The posterior
405 probability distributions obtained through our single-locus approach, described in Supplemental
406 Material, File S3, are shown in Figure 7b, and the posterior probability distributions achieved
407 with our two-locus method, described in Section 2.2, are shown in Figure 7c. Bayesian estimates
408 of the selection coefficients $s_{\mathcal{A}}$ and $s_{\mathcal{B}}$ are summarised in Table 1.

409 One can observe that with a single-locus method, the estimate for the selection coefficient $s_{\mathcal{A}}$
410 is reasonably accurate, but the estimate for the selection coefficient $s_{\mathcal{B}}$ is off by a large amount.

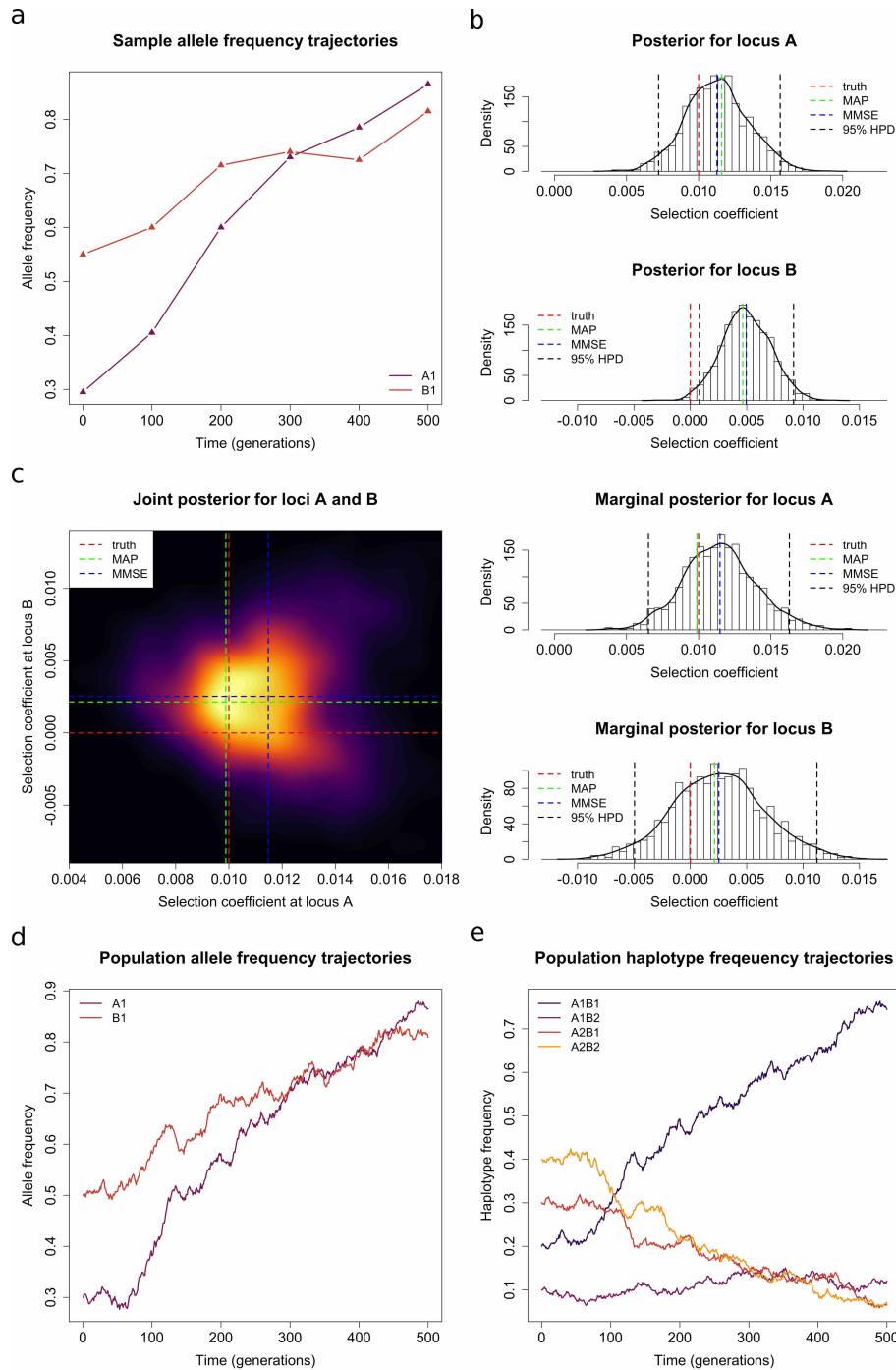


Figure 7: A comparison of the performance differences of the single-locus method and the two-locus method on the simulated dataset of a positively selected locus tightly linked with a selectively neutral locus. (a) Sample mutant allele frequency trajectories. (b) Posteriors obtained with a single-locus method. (c) Posteriors obtained with a two-locus method. (d) Population mutant allele frequency trajectories. (e) Population haplotype frequency trajectories.

411 The true value for the selection coefficient s_B is 0, but the single-locus approach produces an
 412 estimate of roughly 0.005 and a 95% HPD interval that only encompasses positive values, which
 413 is strong evidence for the presence of positive selection. In comparison, the estimates for both

		single-locus method	two-locus method
selection coefficient $s_{\mathcal{A}}$	MAP ($\times 10^{-2}$)	1.160	0.989
	MMSE ($\times 10^{-2}$)	1.127	1.148
	95% HPD ($\times 10^{-2}$)	[0.722, 1.566]	[0.652, 1.630]
selection coefficient $s_{\mathcal{B}}$	MAP ($\times 10^{-2}$)	0.465	0.214
	MMSE ($\times 10^{-2}$)	0.496	0.253
	95% HPD ($\times 10^{-2}$)	[0.080, 0.916]	[-0.495, 1.123]

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 selectively neutral locus.

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

415 To understand the poor performance of the single-locus method in this example, we plot the
416 mutant allele frequency trajectories of the underlying population in Figure 7d and the haplotype
417 frequency trajectories of the underlying population in Figure 7e. The increase in the frequency
418 of the \mathcal{B}_1 allele with time, despite it having a selection coefficient of 0, is caused by the $\mathcal{A}_1\mathcal{B}_1$
419 haplotype, which has a selection coefficient of 0.01. This compensates for the decrease in the
420 frequency of the $\mathcal{A}_2\mathcal{B}_1$ haplotype, resulting in an increasing trajectory for the \mathcal{B}_1 allele, albeit
421 with a slower rate of increase than the \mathcal{A}_1 allele. With the two-locus approach, however, the
422 interplay between all four haplotypes are taken into account and it produces accurate estimates
423 for both of the selection coefficients $s_{\mathcal{A}}$ and $s_{\mathcal{B}}$.

424 In the second example, we consider two positively selected and tightly linked loci \mathcal{A} and \mathcal{B} ,
425 where we take the selection coefficients to be $s_{\mathcal{A}} = 0.01$ and $s_{\mathcal{B}} = 0.005$, respectively, and set
426 the initial haplotype frequencies of the underlying population to be $\mathbf{x}_0 = (0.05, 0.05, 0.7, 0.2)$.
427 The results are illustrated in Figure 8 and summarised in Table 2. In this example, with the
428 single-locus method, the estimate for the selection coefficient $s_{\mathcal{A}}$ is reasonably accurate, but the
429 estimate for the selection coefficient $s_{\mathcal{B}}$ is off by a large amount, *i.e.*, its true value lies outside
430 the 95% HPD interval. In fact, although the \mathcal{B}_1 allele is favoured by natural selection with a
431 selection coefficient of 0.005, the resulting estimate for the selection coefficient $s_{\mathcal{B}}$ is roughly
432 -0.0015 with a 95% HPD interval that includes the value 0, which implies no strong evidence for
433 natural selection. In comparison, the two-locus method again produces fairly accurate estimates
434 for both of the selection coefficients $s_{\mathcal{A}}$ and $s_{\mathcal{B}}$.

435 As shown in Figure 8, the frequency of the \mathcal{A}_1 allele increases with time due to the increase in
436 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,

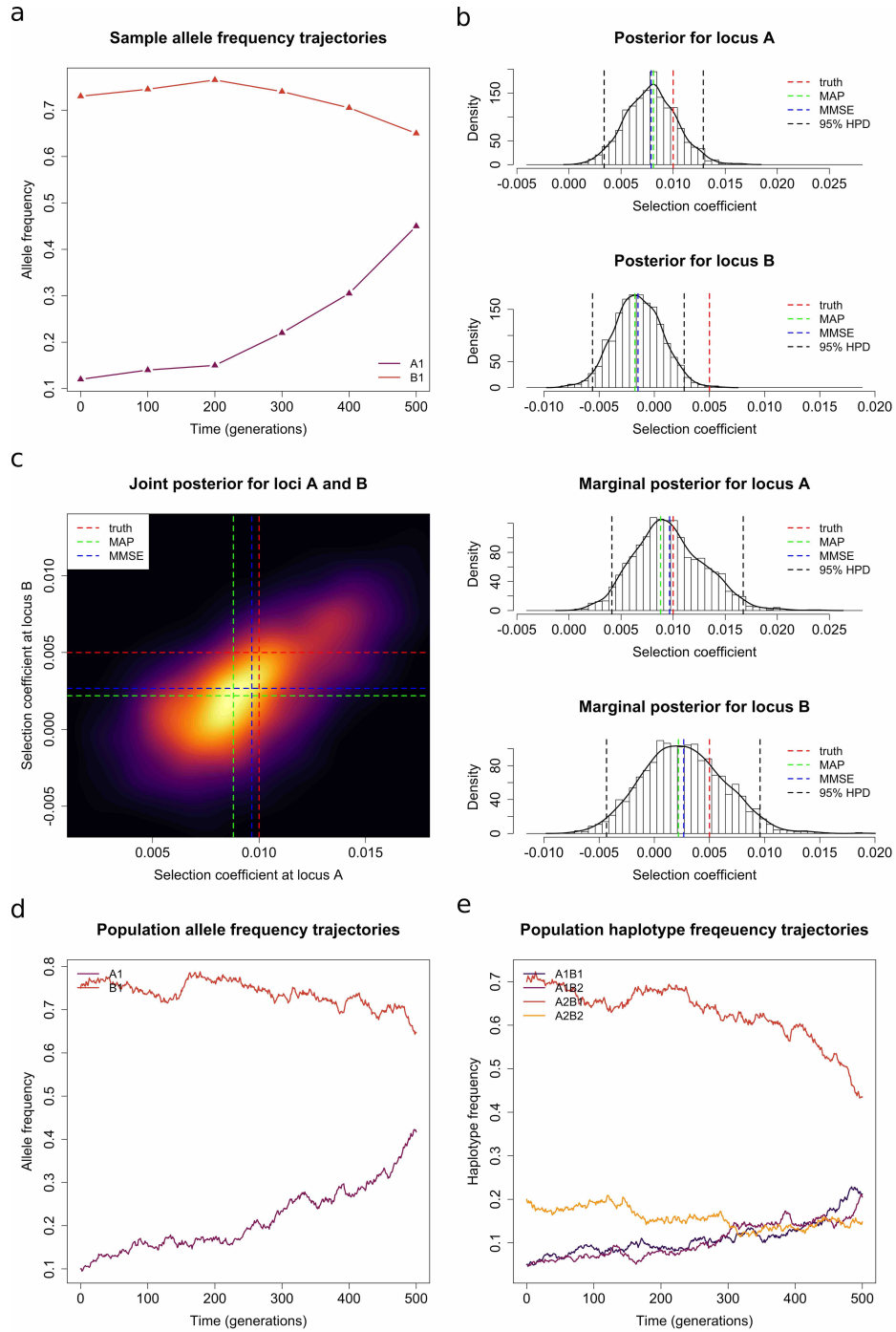


Figure 8: A comparison of the performance differences of the single-locus method and the two-locus method on the simulated dataset of a pair of positively selected and tightly linked loci. (a) Sample mutant allele frequency trajectories. (b) Posteriors obtained with a single-locus method. (c) Posteriors obtained with a two-locus method. (d) Population mutant allele frequency trajectories. (e) Population haplotype frequency trajectories.

437 with the selection coefficients of 0.015 and 0.01, respectively. The \mathcal{B}_1 allele is made up of the
 438 $\mathcal{A}_1\mathcal{B}_1$ and $\mathcal{A}_2\mathcal{B}_1$ haplotypes, with the selection coefficients of 0.01 and 0.005, respectively, which
 439 are the second and third most selected haplotypes. As a result of their initial conditions and

		single-locus method	two-locus method
selection coefficient $s_{\mathcal{A}}$	MAP ($\times 10^{-2}$)	0.811	0.879
	MMSE ($\times 10^{-2}$)	0.789	0.966
	95% HPD ($\times 10^{-2}$)	[0.338, 1.290]	[0.412, 1.673]
selection coefficient $s_{\mathcal{B}}$	MAP ($\times 10^{-2}$)	-0.176	0.219
	MMSE ($\times 10^{-2}$)	-0.148	0.267
	95% HPD ($\times 10^{-2}$)	[-0.560, 0.271]	[-0.433, 0.958]

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.

440 selection coefficients, the frequency of the \mathcal{B}_1 allele roughly holds constant in time, since it is
441 somewhat out-competed by the \mathcal{A}_1 allele. Viewing the trajectory of the \mathcal{B}_1 allele in isolation
442 does not give strong evidence that it is selectively advantageous, which results in an estimate of
443 roughly 0 in its selection coefficient through the single-locus approach. Moreover, even the 95%
444 HPD interval for the single-locus method does not include the true selection coefficient of 0.005
445 for the \mathcal{B}_1 allele. Using the two-locus approach, we are again able to obtain accurate estimates
446 for both of the selection coefficients $s_{\mathcal{A}}$ and $s_{\mathcal{B}}$.

447 In these two examples, we choose a uniform prior over the interval $[-1, 1]$ for the selection
448 coefficients, and we select a flat Dirichlet prior for the initial population haplotype frequencies
449 in the two-locus method and a uniform prior over the interval $[0, 1]$ for the initial population
450 allele frequency in the single-locus method, respectively. Other settings in the Euler-Maruyama
451 scheme and the PMMH algorithm are the same as we applied in the empirical studies in Sec-
452 tion 3.1.1. Compared to existing single-locus approaches, our two-locus method explicitly in-
453 corporates the effect of genetic recombination and the information of local linkage through the
454 two-locus Wright-Fisher diffusion with selection. Indeed, the dynamics of the two-locus Wright-
455 Fisher diffusion with selection can demonstrate complex and unpredictable behaviour (see, *e.g.*,
456 Yu & Etheridge, 2010; Cuthbertson et al., 2012), which can result in inaccurate estimates of
457 the selection coefficients if one simply employs a single-locus approach. In contrast, applying
458 our two-locus method can yield precise estimates of the selection coefficients at both loci.

459 3.2. Analysis of real data

460 We apply our Bayesian inference method to real data by re-analysing time serial samples
461 of segregating alleles of the equine homologue of proto-oncogene *c-kit* (*KIT*). These data come
462 from previous studies of Ludwig et al. (2009), Pruvost et al. (2011) and Wutke et al. (2016), and

463 the sample information and genotyping results for all successfully typed horses can be found in
464 Wutke et al. (2016), which are summarised in Table 3. The *KIT* gene in horses resides on the
465 long arm of chromosome 3 and lies in two intervals associated with white spotting patterns, one
466 in the intron 13 which codes for tobiano (*KIT13*), with the other in intron 16 which codes for
467 sabino (*KIT16*). At the *KIT13* locus, the ancestral allele is designated *KM0*, while the mutant
468 allele, associated with the tobiano pattern and acting as dominant (Brooks et al., 2007), is
469 designated *KM1*. The tobiano pattern is characterised by depigmented patches of skin and
470 associated hair that often cross the dorsal midline and cover the legs. At the *KIT16* locus, the
471 ancestral allele is designated *sb1*, while the mutant allele associated with the sabino pattern
472 and acting as semi-dominant (Brooks & Bailey, 2005), is designated *SB1*. The sabino pattern
473 is characterised by irregularly bordered white patches of skin and associated hair that begin at
474 the extremities and face and may extend up to the belly and midsection.

sample time	sample size	<i>KIT13</i>	<i>KIT16</i>
		<i>KM0/KM1</i>	<i>sb1/SB1</i>
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 ancient horse samples of segregating alleles at the *KIT13* and *KIT16* loci. The unit of the sampling time is the year before present (BP).

475 We set the dominance parameters $h = 0$ for *KIT13* as the *KM1* allele is dominant, and
476 $h = 0.5$ for *KIT16* as the *SB1* allele is semi-dominant. Following Der Sarkissian et al. (2015),
477 we take the population size to be $N = 16000$ and the average length of a generation of horse to
478 be 8 years, the same as in Schraiber et al. (2016). As can be seen in Table 3, there are various
479 sampling time points when the sequencing of the aDNA material yielded a number of unknown
480 alleles at loci *KIT13* and/or *KIT16*. We show all possible mutant allele frequency trajectories
481 of the sample at the *KIT13* and *KIT16* loci in Figure 9. Neither mutant allele was found in the
482 first two samples dated 17146 and 7029 years BP. Indeed, both sabino and tobiano patterns are
483 only present in domestic horses (Wutke et al., 2016). We assume that both mutant alleles, *KM1*

484 and *SB1*, arose after the domestication of the horse, which is thought to have started in the
 485 Eurasian Steppes around 5500 years BP (Outram et al., 2009). We therefore discard the first
 486 two samples from our analysis in this section, but for completeness, in Supplemental Material,
 487 Figures S13-S18, we also present the results of the inference when these two samples are taken
 488 into account.

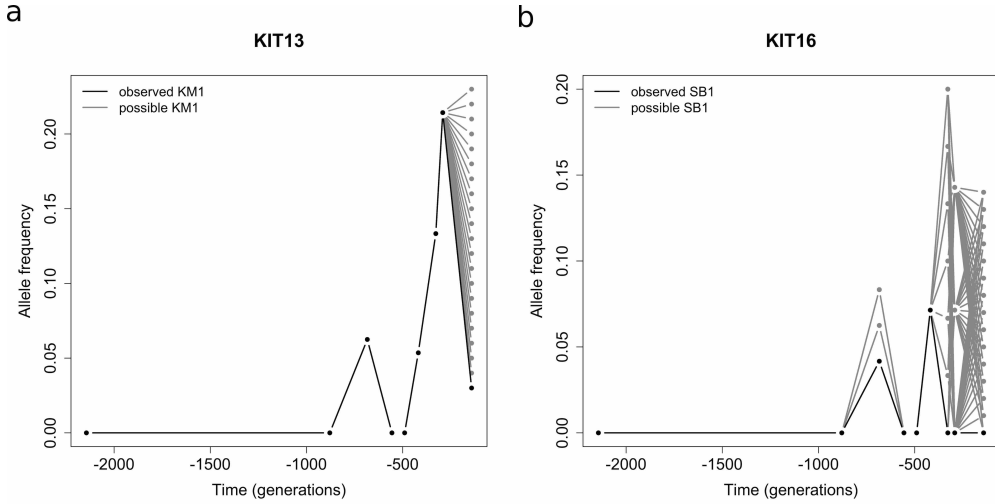


Figure 9: Potential changes in the mutant allele frequencies of the sample over time at loci (a) *KIT13* and (b) *KIT16*. Ancient horse samples were taken at generations -2144, -879, -684, -556, -490, -419, -328, -292 and -142.

489 As a result of the low quality of the *KIT* dataset, it becomes difficult to intuit whether either
 490 or both mutant alleles at the *KIT13* and *KIT16* loci are selected by simply inspecting the mutant
 491 allele frequency trajectories of the sample. Using our two-locus Bayesian inference procedure,
 492 described in Section 2.2, we jointly estimate the selection coefficients for the mutant alleles at
 493 the *KIT13* and *KIT16* loci under the case that sampled chromosomes contain variants with
 494 unknown alleles. For the recombination rate, we choose three average rates of recombination,
 495 5×10^{-9} , 1×10^{-8} and 5×10^{-8} crossovers/bp, as suggested in Dumont & Payseur (2008), and
 496 multiply them by the genetic distance 4688 bp to get the recombination rates between the *KIT13*
 497 and *KIT16* loci. All settings in the Euler-Maruyama scheme and the PMMH algorithm are the
 498 same as we applied in the previous section. The resulting posterior probability distributions are
 499 shown in Figure 10, and the MAP and MMSE estimates, as well as the 95% HPD intervals, are
 500 summarised in Table 4.

501 As can be found in Table 4, the MMSE estimates with different values of the recombination
 502 rate are essentially unchanged, while the MAP estimates vary a bit more than the MMSE

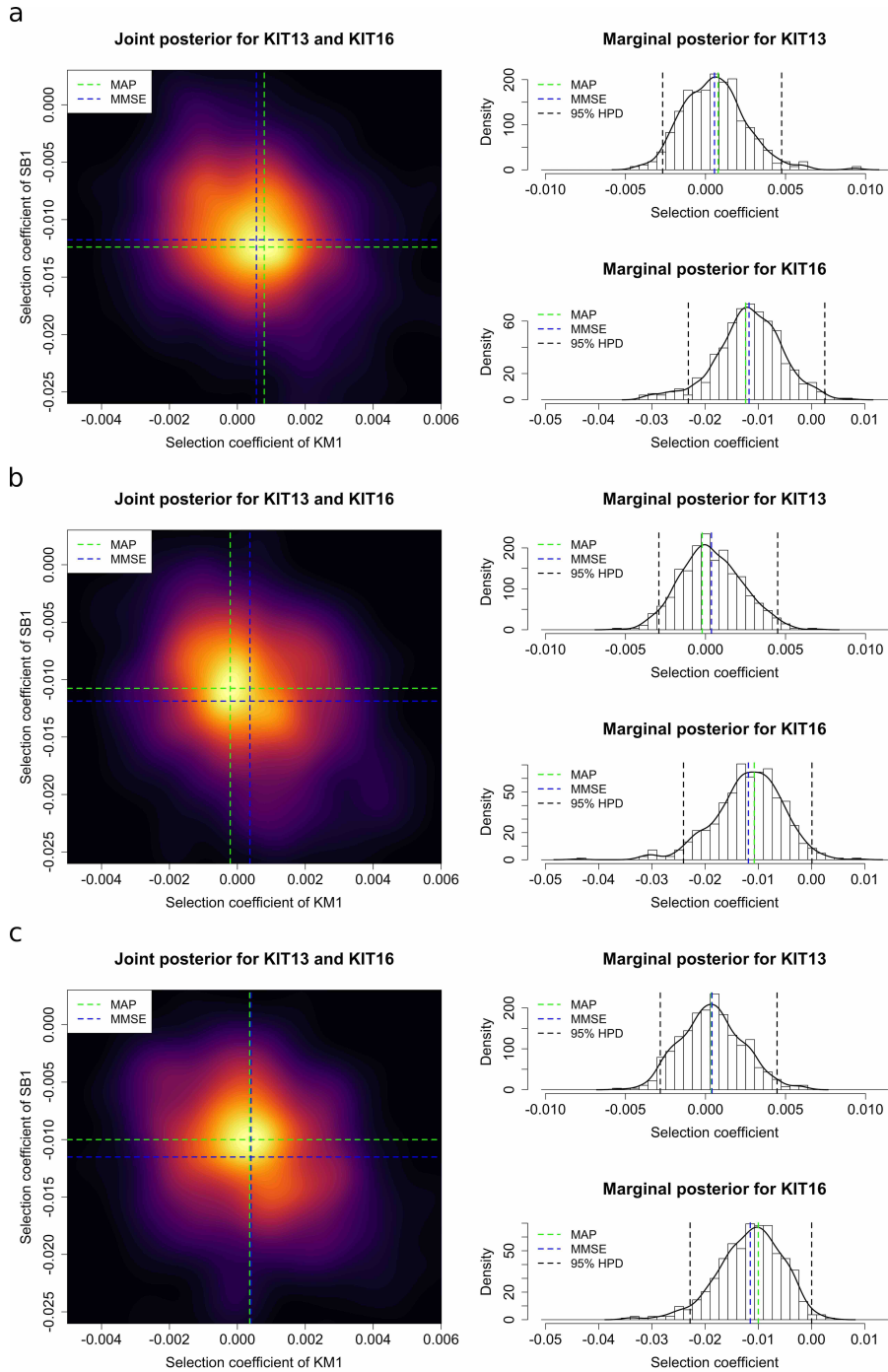


Figure 10: Posterior probability distributions for *KIT13* and *KIT16* obtained by using the two-locus method from the samples dated from 5472 years BP (the third sampling time point) with the population size of 16000 and the average rate of recombination (a) 5×10^{-9} crossovers/bp, (b) 1×10^{-8} crossovers/bp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

503 estimates. This may be caused by the way we achieve our MAP estimates, where the posterior
 504 probability distribution is approximated through the two-dimensional kernel density estimation
 505 with an axis-aligned bivariate normal kernel (Venables & Ripley, 2002). Therefore, the MAP

	recombination rate	MAP ($\times 10^{-2}$)	MMSE ($\times 10^{-2}$)	95% HPD ($\times 10^{-2}$)
<i>KIT13</i>	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]
<i>KIT16</i>	0.234×10^{-4}	-1.238	-1.175	[-2.316, 0.250]
	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 from the samples dated from 5472 years BP (the third sampling time point) with the population size of 16000.

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, but 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 the *KM1* allele at the *KIT13* locus 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 S19 and S20.

We also used our single-locus Bayesian inference procedure, described in Supplemental Material, File S3, to independently estimate the selection coefficients for the mutant alleles at the *KIT13* and *KIT16* loci under the case that sampled chromosomes contain unknown alleles. All settings in the Euler-Maruyama scheme and the PMMH algorithm are the same as we applied in the previous section. The resulting posterior probability distributions are shown in Figure 11, 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 illustrated in Figure 5, the posterior probability distributions for the *KIT13* and *KIT16* loci are both roughly symmetric about 0. This indicates that there is no evidence to support the *KM1* allele at the *KIT13* locus or the *SB1* allele at the *KIT16* locus being selected, which is consistent with the findings of Ludwig et al. (2009) obtained by using the approach of Bollback et al. (2008). Compared to the results

529 shown in Figure 10 and Table 4, we fail to tease apart negative selection at the *KIT16* locus
 530 without considering genetic recombination effect and local linkage information. We present an
 531 example that mimics the *KIT13* and *KIT16* loci, *i.e.*, a negatively selected locus tightly linked
 532 with a selectively neutral locus, which shows similar results to those using the real dataset (see
 533 Supplemental Material, Figure S21 and Table S13).

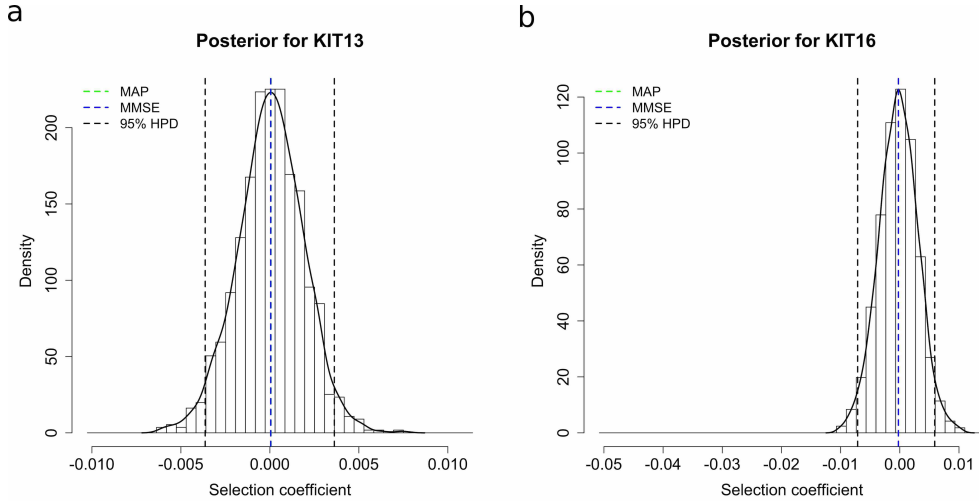


Figure 11: Posterior probability distributions for (a) *KIT13* and (b) *KIT16* obtained by using the single-locus method from the samples dated from 5472 years BP (the third sampling time point) with the population size of 16000.

	MAP ($\times 10^{-2}$)	MMSE ($\times 10^{-2}$)	95% HPD ($\times 10^{-2}$)
<i>KIT13</i>	0.006	0.005	[-0.363, 0.362]
<i>KIT16</i>	-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 from the samples dated from 5472 years BP (the third sampling time point) with the population size of 16000.

534 3.3. Computational issues

535 In the PMMH algorithm, it is desirable to generate a large number of particles in the boot-
 536 strap particle filter to yield an accurate estimate of the marginal likelihood $p(\mathbf{u}_{1:K}, \mathbf{v}_{1:K} \mid \boldsymbol{\vartheta})$.
 537 However, this can be computational burdensome since each iteration of the PMMH algorithm
 538 requires a run of the bootstrap particle filter even though fewer iterations are required. Bal-
 539 ancing the particle number and the MCMC iteration number to obtain good performance at a
 540 reasonable computational cost was investigated by Pitt et al. (2012) and Doucet et al. (2015). In
 541 pseudo-marginal algorithms, if the estimates of the marginal likelihood are too noisy, the chain

542 tends to be ‘sticky’ with excessive autocorrelation (Beaumont, 2003). A simple rule-of-thumb
543 is to select a particle number such that the standard deviation of the log-likelihood estimates is
544 in the range from 1.0 to 1.7. Nevertheless, the PMMH algorithm exactly targets the marginal
545 posterior $p(\boldsymbol{\vartheta} \mid \mathbf{u}_{1:K}, \mathbf{v}_{1:K})$ for any number of particles.

546 In each run of the bootstrap particle filter, we simulate the particles according to the two-
547 locus Wright-Fisher diffusion with selection using the Euler-Maruyama scheme. It is desirable to
548 take a large L in the Euler-Maruyama scheme to get an accurate approximation of the Wright-
549 Fisher diffusion, but large L increases the computational load. Stramer & Bogner (2011) sug-
550 gested choosing L to be L^* such that the estimates of the marginal likelihood are approximately
551 the same for any $L > L^*$, where L^* can be obtained through extensive simulations.

552 In practice, we divide each generation into 5 subintervals in the Euler-Maruyama scheme,
553 *i.e.*, $L = 5$. Our running time for a single iteration of the PMMH algorithm with 1500 particles
554 (see Figure 12), achieving the standard deviation of the log-likelihood at approximately 1.504,
555 on a single core of an Intel Core i7 processor at 4.2 GHz, is around 12.360 seconds for the *KIT*
556 dataset. In principle, every particle can be simulated in parallel on a different core. Running
557 10000 iterations of the PMMH is sufficient for a relatively smooth resulting posterior surface, as
558 shown in Figure 10. We discard the initial 2000 iterations as the burn-in period and then thin
559 the remaining PMMH output, taking every fourth value and regarding these as independent.
560 Dahlin & Schön (2015) outlined a selected number of possible improvements and best practices
561 for implementation. All of our code in this work is written in R with C++ by using Rcpp and
562 RcppArmadillo.

563 Exact-approximate particle filtering approaches such as the PMMH algorithm we use in this
564 work seem to be useful for the inference of population genetic parameters from time series data
565 of allele frequencies. This methodology can be generalised to a range of complex evolutionary
566 scenarios, *e.g.*, changing population size. Although computationally demanding, improvements
567 to the PMMH algorithm continue to be developed (*e.g.*, Yıldırım et al., 2018).

568 4. Discussion

569 In this work, we have developed a novel MCMC-based approach to jointly estimate natural
570 selection at two linked loci from time series genetic data while explicitly accounting for genetic

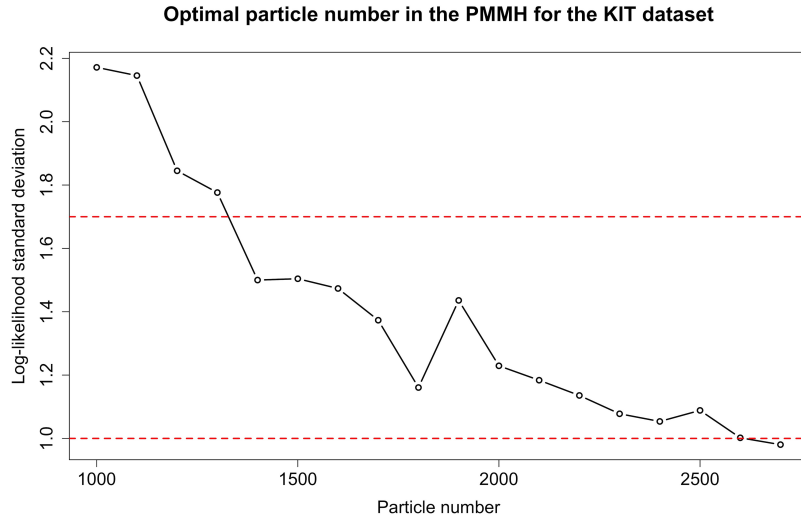


Figure 12: Changes in the standard deviation of the log-likelihood with the number of particles adopted in the PMMH algorithm for the *KIT* dataset.

571 recombination and local linkage. Our Bayesian inference procedure is built on an HMM frame-
 572 work incorporating the two-locus Wright-Fisher diffusion with selection. Our Bayesian estimates
 573 of selection coefficients are achieved with the PMMH algorithm. We have demonstrated that
 574 our method can accurately and efficiently estimate selection coefficients from simulated data,
 575 regardless of whether sampled chromosomes contain unknown alleles or not. We have found that
 576 under certain circumstances, especially in the case of tightly linked loci, existing single-locus
 577 approaches fail to deliver precise estimates for selection coefficients, but our two-locus method
 578 still works well. We have applied our Bayesian inference procedure to the *KIT* gene in horses,
 579 which is involved in the formation of white spotting patterns.

580 As noted earlier, the ancient horse DNA dataset has been the subject of earlier analyses by
 581 Malaspinas et al. (2012), Steinrücken et al. (2014), Schraiber et al. (2016) and He et al. (2019).
 582 Compared with many datasets describing experimental evolution under controlled laboratory
 583 or field mesocosm conditions, aDNA datasets are more likely to be composed of short degraded
 584 DNA fragments, typically with a high degree of genotyping error (Racimo et al., 2016). However,
 585 aDNA data provide an opportunity to investigate the chronology and tempo of natural selection
 586 across evolutionary timescales, which has an advantage of being associated with an interesting
 587 narrative (MacHugh et al., 2017). A motivation for the analysis is to see whether the statistical
 588 developments described here can shed further light on these data. We have found strong evidence
 589 showing that the sabino pattern caused by the *SB1* allele at locus *KIT16* has been selectively

590 deleterious but no evidence showing that the tobiano pattern caused by the *KM1* allele at locus
591 *KIT13* has been selectively advantageous. One explanation for our findings may be that there
592 was a decreasing attractiveness of spotted horses since the Middle Ages due to the religious and
593 cultural ideas (Wutke et al., 2016). Based on ancient Roman records, solid horses were preferred
594 to spotted horses as the latter was considered to be of inferior quality. During medieval times,
595 spotted horses had a negative connotation after several epidemics, resulting in a lower religious
596 prestige for these patterns. Additionally, people might no longer see the need to distinguish wild
597 (solid) horses from domesticated (spotted) horses as wild populations gradually became scarcer
598 and approached extinction. Further reasons for the spotted horses being selectively deleterious
599 might have been novel developments in weaponry such as the longbow, with these horses being
600 an easier target than solid ones (see Wutke et al., 2016, and references therein).

601 In addition to our method, Terhorst et al. (2015) is the only existing approach that can
602 model linked loci and genetic drift for the inference of natural selection from temporal changes
603 in allele frequencies. In Terhorst et al. (2015), the underlying population dynamics at multiple
604 linked loci was modelled using the Wright-Fisher model in their HMM framework, and the
605 likelihood computation was carried out by approximating the Wright-Fisher model through a
606 deterministic path with added Gaussian noises, which aims to fit a mathematically convenient
607 transition probability density function by equating the first two moments of the Wright-Fisher
608 model. Such a moment-based approximation works well for many applications when modelling
609 the allele frequencies with intermediate values. However, as soon as the allele frequencies get
610 close to their boundaries 0 or 1 (*i.e.*, allele loss or fixation), the Wright-Fisher model will be
611 poorly approximated due to the infinite support of the Gaussian distribution that will leak
612 probability mass into the values of the allele frequency that are smaller than 0 or larger than
613 1, which is not mathematically possible. This issue becomes more problematic in the inference
614 of natural selection since natural selection is expected to rapidly drive the allele frequencies
615 towards the boundaries.

616 The MCMC-based method we have developed in this work is built on the standard diffusion
617 limit of the Wright-Fisher model of the stochastic evolutionary dynamics under natural selection
618 at a pair of linked loci, which is shown to be a good approximation even if the allele frequencies
619 get close to their boundaries 0 or 1 (He et al., 2020). The diffusion approximation enables our

620 approach to work well for the allele frequencies with all possible values. Our method can handle
621 sampled chromosomes that contain unknown alleles, which one might expect to encounter in
622 real data, especially in aDNA studies. Even though we have only illustrated the utility of our
623 method on aDNA data in this work, our Bayesian inference procedure can also be used to analyse
624 Pool-Seq time series data from E&R experiments, as in Terhorst et al. (2015). Given the PMMH
625 algorithm, we have used to infer natural selection in this work, our method lends itself naturally
626 to joint estimates of the haplotype frequency trajectories of the underlying population without
627 any increase in computational complexity. Furthermore, our method can be readily extended to
628 model a range of complex evolutionary scenarios, *e.g.*, time-varying population size and selection
629 coefficients, as it is built on simulating the Wright-Fisher diffusion.

630 One limitation of our approach is that we assume that mutant alleles were created before
631 the initial sampling time point. Once a sample contains at least one copy of the mutant allele,
632 we can reasonably assume that the mutant allele arose before the time of that sample. However
633 in the case of earlier samples without any mutant allele, there is uncertainty in pinpointing
634 when the mutant allele arose. This problem can be remedied by co-estimating the allele age as
635 in *e.g.*, Malaspinas et al. (2012), Schraiber et al. (2016) and He et al. (2019), but these works
636 all investigate natural selection at a single locus. Jointly estimating the selection coefficients at
637 linked loci along with the allele ages can be expected to be cumbersome as there are many cases
638 to take into account. In the case of the ancient horse DNA data, we did not wish to make the
639 assumption that the mutant alleles, *KM1* and *SB1*, arose earlier than the time that horses were
640 domesticated. However, we can compare the inference results obtained with different choices of
641 the initial sampling time point (see Supplemental Material, Tables S14-S16) and reach the same
642 conclusion that there is no strong evidence for the *KM1* allele at locus *KIT13* to be positively
643 selected, but there is strong evidence for the *SB1* allele at locus *KIT16* to be negatively selected.

644 Our Bayesian statistical framework lends itself to being extended to infer natural selection at
645 multiple linked loci from time series data of allele frequencies, which might further improve the
646 inference results of natural selection. The challenge is that with the increase in the number of
647 linked loci, modelling the underlying population dynamics subject to natural selection becomes
648 increasingly difficult. For example, there are eight haplotypes to take into account in the case
649 of three linked loci each with two alleles. As a tractable alternative, we can apply our approach

650 to multiple linked loci in a pairwise manner by using the PMMH algorithm within the Gibbs
651 sampler, but it might only work for a small number of linked loci due to the computational cost
652 of our two-locus approach. In practice, it will be necessary to find a good approximation of the
653 Wright-Fisher model for the method to be computationally feasible, which will be the topic of
654 future investigation. An important consideration is to what degree the results of the inference
655 of natural selection are affected by the choice of stochastic or deterministic dynamics for the
656 allele frequency trajectories (Jewett et al., 2016), and whether in many scenarios approximation
657 with a deterministic model is satisfactory.

658 **Acknowledgements**

659 We are grateful to the anonymous reviewers and the communicating editors for their helpful
660 comments on the earlier version of this work. This work was funded in part by the Engineering
661 and Physical Sciences Research Council (EPSRC) Grant EP/I028498/1 to F.Y.

662 **References**

- 663 Andrieu, C., Doucet, A., & Holenstein, R. (2010). Particle Markov chain Monte Carlo methods.
664 *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*, 269–342.
- 665 Andrieu, C., & Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo
666 computations. *The Annals of Statistics*, *37*, 697–725.
- 667 Andrieu, C., & Vihola, M. (2016). Establishing some order amongst exact approximations of
668 MCMCs. *The Annals of Applied Probability*, *26*, 2661–2696.
- 669 Bank, C., Ewing, G. B., Ferrer-Admettla, A., Foll, M., & Jensen, J. D. (2014). Thinking too
670 positive? Revisiting current methods of population genetic selection inference. *Trends in*
671 *Genetics*, *30*, 540–546.
- 672 Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored
673 populations. *Genetics*, *164*, 1139–1160.
- 674 Bollback, J. P., & Huelsenbeck, J. P. (2007). Clonal interference is alleviated by high mutation
675 rates in large populations. *Molecular Biology and Evolution*, *24*, 1397–1406.

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

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

680 Brooks, S. A., Lear, T. L., Adelson, D. L., & Bailey, E. (2007). A chromosome inversion near
681 the KIT gene and the Tobiano spotting pattern in horses. *Cytogenetic and Genome Research*,
682 *119*, 225–230.

683 Burke, M. K., Dunham, J. P., Shahrestani, P., Thornton, K. R., Rose, M. R. et al. (2010).
684 Genome-wide analysis of a long-term evolution experiment with *Drosophila*. *Nature*, *467*,
685 587–590.

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

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

690 Der Sarkissian, C., Ermini, L., Schubert, M., Yang, M. A., Librado, P. et al. (2015). Evolutionary
691 genomics and conservation of the endangered Przewalski’s horse. *Current Biology*, *25*, 2577–
692 2583.

693 Doucet, A., Pitt, M. K., Deligiannidis, G., & Kohn, R. (2015). Efficient implementation of
694 Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, *102*,
695 295–313.

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

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

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

702 Ferrer-Admetlla, A., Leuenberger, C., Jensen, J. D., & Wegmann, D. (2016). An approximate
703 Markov model for the Wright-Fisher diffusion and its application to time series data. *Genetics*,
704 *203*, 831–846.

705 Fisher, R. A. (1922). On the dominance ratio. *Proceedings of the Royal Society of Edinburgh*,
706 *42*, 321–341.

707 Foll, M., Poh, Y.-P., Renzette, N., Ferrer-Admetlla, A., Bank, C. et al. (2014). Influenza
708 virus drug resistance: a time-sampled population genetics perspective. *PLoS Genetics*, *10*,
709 e1004185.

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

713 Gordon, N. J., Salmond, D. J., & Smith, A. F. M. (1993). Novel approach to nonlinear/non-
714 Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, *140*,
715 107–113.

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

719 He, Z., Beaumont, M. A., & Yu, F. (2017). Effects of the ordering of natural selection and
720 population regulation mechanisms on Wright-Fisher models. *G3: Genes, Genomes, Genetics*,
721 *7*, 2095–2106.

722 He, Z., Beaumont, M. A., & Yu, F. (2020). Numerical simulation of the two-locus Wright-Fisher
723 stochastic differential equation with application to approximating transition probability den-
724 sities. *bioRxiv*, (p. 213769).

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

727 Holder, K. K., & Bull, J. J. (2001). Profiles of adaptation in two similar viruses. *Genetics*, *159*,
728 1393–1404.

- 729 Hummel, S., Schmidt, D., Kremeyer, B., Herrmann, B., & Oppermann, M. (2005). Detection
730 of the CCR5- Δ 32 HIV resistance gene in Bronze Age skeletons. *Genes and Immunity*, *6*,
731 371–374.
- 732 Izenman, A. J. (1991). Recent developments in nonparametric density estimation. *Journal of*
733 *the American Statistical Association*, *86*, 205–224.
- 734 Jewett, E. M., Steinrücken, M., & Song, Y. S. (2016). The effects of population size histories
735 on estimates of selection coefficients from time-series genetic data. *Molecular Biology and*
736 *Evolution*, *33*, 3002–3027.
- 737 Kloeden, P. E., & Platen, E. (1992). *Numerical Solution of Stochastic Differential Equations*.
738 Berlin: Springer-Verlag.
- 739 Lacerda, M., & Seoighe, C. (2014). Population genetics inference for longitudinally-sampled
740 mutants under strong selection. *Genetics*, *198*, 1237–1250.
- 741 Lang, G. I., Rice, D. P., Hickman, M. J., Sodergren, E., Weinstock, G. M. et al. (2013). Pervasive
742 genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature*, *500*,
743 571–574.
- 744 Lewontin, R. C., & Kojima, K.-i. (1960). The evolutionary dynamics of complex polymorphisms.
745 *Evolution*, *14*, 458–472.
- 746 Ludwig, A., Pruvost, M., Reissmann, M., Benecke, N., Brockmann, G. A. et al. (2009). Coat
747 color variation at the beginning of horse domestication. *Science*, *324*, 485–485.
- 748 MacHugh, D. E., Larson, G., & Orlando, L. (2017). Taming the past: ancient DNA and the
749 study of animal domestication. *Annual Review of Animal Biosciences*, *5*, 329–351.
- 750 Malaspinas, A.-S. (2016). Methods to characterize selective sweeps using time serial samples:
751 an ancient DNA perspective. *Molecular Ecology*, *25*, 24–41.
- 752 Malaspinas, A.-S., Malaspinas, O., Evans, S. N., & Slatkin, M. (2012). Estimating allele age
753 and selection coefficient from time-serial data. *Genetics*, *192*, 599–607.
- 754 Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N. et al. (2015). Genome-wide
755 patterns of selection in 230 ancient Eurasians. *Nature*, *528*, 499–503.

- 756 Mathieson, I., & McVean, G. (2013). Estimating selection coefficients in spatially structured
757 populations from time series data of allele frequencies. *Genetics*, *193*, 973–984.
- 758 Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A. et al. (2013). Recalibrating
759 Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*,
760 *499*, 74–78.
- 761 Orozco-terWengel, P., Kapun, M., Nolte, V., Kofler, R., Flatt, T. et al. (2012). Adaptation of
762 *Drosophila* to a novel laboratory environment reveals temporally heterogeneous trajectories
763 of selected alleles. *Molecular Ecology*, *21*, 4931–4941.
- 764 Outram, A. K., Stear, N. A., Bendrey, R., Olsen, S., Kasparov, A. et al. (2009). The earliest
765 horse harnessing and milking. *Science*, *323*, 1332–1335.
- 766 Paris, C., Servin, B., & Boitard, S. (2019). Inference of selection from genetic time series
767 using various parametric approximations to the Wright-Fisher model. *G3: Genes, Genomes,*
768 *Genetics*, *9*, 4073–4086.
- 769 Pitt, M. K., dos Santos Silva, R., Giordani, P., & Kohn, R. (2012). On some properties of Markov
770 chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*,
771 *171*, 134–151.
- 772 Pruvost, M., Bellone, R., Benecke, N., Sandoval-Castellanos, E., Cieslak, M. et al. (2011).
773 Genotypes of predomestic horses match phenotypes painted in Paleolithic works of cave art.
774 *Proceedings of the National Academy of Sciences*, *108*, 18626–18630.
- 775 Racimo, F., Renaud, G., & Slatkin, M. (2016). Joint estimation of contamination, error and
776 demography for nuclear DNA from ancient humans. *PLoS Genetics*, *12*, e1005972.
- 777 Ragsdale, A. P., & Gutenkunst, R. N. (2017). Inferring demographic history using two-locus
778 statistics. *Genetics*, *206*, 1037–1048.
- 779 Sato, K.-I. (1976). Diffusion processes and a class of Markov chains related to population
780 genetics. *Osaka Journal of Mathematics*, *13*, 631–659.
- 781 Schraiber, J. G., Evans, S. N., & Slatkin, M. (2016). Bayesian inference of natural selection
782 from allele frequency time series. *Genetics*, *203*, 493–511.

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

786 Steinrücken, M., Bhaskar, A., & Song, Y. S. (2014). A novel spectral method for inferring
787 general diploid selection from time series genetic data. *The Annals of Applied Statistics, 8*,
788 2203–2222.

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

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

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

795 Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S-PLUS*. New York:
796 Springer-Verlag.

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

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

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

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

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

806 Wutke, S., Benecke, N., Sandoval-Castellanos, E., Döhle, H.-J., Friederich, S. et al. (2016).
807 Spotted phenotypes in horses lost attractiveness in the Middle Ages. *Scientific Reports, 6*,
808 38548.

- 809 Yildirim, S., Andrieu, C., & Doucet, A. (2018). Scalable Monte Carlo inference for state-space
810 models. *arXiv preprint arXiv:1809.02527*, .
- 811 Yu, F., & Etheridge, A. (2010). The fixation probability of two competing beneficial mutations.
812 *Theoretical Population Biology*, 78, 36–45.
- 813 Zhang, X. (2006). Euler-Maruyama approximations for SDEs with non-Lipschitz coefficients
814 and applications. *Journal of Mathematical Analysis and Applications*, 316, 447–458.